**А.И.Орлов**

# ПРИКЛАДНАЯ СТАТИСТИКА

*Учебник для вузов*

*Издательство*

***«ЭКЗАМЕН»***

МОСКВА

2004

Орлов А.И.
Прикладная статистика. Учебник. / А.И.Орлов.- М.: Издательство «Экзамен», 2004. - 656 с.

## Аннотация

Учебник посвящен основным методам современной прикладной статистики. В первой части рассмотрен вероятностно-статистический фундамент прикладной статистики. Основные проблемы прикладной статистики – описание данных, оценивание, проверка гипотез – разобраны во второй части. Методам статистического анализа числовых величин, многомерного статистического анализа, временных рядов, статистики нечисловых и интервальных данных посвящена третья часть учебника. Обсуждается методология прикладной статистики, ее современное состояние и перспективы развития. Изложение соответствует рекомендациям Российской академии статистических методов.

Каждая глава учебника – это введение в большую область прикладной статистики. Приведенные литературные ссылки помогут выйти на передний край теоретических и прикладных работ, познакомиться с доказательствами теорем, помещенных в учебник.

Для студентов и преподавателей вузов, слушателей институтов повышения квалификации, структур второго образования и программ МВА («Мастер делового администрирования»), инженеров различных специальностей, менеджеров, экономистов, социологов, научных и практических работников, связанных с анализом данных.

# ОГЛАВЛЕНИЕ

## Часть 3. Методы прикладной статистики

## Часть 4. Заключение. Современная прикладная статистика



## Приложения

# ПРЕДИСЛОВИЕ

Прикладная статистика – это наука о том, как обрабатывать данные. Методы прикладной статистики активно применяются в технических исследованиях, экономике, теории и практике управления (менеджмента), социологии, медицине, геологии, истории и т.д. С результатами наблюдений, измерений, испытаний, опытов, с их анализом имеют дело специалисты во всех отраслях практической деятельности, почти во всех областях теоретических исследований. Настоящий учебник позволяет овладеть современными методами прикладной статистики на уровне, достаточном для использования этих методов в научной и практической деятельности.

*Содержание учебника.* Учебник посвящен основным методам современной прикладной статистики и состоит из четырех частей. В первой части рассмотрен вероятностно-статистический фундамент прикладной статистики. Для удобства читателей включены основы современной теории вероятностей и математической статистики, на которых базируется прикладная статистика.

Основные проблемы прикладной статистики – описание данных, оценивание, проверка гипотез – разобраны во второй части. Методам статистического анализа числовых величин, многомерного статистического анализа, временных рядов, статистики нечисловых и интервальных данных посвящена третья часть учебника. В заключительной четвертой части обсуждаются перспективы развития прикладной статистики и ее методология. В конце каждой главы приведены процитированные в ней литературные источники, контрольные вопросы и задачи, а также темы докладов, рефератов, исследовательских работ. Нумерация таблиц, рисунков, формул, теорем, примеров проводится по главам, в отдельных случаях – по подразделам (параграфам, пунктам).

Общее количество статей и книг по прикладной статистике давно превысило $10^6$, из них актуальными к настоящему времени являются не менее $10^5$. Конкретный специалист может овладеть несколькими тысячами из них. Следовательно, ни один исследователь не может претендовать на знакомство более чем с 2-3% актуальных публикаций, и в любом учебнике содержится лишь небольшая часть знаний, накопленных в прикладной статистике. Однако автор надеется, что наиболее важные подходы, идеи, результаты и алгоритмы расчетов включены в учебник. Эта надежда основана на более чем тридцатилетнем опыте теоретической и практической работы в прикладной статистике, на совокупном опыте членов научных сообществ, скрупулезном анализе положения в прикладной статистике при создании Всесоюзной статистической ассоциации, Российской ассоциации статистических методов и Российской академии статистических методов.

В отличие от учебной литературы по математическим дисциплинам, в настоящей книге практически отсутствуют доказательства. Однако в нескольких случаях мы сочли целесообразным их привести. При первом чтении доказательства теорем можно пропустить.

О роли литературных ссылок в учебнике необходимо сказать достаточно подробно. Прежде всего, эта книга представляет собой замкнутый текст, не требующий для своего понимания ничего, кроме знания стандартных учебных курсов высшей математике. Зачем же нужны ссылки? Доказательства всех приведенных в учебнике теорем приведены в ранее опубликованных статьях и монографиях. Дотошный читатель, в частности, при подготовке рефератов и при желании глубже проникнуть в материал учебника, может обратиться к приведенным в каждой главе спискам цитированной литературы. Каждая глава учебника – это введение в большую область прикладной статистики. Приведенные литературные ссылки помогут читателям выйти на передний край теоретических и прикладных работ, познакомиться с доказательствами теорем, включенных в учебник. За многие десятилетия накопились большие книжные богатства, и их надо активно использовать.

Включенные в учебник материалы прошли многолетнюю и всестороннюю проверку. Кроме МГТУ им. Н.Э.Баумана, они использовались при преподавании во многих других отечественных и зарубежных образовательных структурах. О некоторых из них можно получить представление из справки «Об авторе этой книги» в конце учебника.

В 2002 и 2003 гг. издательством «Экзамен» был выпущен двумя изданиями учебник «Эконометрика» А.И.Орлова. Это говорит об актуальности тематики настоящего учебника, поскольку под эконометрикой понимают применение статистических методов (в том числе прикладной статистики) в экономике и управлении (менеджменте).

*Для кого написан учебник?* Учебник предназначен для студентов различных специальностей, прежде всего технических, управленческих и экономических, слушателей институтов повышения квалификации, структур послевузовского (в том числе второго) образования, в частности, программ МВА («Мастер делового администрирования»), преподавателей вузов. Он будет полезен инженерам, менеджерам, экономистам, социологам, биологам, медикам, психологам, историкам, другим специалистам, самостоятельно повышающим свой научный уровень. Короче, всем научным и практическим работникам, связанным с анализом данных.

Учебник может быть использован при изучении дисциплин, полностью или частично посвященным методам анализа результатов наблюдений (измерений, испытаний, опытов). Типовые названия таких вузов - «Прикладная статистика», «Эконометрика», «Анализ данных», «Многомерный статистический анализ», «Общая теория статистики», «Планирование эксперимента», «Биометрика», «Теория принятия решений», «Управленческие решения», «Экономико-математическое моделирование», «Математические методы прогнозирования», «Прогнозирование и технико-экономическое планирование», «Хемометрия», «Математические методы в социологии», «Математические методы в геологии» и т.п.

Специалистам по теории вероятностей и математической статистике эта книга также может быть интересна и полезна, поскольку в ней описан современный взгляд на прикладную математическую статистику, основные подходы и результаты в этой области, открывающие большой простор для дальнейших математических исследований.

*Отечественная научная школа по прикладной статистике.* В нашей стране прикладная статистика активно развивалась с начала 1980-х годов. В 1990 г. при создании Всесоюзной статистической ассоциации (ВСА) одной из ее четырех секций была секция прикладной статистики, а руководитель этой секции А.И.Орлов был избран вице-президентом ВСА. В XXI в. развитие при5кладной статистики продолжается в рамках Российской ассоциации статистических методов и Российской академии статистических методов.

По ряду причин исторического характера основное место публикаций научных работ по прикладной статистике в нашей стране - отдел "Математические методы исследования" журнала "Заводская лаборатория". В отделе публикуются статьи по статистическим методам анализа технических и технико-экономических данных. Автор искренне благодарен главному редактору журнала академику РАН Н.П.Лякишеву, зам. главного редактора М.Г.Плотницкой, редактору отдела М.Е.Носовой. Автору приятно выразить радость от возможности работать вместе со своими коллегами по секции "Математические методы исследования", прежде всего с заслуженным деятелем науки РФ проф. В.Г.Горским. Автор искренне благодарен своим учителям - академику АН УССР Б.Г. Гнеденко, члену-корреспонденту АН СССР Л.Н. Большеву, проф. В.В. Налимову.

Автор искренне благодарен заведующему кафедрой "Экономика и организация производства" факультета "Инженерный бизнес и менеджмент" Московского государственного технического университета им. Н.Э. Баумана профессору, доктору экономических наук С.Г. Фалько за постоянную поддержку проекта по разработке и внедрению эконометрических курсов. Хотелось бы сказать спасибо всему коллективу кафедры и факультета в целом, декану и членам Ученого Совета, поддержавшим инициативу о введении эконометрики в учебный процесс МГТУ им. Н.Э.Баумана.

С текущей научной информацией по прикладной статистике проще всего познакомиться на сайтах автора www.antorlov.chat.ru, www.newtech.ru/~orlov, www.antorlov.euro.ru , входящих в Интернет. Достаточно большой объем информации содержит электронный еженедельник "Эконометрика", выпускаемый с июля 2000 г. (автор искренне благодарен редактору этого электронного издания А.А.Орлову за многолетний энтузиазм по выпуску еженедельника).

В учебнике изложено представление о прикладной статистике, соответствующее общепринятому в мире. Изложение доведено до современного уровня научных исследований в этой области. Конечно, возможны различные точки зрения по тем или иным частным вопросам. Автор будет благодарен читателям, если они сообщат свои вопросы и замечания по адресу издательства или непосредственно автору по электронной почте E-mail: orlov@professor.ru .

# Введение. Прикладная статистика как область научно-практической деятельности

**Развитие представлений о статистике.** Впервые термин «статистика» мы находим в художественной литературе – в «Гамлете» Шекспира (1602 г., акт 5, сцена 2). Смысл этого слова у Шекспира – знать, придворные. По-видимому, оно происходит от латинского слова *status*, что в оригинале означает «состояние» или «политическое состояние».

В течение следующих 400 лет термин «статистика» понимали и понимают по-разному. В работе [1] собрано более 200 определений этого термина, некоторые из которых приводятся ниже.

Вначале под статистикой понимали описание экономического и политического состояния государства или его части. Например, к 1792 г. относится определение: «Статистика описывает состояние государства в настоящее время или в некоторый известный момент в прошлом». И в настоящее время деятельность государственных статистических служб (в нашей стране – Государственного комитета РФ по статистике) вполне укладывается в это определение.

Однако постепенно термин «статистика» стал использоваться более широко. По Наполеону Бонапарту «Статистика – это бюджет вещей». Тем самым статистические методы были признаны полезными не только для административного управления, но и на уровне отдельного предприятия. Согласно формулировке 1833 г. «цель статистики заключается в представлении фактов в наиболее сжатой форме». Приведем еще два высказывания. Статистика состоит в в наблюдении явлений, которые могут быть подсчитаны или выражены посредством чисел (1895). Статистика – это численное представление фактов из любой области исследования в их взаимосвязи (1909).

В XX в. статистику часто рассматривают прежде всего как самостоятельную научную дисциплина. Статистика есть совокупность методов и принципов, согласно которым проводится сбор, анализ, сравнение, представление и интерпретация числовых данных (1925). В 1954 г. академик АН УССР Б.В. Гнеденко дал следующее определение: «Статистика состоит из трех разделов:

1) сбор статистических сведений, т.е. сведений, характеризующих отдельные единицы каких-либо массовых совокупностей;

2) статистическое исследование полученных данных, заключающееся в выяснении тех закономерностей, которые могут быть установлены на основе данных массового наблюдения;

3) разработка приемов статистического наблюдения и анализа статистических данных. Последний раздел, собственно, и составляет содержание математической статистики».

Термин «статистика» употребляют еще в двух смыслах. Во-первых, в обиходе под «статистикой» часто понимают набор количественных данных о каком-либо явлении или процессе. Во-вторых, статистикой называют функцию от результатов наблюдений, используемую для оценивания характеристик и параметров распределений и проверки гипотез.

Чтобы подойти к термину «прикладная статистика», кратко рассмотрим историю реальных статистических работ.

**Краткая история статистических методов.** Типовые примеры раннего этапа применения статистических методов описаны в Ветхом Завете (см., например, Книгу Чисел). Там, в частности, приводится число воинов в различных племенах. С математической точки зрения дело сводилось к подсчету числа попаданий значений наблюдаемых признаков в определенные градации.

В дальнейшем результаты обработки статистических данных стали представлять в виде таблиц и диаграмм, как это и сейчас делает Госкомстат РФ. Надо признать, что по сравнению с Ветхим Заветом есть прогресс - в Библии не было таблиц и диаграмм. Однако нет продвижения по сравнению с работами российских статистиков конца девятнадцатого - начала двадцатого века (типовой монографией тех времен можно считать книгу [2], которая в настоящее время ещё легко доступна).

Сразу после возникновения теории вероятностей (Паскаль, Ферма, 17 век) вероятностные модели стали использоваться при обработке статистических данных. Например, изучалась частота рождения мальчиков и девочек, было установлено отличие вероятности рождения мальчика от 0.5, анализировались причины того, что в парижских приютах эта вероятность не та, что в самом Париже, и т.д. Имеется достаточно много публикаций по истории теории вероятностей с описанием раннего этапа развития статистических методов исследований, к лучшим из них относится очерк [3].

В 1794 г. (по другим данным - в 1795 г.) К.Гаусс разработал метод наименьших квадратов, один из наиболее популярных ныне статистических методов, и применил его при расчете орбиты астероида Церера - для борьбы с ошибками астрономических наблюдений [4]. В Х1Х веке заметный вклад в развитие практической статистики внес бельгиец Кетле, на основе анализа большого числа реальных данных показавший устойчивость относительных статистических показателей, таких, как доля самоубийств среди всех смертей [5]. Интересно, что основные идеи статистического приемочного контроля и сертификации продукции обсуждались академиком Петербургской АН М.В. Остроградским (1801-1862) и применялись в российской армии ещё в середине Х1Х в. [3]. Статистические методы управления качеством и сертификации продукции сейчас весьма актуальны [6].

Современный этап развития статистических методов можно отсчитывать с 1900 г., когда англичанин К. Пирсон основан журнал «*Biometrika*». Первая треть ХХ в. прошла под знаком параметрической статистики. Изучались методы, основанные на анализе данных из параметрических семейств распределений, описываемых кривыми семейства Пирсона. Наиболее популярным было нормальное (гауссово) распределение. Для проверки гипотез использовались критерии Пирсона, Стьюдента, Фишера. Были предложены метод максимального правдоподобия, дисперсионный анализ, сформулированы основные идеи планирования эксперимента.

Разработанную в первой трети ХХ в. теорию анализа данных называем параметрической статистикой, поскольку ее основной объект изучения - это выборки из распределений, описываемых одним или небольшим числом параметров. Наиболее общим является семейство кривых Пирсона, задаваемых четырьмя параметрами. Как правило, нельзя указать каких-либо веских причин, по которым распределение результатов конкретных наблюдений должно входить в то или иное параметрическое семейство. Исключения хорошо известны: если вероятностная модель предусматривает суммирование независимых случайных величин, то сумму естественно описывать нормальным распределением; если же в модели рассматривается произведение таких величин, то итог, видимо, приближается логарифмически нормальным распределением, и т.д. Однако подобных моделей нет в подавляющем большинстве реальных ситуаций, и приближение реального распределения с помощью кривых из семейства Пирсона или его подсемейств - чисто формальная операция.

Именно из таких соображений критиковал параметрическую статистику академик АН СССР С.Н.Бернштейн в 1927 г. в своем докладе на Всероссийском съезде математиков [7]. Однако эта теория, к сожалению, до сих пор остается основой преподавания статистических методов и продолжает использоваться основной массой прикладников, далеких от новых веяний в статистике. Почему так происходит? Чтобы попытаться ответить на этот вопрос, обратимся к наукометрии.

**Наукометрия статистических исследований.** В рамках движения за создание Всесоюзной статистической ассоциации (учреждена в 1990 г.) был проведен назад анализ статистики как области научно-практической деятельности. Он показал, в частности, что актуальными для специалистов в настоящее время являются не менее чем 100 тысяч публикаций (подробнее см. статьи [8,9]). Реально же каждый из нас знаком с существенно меньшим количеством книг и статей. Так, в известном трехтомнике М Кендалла и А. Стьюарта [10-12] – наиболее полном на русском языке издании по статистическим методам - всего около 2 тысяч литературных ссылок. При всей очевидности соображений о многократном дублировании в публикациях ценных идей приходится признать, что каждый специалист по прикладной статистике владеет лишь небольшой частью

накопленных в этой области знаний. Не удивительно, что приходится постоянно сталкиваться с игнорированием или повторением ранее полученных результатов, с уходом в тупиковые (с точки зрения практики) направления исследований, с беспомощностью при обращении к реальным данным, и т.д. Все это - одно из проявлений адапционного механизма торможения развития науки, о котором еще 30 лет назад писали В.В.Налимов и другие науковеды (см., например, [13]).

Традиционный предрассудок состоит в том, что каждый новый результат, полученный исследователем - это кирпич в непрерывно растущее здание науки, который непременно будет проанализирован и использован научным сообществом, а затем и при решении практических задач. Реальная ситуация - совсем иная. Основа профессиональных знаний исследователя, инженера, экономиста менеджера, социолога, историка, геолога, медика закладывается в период обучения. Затем знания пополняются в том узком направлении, в котором работает специалист. Следующий этап - их тиражирование новому поколению. В результате вузовские учебники отстоят от современного развития на десятки лет. Так, учебники по математической статистике, согласно мнению экспертов, по научному уровню в основном соответствуют 40-60-м годам XX в. А потому середине XX в. соответствует большинство вновь публикуемых исследований и тем более - прикладных работ. Одновременно приходится признать, что результаты, не вошедшие в учебники, независимо от их ценности почти все забываются.

Активно продолжается развитие тупиковых направлений. Психологически это понятно. Приведу пример из своего опыта. В свое время по заказу Госстандарта я разработал методы оценки параметров гамма-распределения [14]. Поэтому мне близки и интересны работы по оцениванию параметров по выборкам из распределений, принадлежащих тем или иным параметрическим семействам, понятия функции максимального правдоподобия, эффективности оценок, использование неравенства Рао-Крамера и т.д. К сожалению, я знаю, что это - тупиковая ветвь теории статистики, поскольку реальные данные не подчиняются каким-либо параметрическим семействам, надо применять иные статистические методы, о которых речь пойдет ниже. Понятно, что специалистам по параметрической статистике, потратившим многие годы на совершенствование в своей области, психологически трудно согласиться с этим утверждением. В том числе и мне. Но необходимо идти вперед. Поэтому настоящий учебник очищен от тупиковых подходов. В том числе и от неравенства Рао-Крамера.

**Появление прикладной статистики**. В нашей стране термин «прикладная статистика» вошел в широкое употребление в 1981 г. после выхода массовым тиражом (33940 экз.) сборника «Современные проблемы кибернетики (прикладная статистика)». В этом сборнике обосновывалась трехкомпонентная структура прикладной статистики [15]. Во-первых, в нее входят ориентированные на прикладную деятельность статистические методы анализа данных (эту область можно назвать прикладной математической статисткой и включать также и в прикладную математику). Однако прикладную статистику нельзя целиком относить к математике. Она включает в себя две внематематические области. Во-первых, методологию организации статистического исследования: как планировать исследование, как собирать данные, как подготавливать данные к обработке, как представлять результаты. Во-вторых, организацию компьютерной обработки данных, в том числе разработку и использование баз данных и электронных таблиц, статистических программных продуктов, например, диалоговых систем анализа данных.

В нашей стране термин «прикладная статистика» использовался и ранее 1981 г., но лишь внутри сравнительно небольших и замкнутых групп специалистов [15].

Прикладная статистика и математическая статистика – это две разные научные дисциплины. Различие четко проявляется и при преподавании. Курс математической статистики состоит в основном из доказательств теорем, как и соответствующие учебные пособия. В курсах прикладной статистики основное - методология анализа данных и алгоритмы расчетов, а теоремы приводятся как обоснования этих алгоритмов, доказательства же, как правило, опускаются (их можно найти в научной литературе).

**Структура современной статистики**. Внутренняя структура статистики как науки была выявлена и обоснована при создании в 1990 г. Всесоюзной статистической

ассоциации [9]. Прикладная статистика - методическая дисциплина, являющаяся центром статистики. При применении методов прикладной статистики к конкретным областям знаний и отраслям народного хозяйства получаем научно-практические дисциплины типа "статистика в промышленности", "статистика в медицине" и др. С этой точки зрения эконометрика - это "статистические методы в экономике" [6]. Математическая статистика играет роль математического фундамента для прикладной статистики.

К настоящему времени очевидно четко выраженное размежевание этих двух научных направлений. Математическая статистика исходит из сформулированных в 1930-50 гг. постановок математических задач, происхождение которых связано с анализом статистических данных. Начиная с 70-х годов XX в. исследования по математической статистике посвящены обобщению и дальнейшему математическому изучению этих задач. Поток новых математических результатов (теорем) не ослабевает, но новые практические рекомендации по обработке статистических данных при этом не появляются. Можно сказать, что математическая статистика как научное направление замкнулась внутри себя.

Сам термин «прикладная статистика» возник как реакция на описанную выше тенденцию. Прикладная статистика нацелена на решение реальных задач. Поэтому в ней возникают новые постановки математических задач анализа статистических данных, развиваются и обосновываются новые методы. Обоснование часто проводится математическими методами, т.е. путем доказательства теорем. Большую роль играет методологическая составляющая - как именно ставить задачи, какие предположения принять с целью дальнейшего математического изучения. Велика роль современных информационных технологий, в частности, компьютерного эксперимента.

Рассматриваемое соотношение математической и прикладной статистик отнюдь не являются исключением. Как правило, математические дисциплины проходят в своем развитии ряд этапов. Вначале в какой-либо прикладной области возникает необходимость в применении математических методов и накапливаются соответствующие эмпирические приемы (для геометрии это - "измерение земли", т.е. землемерие, в Древнем Египте). Затем возникает математическая дисциплина со своей аксиоматикой (для геометрии это - время Евклида). Затем идет внутриматематическое развитие и преподавание (считается, что большинство результатов элементарной геометрии получено учителями гимназий в XIX в.). При этом на запросы исходной прикладной области перестают обращать внимание, и та порождает новые научные дисциплины (сейчас "измерением земли" занимается не геометрия, а геодезия и картография). Затем научный интерес к исходной дисциплине иссякает, но преподавание по традиции продолжается (элементарная геометрия до сих пор изучается в средней школе, хотя трудно понять, в каких практических задачах может понадобиться, например, теорема о том, что высоты треугольника пересекаются в одной точке). Следующий этап - окончательное вытеснение дисциплины из реальной жизни в историю науки (объем преподавания элементарной геометрии в настоящее время постепенно сокращается, в частности, ей все меньше уделяется внимания на вступительных экзаменах в вузах). К интеллектуальным дисциплинам, закончившим свой жизненный путь, относится средневековая схоластика. Как справедливо отмечает проф. МГУ им. М.В. Ломоносова В.Н. Тутубалин [16], теория вероятностей и математическая статистика успешно двигаются по ее пути - вслед за элементарной геометрией.

Подведем итог. Хотя статистические данные собираются и анализируются с незапамятных времен (см., например, Книгу Чисел в Ветхом Завете), современная математическая статистика как наука была создана, по общему мнению специалистов, сравнительно недавно - в первой половине XX в. Именно тогда были разработаны основные идеи и получены результаты, излагаемые ныне в учебных курсах математической статистики. После чего специалисты по математической статистике занялись внутриматематическими проблемами, а для теоретического обслуживания проблем практического анализа статистических данных стала формироваться новая дисциплина - прикладная статистика.

В настоящее время статистическая обработка данных проводится, как правило, с помощью соответствующих программных продуктов. Разрыв между математической и прикладной статистикой проявляется, в частности, в том, что большинство методов,

включенных в статистические пакеты программ (например, в заслуженные *Statgraphics* и *SPSS* или в более новую систему *Statistica*), даже не упоминается в учебниках по математической статистике. В результате специалист по математической статистике оказывается зачастую беспомощным при обработке реальных данных, а пакеты программ применяют (что еще хуже - и разрабатывают) лица, не имеющие необходимой теоретической подготовки. Естественно, что они допускают разнообразные ошибки, в том числе в таких ответственных документах, как государственные стандарты по статистическим методам [17].

**Что дает прикладная статистика народному хозяйству?** Так называлась статья [18], в которой приводились многочисленные примеры успешного использования методов прикладной математической статистики при решении практических задач. Перечень примеров можно продолжать практически безгранично (см., например, недавнюю сводку [19]).

Методы прикладной статистики используются в зарубежных и отечественных экономических и технических исследованиях, работах по управлению (менеджменту), в медицине, социологии, психологии, истории, геологии и других областях. Их применение дает заметный экономический эффект. Например, в США - не менее 20 миллиардов долларов ежегодно только в области статистического контроля качества. В 1988 г. затраты на статистический анализ данных в нашей стране оценивались в 2 миллиарда рублей ежегодно [20]. Согласно расчетам сравнительной стоимости валют на основе потребительских паритетов [5], эту величину можно сопоставить с 2 миллиардами долларов США. Следовательно, объем отечественного "рынка статистических услуг" был на порядок меньше, чем в США, что совпадает с оценками и по другим показателям, например, по числу специалистов.

Публикации по новым статистическим методам, по их применениям в технико-экономических исследованиях, в инженерном деле постоянно появляются, например, в журнале "Заводская лаборатория", в секции "Математические методы исследования". Надо назвать также журналы "Автоматика и телемеханика" (издается Институтом проблем управления Российской академии наук), "Экономика и математические методы" (издается Центральным экономико-математическим институтом РАН).

Однако необходимо констатировать, что для большинства менеджеров, экономистов и инженеров прикладная статистика является пока экзотикой. Это объясняется тем, что в вузах современным статистическим методам почти не учат. Во всяком случае, по состоянию на 2003 г. каждый квалифицированный специалист в этой области - самоучка.

Этому выводу не мешает то, что в вузовских программах обычно есть два курса, связанных со статистическими методами. Один из них - "Теория вероятностей и математическая статистика". Этот небольшой курс обычно читают специалисты с математических кафедр. Они успевают дать лишь общее представление об основных понятиях математической статистики. Кроме того, внимание математиков обычно сосредоточено на внутриматематических проблемах, их больше интересует доказательства теорем, а не применение современных статистических методов в задачах экономики и менеджмента. Другой курс - "Статистика" или "Общая теория статистики", входящий в стандартный блок экономических дисциплин. Фактически он является введением в прикладную статистику и содержит первые начала эконометрических методов (по состоянию на 1900 г.).

Прикладная статистика как учебный предмет опирается на два названных вводных курса. Она призвана вооружить специалиста современным статистическим инструментарием. Специалист – это инженер, экономист, менеджер, геолог, медик, социолог, психолог, историк, химик, физик и т.д. Во многих странах мира - Японии и США, Франции и Швейцарии, Перу и Ботсване и др. - статистическим методам обучают в средней школе. ЮНЕСКО постоянно проводят конференции по вопросам такого обучения [21]. В СССР и СЭВ, а теперь - по плохой традиции - и в России игнорируют этот предмет в средней школе и лишь слегка затрагивают его в высшей. Результат на рынке труда очевиден - снижение конкурентоспособности специалистов.

Проблемы прикладной статистики постоянно обсуждаются специалистами. Широкий интерес вызвала дискуссия в журнале «Вестник статистики», в рамках которой были, в частности, опубликованы статьи [9, 18]. На появление в нашей стране прикладной статистики отреагировали и в США [22].

В нашей стране получены многие фундаментальные результаты прикладной статистики. Огромное значение имеют работы академика РАН А.Н. Колмогорова [23]. Во многих случаях именно его работы дали первоначальный толчок дальнейшему развитию ряда направлений прикладной статистики. Зачастую еще 50-70 лет назад А.Н. Колмогоров рассматривал те проблемы, которые только сейчас начинают широко обсуждаться. Как правило, его работы не устарели и сейчас. Свою жизнь посвятили прикладной статистике члены-корреспонденты АН СССР Н.В. Смирнов и Л.Н. Большев. В настоящем учебнике постоянно встречаются ссылки на лучшую публикацию XX в. по прикладной статистике – составленные ими подробно откомментированные «Таблицы …» [24].

**Структура учебника.** Настоящий учебник состоит из четырех основных частей. Первая из них посвящена фундаменту здания современной прикладной статистики. Анализируются различные виды статистических данных - количественных и категоризованных (качественных), нечисловых и нечетких, соответствующих тем или иным шкалам измерения. Современная прикладная статистика позволяет анализировать данные в пространствах произвольной природы, при этом ее математический аппарат опирается на использование расстояний в таких пространствах. Дается представление о введении расстояний с помощью тех или иных систем аксиом.

Современная прикладная статистика основана на использовании вероятностных моделей. Поэтому мы сочли полезным включить в учебник главу 1.2, посвященную основам вероятностно-статистических методов описания неопределенностей в прикладной статистике. Обсуждаются понятия вероятностного пространства, случайной величины, ее распределения и характеристик. Дается представление об основных проблемах прикладной статистики - описании данных, оценивании, проверке гипотез. Следующая глава посвящена выборочным исследованиям. Рассматриваются примеры применения случайных выборок при оценивании функции спроса и изучении предпочтений потребителей.

Ряд результатов теории вероятностей, составляющих теоретическую базу прикладной статистики, приведен в главе 1.4. Рассмотрены законы больших чисел, центральные предельные теоремы, теоремы о наследовании сходимости, метод линеаризации и принцип инвариантности. Показано, что нечеткие множества можно рассматривать как проекции случайных множеств. Обсуждаются проблемы устойчивости статистических выводов.

Основным проблемам прикладной статистики посвящена вторая часть. Начинаем с описания данных. При обсуждении моделей порождения данных, показано, в частности, что распределения реальных данных, как правило, не являются нормальными. Рассмотрено построение таблиц и использование выборочных характеристик. Выбор средних величин увязан со шкалами измерения данных и видом соответствующих инвариантных алгоритмов. В рамках вероятностных моделей порождения нечисловых данных введены эмпирические и теоретические средние в пространствах произвольной природы, для них доказаны законы больших чисел. В прикладной статистике широко используются непараметрические ядерные оценки плотности, в том числе в дискретных пространствах.

Среди методов оценивания параметров предпочтение отдается одношаговым оценкам. Установлено поведение решений экстремальных статистических задач при росте объемов выборок. Эти результаты позволяют установить состоятельность обычно используемых оценок. В рамках теории робастности статистических процедур изучается устойчивость оценок к малым отклонениям от исходных предпосылок.

Завершающая глава второй части посвящена проверке гипотез. Обоснован метод моментов проверки гипотез. Продемонстрирована неустойчивость параметрических методов отбраковки выбросов. Развита предельная теория непараметрических критериев. На основе теории несмещенных оценок разработан метод проверки гипотез по

совокупности малых выборок. Обсуждается проблема множественных проверок статистических гипотез.

В третьей части рассмотрены конкретные методы прикладной статистики, сгруппированные по типу обрабатываемых данных. Статистический анализ числовых величин начинается с оценивания основных характеристик распределения. Затем обсуждаются методы проверки однородности характеристик двух независимых выборок, в том числе двухвыборочный критерий Вилкоксона и состоятельные критерии проверки однородности независимых выборок. Среди различных методов проверки однородности связанных выборок выделяются ориентированные на проверку гипотезы симметрии распределения.

В многомерном статистическом анализе от коэффициентов корреляции переходим к основам линейного регрессионного анализа, рассматриваемым в основном на примере восстановления линейной зависимости между двумя переменными. Уделено внимание основам теории классификации и статистическим методам классификации, методам снижения размерности, индексам и их применению (на примере индекса инфляции).

В следующей главе рассмотрены методы анализа и прогнозирования временных рядов. Внимание уделено оцениванию длины периода и периодической составляющей. Рассмотрен один из наиболее современных методов статистики временных рядов - метод ЖОК оценки результатов взаимовлияний факторов. Обсуждаются вопросы моделирования и анализа многомерных временных рядов, в том числе с учетом балансовых соотношений.

Одно из центральных мест в учебнике занимает статистика нечисловых данных. Рассмотрена структура этой области прикладной статистики. Развиваются теория случайных толерантностей и теория люсианов. Проанализированы метод парных сравнений и статистика нечетких множеств. Обсуждается применение статистики нечисловых данных в теории и практике экспертных оценках.

Заключительная глава третьей части посвящена развитой в течение последних 25 лет статистике интервальных данных. После обсуждения основных идей статистики интервальных данных рассмотрены интервальные варианты основных методов прикладной статистики. Речь идет об оценивании характеристик и параметров распределения, задачах проверки гипотез, линейном регрессионном анализе интервальных данных, интервальном дискриминантном анализе и интервальном кластер-анализе. В качестве примера практического использования разобрано применение статистики интервальных данных для оценки погрешностей характеристик финансовых потоков инвестиционных проектов. Завершается глава обсуждением места статистики интервальных данных в прикладной статистике.

В заключительной четвертой части учебника речь идет об основных проблемах современной прикладной статистики. Выделены «точки роста» этой научно-практической дисциплины. Обсуждаются вопросы развития и внедрения высоких статистических технологий. Рассмотрена роль компьютеров при вероятностно-статистическом моделировании реальных явлений и процессов и их использование при изучении теоретических проблем анализа статистических данных. В конце четвертой части сформулированы основные нерешенные проблемы современной прикладной статистики.

К учебнику даны три приложения. В первом рассмотрены методологические вопросы прикладной статистики. Во втором рассказывается о дискуссии по основным проблемам прикладной статистики, прошедшей в нашей стране в 1980-е годы, и последовавших затем событиях. Для большей объективности отражения дискуссии в качестве приложения 2 использовано изложение статьи [22] в журнале Американской статистической ассоциации. Наконец, в приложении 3 приведены основные сведения о научной и преподавательской деятельности автора настоящего учебника, поясняющие положенные в основу учебника идеи.

Таким образом, настоящий учебник построен на основе обобщения опыта многих специалистов по анализу конкретных технических, экономических, медицинских и иных данных и отражает современное представление о прикладной статистике как самостоятельной научно-практической дисциплине.

## Литература

1. Никитина Е.П., Фрейдлина В.Д., Ярхо А.В. Коллекция определений термина «статистика». – М.: МГУ, 1972. – 46 с.
2. Ленин В.И. Развитие капитализма в России. Процесс образования внутреннего рынка для крупной промышленности. - М.: Политиздат, 1986. - XII, 610 с.
3. Гнеденко Б.В. Очерк по истории теории вероятностей. – М.:УРСС, 2001. – 88 с.
4. Клейн Ф. Лекции о развитии математики в XIX столетии. Часть I. - М.-Л.: Объединенное научно-техническое издательство НКТП СССР, 1937. - 432 с.
5. Плошко Б.Г., Елисеева И.И. История статистики: Учеб. пособие. - М.: Финансы и статистика. 1990. - 295 с.
6. Орлов А.И. Эконометрика. Учебник для вузов. Изд. 2-е, исправленное и дополненное. - М.: Изд-во "Экзамен", 2003. – 576 с.
7. Бернштейн С.Н. Современное состояние теории вероятностей и ее приложений. - В сб.: Труды Всероссийского съезда математиков в Москве 27 апреля - 4 мая 1927 г. - М.-Л.: ГИЗ, 1928. С.50-63.
8. Орлов А.И. О современных проблемах внедрения прикладной статистики и других статистических методов. / Заводская лаборатория. 1992. Т.58. № 1. С.67-74.
9. Орлов А.И. О перестройке статистической науки и её применений. / Вестник статистики. 1990. № 1. С.65 - 71.
10. Кендалл М., Стьюарт А. Теория распределений. - М.: Наука, 1966. - 566 с.
11. Кендалл М., Стьюарт А. Статистические выводы и связи. - М.: Наука, 1973. - 899 с.
12. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. - М.: Наука, 1976. - 736 с.
13. Налимов В.В., Мульченко З.М. Наукометрия. Изучение развития науки как информационного процесса. - М.: Наука, 1969. - 192 с.
14. ГОСТ 11.011-83. Прикладная статистика. Правила определения оценок и доверительных границ для параметров гамма-распределения. - М.: Изд-во стандартов. 1984. - 53 с.
15. Орлов А.И. О развитии прикладной статистики. - В сб.: Современные проблемы кибернетики (прикладная статистика). - М.: Знание, 1981, с.3-14.
16. Тутубалин В.Н. Границы применимости (вероятностно-статистические методы и их возможности). - М.: Знание, 1977. - 64 с.
17. Орлов А.И. Сертификация и статистические методы. - Журнал "Заводская лаборатория". 1997. Т.63. № 3. С.55-62.
18. Орлов А.И. Что дает прикладная статистика народному хозяйству? – Журнал «Вестник статистики». 1986, No.8. С.52 – 56.
19. Орлов А.И., Орлова Л.А. Применение эконометрических методов при решении задач контроллинга. – Журнал «Контроллинг». 2003. №4.
20. Комаров Д.М., Орлов А.И. Роль методологических исследований в разработке методоориентированных экспертных систем (на примере оптимизационных и статистических методов). - В сб.: Вопросы применения экспертных систем. - Минск: Центросистем, 1988. С.151-160.
21. The teaching of statistics / Studies in mathematical education, vol.7. - Paris, UNESCO, 1991. - 258 pp.
22. Котц С., Смит К. Пространство Хаусдорфа и прикладная статистика: точка зрения ученых СССР. - The American Statistician. November 1988. Vol. 42. № 4. P. 241-244.
23. Кудлаев Э.М., Орлов А.И. Вероятностно-статистические методы исследования в работах А.Н.Колмогорова. – Журнал «Заводская лаборатория». 2003. Т.69. № 5. С.55-61.
24. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1965 (1-е изд.), 1968 (2-е изд.), 1983 (3-е изд.).

# Часть 1. Фундамент прикладной статистики

## 1.1. Различные виды статистических данных

### 1.1.1. Количественные и категоризованные данные

Методы прикладной статистики – это методы анализа данных, причем обычно достаточно большого количества данных. Статистические данные могут иметь различную природу. Исторически самыми ранними были два вида данных – сведения о числе объектов, удовлетворяющих тем или иным условиям, и числовые результаты измерений.

Первый из этих видов данных до сих пор главенствует в статистических сборниках Госкомстата РФ. Такого рода данные часто называют *категоризованными*, поскольку о каждом из рассматриваемых объектов известно, в какую из нескольких заранее заданных категорий он попадает. Примером является информация Госкомстата РФ о населении страны, с разделением по возрастным категориям и полу. Часто при составлении таблиц жертвуют информацией, заменяя точное значение измеряемой величины на указание интервала группировки, в которую это значение попадает. Например, вместо точного возраста человека используют лишь один из указанных в таблице возрастных интервалов.

Второй наиболее распространенный вид данных – количественные данные, рассматриваемые как действительные числа. Таковы результаты измерений, наблюдений, испытаний, опытов, анализов. Количественные данные обычно описываются набором чисел (выборкой), а не таблицей.

Нельзя утверждать, что категоризованные данные соответствуют первому этапу исследования, а числовые – следующему, на котором используются более совершенные методы измерения. Дело в том, что человеку свойственно давать качественные ответы на возникающие в его практической деятельности вопросы. Примером является используемая А.А. Пивнем таблица сильных и слабых сторон внутренней среды Компании (табл.1).

Таблица 1

Оценка сильных и слабых сторон внутренней среды Компании

| Показатели Компании | Оценка показателя (По отношению к предприятиям отрасли) | | | | | Важность (вес) | | |
|---|---|---|---|---|---|---|---|---|
| | Очень высо-кая | Выс о-кая | Средн яя | Низк ая | Очень низ-кая | Вы-со-кая | Средн яя | Низкая |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| **Финансы** | | | | | | | | |
| 1.Оценка структуры активов | | | X | | | X | | |
| 2.Инвестиционная привлекательность | | | X | | | X | | |
| 3.Доход на активы | | | | X | | X | | |
| 4.Норма прибыли | | | | | X | X | | |
| 5.Доход на вложенный капитал | | | | X | | | X | |
| **Производство** | | | | | | | | |
| 1.Использования оборудования | | | X | | | | X | |
| 2.Производственные мощности | | | X | | | | | X |
| 3.Численность | | | X | | | | X | |
| 4.Система контроля качества | | X | | | | X | | |
| 5.Возможность расширения производства | | | X | | | X | | |
| 6.Износ оборудования | | | | X | | X | | |
| **Организация и управление** | | | | | | | | |
| 1.Численность ИТР и управленческого персонала | | | X | | | X | | |
| 2.Скорость реакции управления на изменения во внешней среде | | | X | | | X | | |

| Показатели Компании | Оценка показателя (По отношению к предприятиям отрасли) | | | | | Важность (вес) | | |
|---|---|---|---|---|---|---|---|---|
| | Очень высо-кая | Выс о-кая | Средн яя | Низк ая | Очень низ-кая | Вы-со-кая | Средн яя | Низкая |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 3.Четкость разделения полномочий и функций | | | | X | | | X | |
| 4.Качество используемой в управлении информации | | | X | | | X | | |
| 5.Гибкость оргструктуры управления | | X | | | | X | | |
| **Маркетинг** | | | | | | | | |
| 1.Доля рынка | | X | | | | X | | |
| 2.Репутация Компании | | X | | | | X | | |
| 3.Престиж торговой марки | | | X | | | | X | |
| 4.Стимулирование сбыта | | X | | | | X | | |
| 5.Численность сбытового персонала | | | | X | | | | X |
| 6.Уровень цен | | | X | | | X | | |
| 7.Уровень сервиса | | X | | | | X | | |
| 8.Число клиентов | | X | | | | | X | |
| 9.Качество поступающей информации | | | X | | | | X | |
| **Кадровый состав** | | | | | | | | |
| 1.Уровень квалификации производственного персонала | | X | | | | X | | |
| 2.Расходы по подготовке и переподготовке персонала | | X | | | | X | | |
| 3.Уровень подготовки сбытового персонала в технической области | | | | X | | | X | |
| **Технология** | | | | | | | | |
| 1.Применяемые стандарты | | X | | | | | | X |
| 2.Новые продукты | | | X | | | | X | |
| 3.Расходы на НИОКР | | X | | | | | X | |

Ясно, что вполне можно превратить в числа значения признаков, названия которых приведены в столбце «Показатели Компании», однако этот переход будет зависеть от исследователя, носить неизбежный налет субъективизма.

Иногда не удается однозначно отнести данные к категоризованным или количественным. Например, в Ветхом Завете, в Четвертой книге Моисеева «Числа» указывается количество воинов в различных коленах. С одной стороны, это типичные категоризованные данные, градациями служат названия колен. С другой стороны, эти данные можно рассматривать как количественные, как выборку, их вполне естественно складывать, вычислять среднее арифметическое и т.п.

Описанная ситуация типична. Существует весьма много различных видов статистических данных. Это связано, в частности, со способами их получения. Например, если испытания некоторых технических устройств продолжаются до определенного момента, то получаем т.н. *цензурированные* данные, состоящие из набора чисел – продолжительности работы ряда устройств до отказа, и информации о том, что остальные устройства продолжали работать в момент окончания испытания. Такого рода данные часто используются при оценке и контроле надежности технических устройств.

Описание вида данных и, при необходимости, механизма их порождения – начало любого статистического исследования.

В простейшем случае статистические данные – это значения некоторого признака, свойственного изучаемым объектам. Значения могут быть количественными или представлять собой указание на категорию, к которой можно отнести объект. Во втором случае говорят о

качественном признаке. Используют и более сложные признаки, перечень которых будет расширяться по мере развертывания изложения в учебнике.

При измерении по нескольким количественным или качественным признакам в качестве статистических данных об объекте получаем вектор. Его можно рассматривать как новый вид данных. В таком случае выборка состоит из набора векторов. Есть часть координат – числа, а часть – качественные (категоризованные) данные, то говорим о векторе разнотипных данных.

Одним элементом выборки, т.е. одним измерением, может быть и функция в целом. Например, электрокардиограмма больного или амплитуда биений вала двигателя. Или временной ряд, описывающий динамику показателей определенной фирмы. Тогда выборка состоит из набора функций.

Элементами выборки могут быть и бинарные отношения. Например, при опросах экспертов часто используют упорядочения (ранжировки) объектов экспертизы – образцов продукции, инвестиционных проектов, вариантов управленческих решений. В зависимости от регламента экспертного исследования элементами выборки могут быть различные виды бинарных отношений (упорядочения, разбиения, толерантности), множества, нечеткие множества и т.д.

Итак, математическая природы элементов выборки в различных задачах прикладной статистики может быть самой разной. Однако можно выделить два класса статистических данных – числовые и нечисловые. Соответственно прикладная статистика разбивается на две части – числовую статистику и нечисловую статистику.

Числовые статистические данные – это числа, вектора, функции. Их можно складывать, умножать на коэффициенты. Поэтому в числовой статистике большое значение имеют разнообразные суммы. Математический аппарат анализа сумм случайных элементов выборки – это (классические) законы больших чисел и центральные предельные теоремы (см. главу 1.3).

Нечисловые статистические данные – это категоризованные данные, вектора разнотипных признаков, бинарные отношения, множества, нечеткие множества и др. Их нельзя складывать и умножать на коэффициенты. Поэтому не имеет смысла говорить о суммах нечисловых статистических данных. Они являются элементами нечисловых математических пространств (множеств). Математический аппарат анализа нечисловых статистических данных основан на использовании расстояний между элементами (а также мер близости, показателей различия) в таких пространствах. С помощью расстояний определяются эмпирические и теоретические средние, доказываются законы больших чисел, строятся непараметрические оценки плотности распределения вероятностей, решаются задачи диагностики и кластерного анализа, и т.д. (см. главу 3.4).

Сведем информацию об основных областях прикладной статистики в табл.2. Отметим, что модели порождения цензурированных данных входят в состав каждой из рассматриваемых областей.

Таблица 2.

Области прикладной статистики

| № п/п | Вид статистических данных | Область прикладной статистики |
|---|---|---|
| 1 | Числа | Статистика (случайных) величин |
| 2 | Конечномерные вектора | Многомерный статистический анализ |
| 3 | Функции | Статистика случайных процессов и временных рядов |
| 4 | Объекты нечисловой природы | Статистика нечисловых данных (статистика объектов нечисловой природы) |

### 1.1.2. Основные шкалы измерения

**Почему необходима теория измерений?** Теория измерений (в дальнейшем сокращенно ТИ) является одной из составных частей прикладной статистики. Она входит в состав *статистики объектов нечисловой природы*.

Использование чисел в жизни и хозяйственной деятельности людей отнюдь не всегда предполагает, что эти числа можно складывать и умножать, производить иные арифметические действия. Что бы вы сказали о человеке, который занимается умножением телефонных номеров? И отнюдь не всегда 2+2=4. Если вы вечером поместите в клетку двух животных, а потом еще двух, то отнюдь не всегда можно утром найти в этой клетке четырех животных. Их может быть и много больше - если вечером вы загнали в клетку овцематок или беременных кошек. Их может быть и меньше - если к двум волкам вы поместили двух ягнят. Числа используются гораздо шире, чем арифметика.

Так, например, мнения экспертов часто выражены в *порядковой шкале* (подробнее о шкалах говорится ниже), т.е. эксперт может сказать (и обосновать), что один показатель качества продукции более важен, чем другой, первый технологический объект более опасен, чем второй, и т.д. Но он не в состоянии сказать, *во сколько раз* или *на сколько* более важен, соответственно, более опасен. Экспертов часто просят дать ранжировку (упорядочение) объектов экспертизы, т.е. расположить их в порядке возрастания (или убывания) интенсивности интересующей организаторов экспертизы характеристики. Ранг - это номер (объекта экспертизы) в упорядоченном ряду значений характеристики у различных объектов. Такой ряд в статистике называется вариационным. Формально ранги выражаются числами 1, 2, 3, ..., но с этими числами нельзя делать привычные арифметические операции. Например, хотя в арифметике 1 + 2 = 3, но нельзя утверждать, что для объекта, стоящем на третьем месте в упорядочении, интенсивность изучаемой характеристики равна сумме интенсивностей объектов с рангами 1 и 2. Так, один из видов экспертного оценивания - оценки учащихся. Вряд ли кто-либо будет утверждать, что знания отличника равны сумме знаний двоечника и троечника (хотя 5 = 2 + 3), хорошист соответствует двум двоечникам (2 + 2 = 4), а между отличником и троечником такая же разница, как между хорошистом и двоечником (5 - 3 = 4 - 2). Поэтому очевидно, что для анализа подобного рода качественных данных необходима не всем известная арифметика, а другая теория, дающая базу для разработки, изучения и применения конкретных методов расчета. Это и есть ТИ.

При чтении литературы надо иметь в виду, что в настоящее время термин "теория измерений" применяется для обозначения целого ряда научных дисциплин. А именно, классической метрологии (науки об измерениях физических величин), рассматриваемой здесь ТИ, некоторых других направлений, например, алгоритмической теории измерений. Обычно из контекста понятно, о какой конкретно теории идет речь.

**Краткая история теории измерений.** Сначала ТИ развивалась как теория психофизических измерений. В послевоенных публикациях американский психолог С.С. Стивенс основное внимание уделял шкалам измерения. Во второй половине XX в. сфера применения ТИ стремительно расширяется. Посмотрим, как это происходило. Один из томов выпущенной в США в 1950-х годах "Энциклопедии психологических наук" назывался "Психологические измерения". Значит, составители этого тома расширили сферу применения РТИ с психофизики на психологию в целом. А в основной статье в этом сборнике под названием, обратите внимание, "Основы теории измерений", изложение шло на абстрактно-математическом уровне, без привязки к какой-либо конкретной области применения. В этой статье [1] упор был сделан на "гомоморфизмах эмпирических систем с отношениями в числовые" (в эти математические термины здесь вдаваться нет необходимости), и математическая сложность изложения возросла по сравнению с работами С.С. Стивенса.

Уже в одной из первых отечественных статей по РТИ (конец 1960-х годов) было установлено, что баллы, присваиваемые экспертами при оценке объектов экспертизы, как правило, измерены в порядковой шкале. Отечественные работы, появившиеся в начале 1970-х годов, привели к существенному расширению области использования РТИ. Ее применяли к педагогической квалиметрии (измерению качества знаний учащихся), в системных

исследованиях, в различных задачах теории экспертных оценок, для агрегирования показателей качества продукции, в социологических исследованиях, и др.

Итоги этого этапа были подведены в монографии [2]. В качестве двух основных проблем РТИ наряду с *установлением типа шкалы* измерения конкретных данных был выдвинут поиск алгоритмов анализа данных, результат работы которых не меняется при любом допустимом преобразовании шкалы (т.е. является *инвариантным* относительно этого преобразования).

Метрологи вначале резко возражали против использования термина "измерение" для качественных признаков. Однако постепенно возражения сошли на нет, и к концу XX в. ТИ стала рассматриваться как общенаучная теория.

**Шесть типов шкал.** В соответствии с ТИ при математическом моделировании реального явления или процесса следует прежде всего установить *типы шкал*, в которых измерены те или иные переменные. Тип шкалы задает *группу допустимых преобразований шкалы*. Допустимые преобразования не меняют соотношений между объектами измерения. Например, при измерении длины переход от аршин к метрам не меняет соотношений между длинами рассматриваемых объектов - если первый объект длиннее второго, то это будет установлено и при измерении в аршинах, и при измерении в метрах. Обратите внимание, что при этом численное значение длины в аршинах отличается от численного значения длины в метрах - не меняется лишь результат сравнения длин двух объектов.

Укажем основные виды шкал измерения и соответствующие группы допустимых преобразований.

В *шкале наименований* (другое название этой шкалы - *номинальная;* это - переписанное русскими буквами английское название шкалы) **допустимыми** являются все взаимно-однозначные преобразования. В этой шкале числа используются лишь как метки. Примерно так же, как при сдаче белья в прачечную, т.е. лишь для различения объектов. В шкале наименований измерены, например, номера телефонов, автомашин, паспортов, студенческих билетов. Номера страховых свидетельств государственного пенсионного страхования, медицинского страхования, ИНН (индивидуальный номер налогоплательщика) измерены в шкале наименований. Пол людей тоже измерен в шкале наименований, результат измерения принимает два значения - мужской, женский. Раса, национальность, цвет глаз, волос - номинальные признаки. Номера букв в алфавите - тоже измерения в шкале наименований. Никому в здравом уме не придет в голову складывать или умножать номера телефонов, такие операции не имеют смысла. Сравнивать буквы и говорить, например, что буква П лучше буквы С, также никто не будет. Единственное, для чего годятся измерения в шкале наименований - это различать объекты. Во многих случаях только это от них и требуется. Например, шкафчики в раздевалках для взрослых различают по номерам, т.е. числам, а в детских садах используют рисунки, поскольку дети еще не знают чисел.

В *порядковой шкале* числа используются не только для различения объектов, но и для установления порядка между объектами. Простейшим примером являются оценки знаний учащихся. Символично, что в средней школе применяются оценки 2, 3, 4, 5, а в высшей школе ровно тот же смысл выражается словесно - неудовлетворительно, удовлетворительно, хорошо, отлично. Этим подчеркивается "нечисловой" характер оценок знаний учащихся. В порядковой шкале **допустимыми** являются все строго возрастающие преобразования.

Установление типа шкалы, т.е. задания группы допустимых преобразований шкалы измерения - дело специалистов соответствующей прикладной области. Так, оценки привлекательности профессий мы в монографии [2], выступая в качестве социологов, считали измеренными в порядковой шкале. Однако отдельные социологи не соглашались с нами, полагая, что выпускники школ пользуются шкалой с более узкой группой допустимых преобразований, например, интервальной шкалой. Очевидно, эта проблема относится не к математике, а к наукам о человеке. Для ее решения может быть поставлен достаточно трудоемкий эксперимент. Пока же он не поставлен, целесообразно принимать порядковую шкалу, так как это гарантирует от возможных ошибок.

Оценки экспертов, как уже отмечалось, часто следует считать измеренными в порядковой шкале. Типичным примером являются задачи ранжирования и классификации промышленных объектов, подлежащих экологическому страхованию.

Почему мнения экспертов естественно выражать именно в порядковой шкале? **Как показали многочисленные опыты, человек более правильно (и с меньшими затруднениями) отвечает на вопросы качественного, например, сравнительного, характера, чем количественного.** Так, ему легче сказать, какая из двух гирь тяжелее, чем указать их примерный вес в граммах.

В различных областях человеческой деятельности применяется много других видов порядковых шкал. Так, например, в минералогии используется шкала Мооса, по которому минералы классифицируются согласно критерию твердости. А именно: тальк имеет балл 1, гипс - 2, кальций - 3, флюорит - 4, апатит - 5, ортоклаз - 6, кварц - 7, топаз - 8, корунд - 9, алмаз - 10. Минерал с большим номером является более твердым, чем минерал с меньшим номером, при нажатии царапает его.

Порядковыми шкалами в географии являются - бофортова шкала ветров ("штиль", "слабый ветер", "умеренный ветер" и т.д.), шкала силы землетрясений. Очевидно, нельзя утверждать, что землетрясение в 2 балла (лампа качнулась под потолком - такое бывает и в Москве) ровно в 5 раз слабее, чем землетрясение в 10 баллов (полное разрушение всего на поверхности земли).

В медицине порядковыми шкалами являются - шкала стадий гипертонической болезни (по Мясникову), шкала степеней сердечной недостаточности (по Стражеско-Василенко-Лангу), шкала степени выраженности коронарной недостаточности (по Фогельсону), и т.д. Все эти шкалы построены по схеме: заболевание не обнаружено; первая стадия заболевания; вторая стадия; третья стадия… Иногда выделяют стадии 1а, 1б и др. Каждая стадия имеет свойственную только ей медицинскую характеристику. При описании групп инвалидности числа используются в противоположном порядке: самая тяжелая - первая группа инвалидности, затем - вторая, самая легкая - третья.

Номера домов также измерены в порядковой шкале - они показывают, в каком порядке стоят дома вдоль улицы. Номера томов в собрании сочинений писателя или номера дел в архиве предприятия обычно связаны с хронологическим порядком их создания.

При оценке качества продукции и услуг, в т.н. квалиметрии (буквальный перевод: измерение качества) популярны порядковые шкалы. А именно, единица продукции оценивается как годная или не годная. При более тщательном анализе используется шкала с тремя градациями: есть значительные дефекты - присутствуют только незначительные дефекты - нет дефектов. Иногда применяют четыре градации: имеются критические дефекты (делающие невозможным использование) - есть значительные дефекты - присутствуют только незначительные дефекты - нет дефектов. Аналогичный смысл имеет сортность продукции - высший сорт, первый сорт, второй сорт,…

При оценке экологических воздействий первая, наиболее обобщенная оценка - обычно порядковая, например: природная среда стабильна - природная среда угнетена (деградирует). Аналогично в эколого-медицинской шкале: нет выраженного воздействия на здоровье людей - отмечается отрицательное воздействие на здоровье.

Порядковая шкала используется и во многих иных областях. В эконометрике это прежде всего различные методы экспертных оценок. (см. посвященный им материал в части 3).

Все шкалы измерения делят на две группы - шкалы качественных признаков и шкалы количественных признаков.

**Порядковая шкала и шкала наименований - основные шкалы качественных признаков**. Поэтому во многих конкретных областях результаты качественного анализа можно рассматривать как измерения по этим шкалам.

**Шкалы количественных признаков - это шкалы интервалов, отношений, разностей, абсолютная**. По шкале *интервалов* измеряют величину потенциальной энергии или координату точки на прямой. В этих случаях на шкале нельзя отметить ни естественное начало отсчета, ни естественную единицу измерения. Исследователь должен сам задать точку отсчета и сам выбрать единицу измерения. Допустимыми преобразованиями в шкале интервалов

являются линейные возрастающие преобразования, т.е. линейные функции. Температурные шкалы Цельсия и Фаренгейта связаны именно такой зависимостью: ${}^0C = 5/9\ ({}^0F - 32)$, где ${}^0C$ - температура (в градусах) по шкале Цельсия, а ${}^0F$ - температура по шкале Фаренгейта.

Из количественных шкал наиболее распространенными в науке и практике являются шкалы *отношений.* В них есть естественное начало отсчета - нуль, т.е. отсутствие величины, но нет естественной единицы измерения. По шкале отношений измерены большинство физических единиц: масса тела, длина, заряд, а также цены в экономике. Допустимыми преобразованиями шкале отношений являются подобные (изменяющие только масштаб). Другими словами, линейные возрастающие преобразования без свободного члена. Примером является пересчет цен из одной валюты в другую по фиксированному курсу. Предположим, мы сравниваем экономическую эффективность двух инвестиционных проектов, используя цены в рублях. Пусть первый проект оказался лучше второго. Теперь перейдем на валюту самой экономически мощной державы мира - юани, используя фиксированный курс пересчета. Очевидно, первый проект должен опять оказаться более выгодным, чем второй. Это очевидно из общих соображений. Однако алгоритмы расчета не обеспечивают автоматически выполнения этого очевидного условия. Надо проверять, что оно выполнено. Результаты подобной проверки для средних величин описаны ниже (раздел 2.1.3).

В шкале разностей есть естественная единица измерения, но нет естественного начала отсчета. Время измеряется по шкале *разностей*, если год (или сутки - от полудня до полудня) принимаем естественной единицей измерения, и по шкале интервалов в общем случае. На современном уровне знаний естественного начала отсчета указать нельзя. Дату сотворения мира различные авторы рассчитывают по-разному, равно как и момент рождества Христова. Так, согласно новой статистической хронологии [3], разработанной группой известного историка акад. РАН А.Т.Фоменко, Господь Иисус Христос родился примерно в 1054 г. по принятому ныне летоисчислению в Стамбуле (он же - Царьград, Византия, Троя, Иерусалим, Рим).

Только для *абсолютной* шкалы результаты измерений - числа в обычном смысле слова. Примером является число людей в комнате. Для абсолютной шкалы допустимым является только тождественное преобразование.

В процессе развития соответствующей области знания тип шкалы может меняться. Так, сначала температура измерялась по *порядковой* шкале (холоднее - теплее). Затем - по *интервальной* (шкалы Цельсия, Фаренгейта, Реомюра). Наконец, после открытия абсолютного нуля температуру можно считать измеренной по шкале *отношений* (шкала Кельвина). Надо отметить, что среди специалистов иногда имеются разногласия по поводу того, по каким шкалам следует считать измеренными те или иные реальные величины. Другими словами, процесс измерения включает в себя и определение типа шкалы (вместе с обоснованием выбора определенного типа шкалы). Кроме перечисленных шести основных типов шкал, иногда используют и иные шкалы.

Обсуждение шкал измерения будет продолжено далее в более широком контексте – как одного из понятий статистики нечисловых данных.

### 1.1.3. Нечисловые данные

Статистика нечисловых данных - это направление в прикладной статистике, в котором в качестве исходных статистических данных (результатов наблюдений) рассматриваются объекты нечисловой природы. Так принято называть объекты, которые нецелесообразно описывать числами, в частности элементы нелинейных пространств. Примерами являются бинарные отношения (ранжировки, разбиения, толерантности и др.), результаты парных и множественных сравнений, множества, нечеткие множества, измерение в шкалах, отличных от абсолютных. Этот перечень примеров не претендует на законченность. Он складывался постепенно, по мере того, как развивались теоретические исследования в области статистики нечисловых данных и расширялся опыт применений этого направления прикладной статистики.

Объекты нечисловой природы широко используются в теоретических и прикладных исследованиях по экономике, менеджменту и другим проблемам управления, в частности управления качеством продукции, в технических науках, социологии, психологии, медицине и т.д., а также практически во всех отраслях народного хозяйства.

Начнем с первоначального знакомства с основными видами объектов нечисловой природы.

**Результаты измерений в шкалах, отличных от абсолютной.** Рассмотрим подробнее, чем раньше, конкретное исследование в области маркетинга образовательных услуг, послужившее поводом к развитию отечественных исследований по теории измерений. При изучении привлекательности различных профессий для выпускников новосибирских школ был составлен список из 30 профессий. Опрашиваемых просили оценить каждую из этих профессий одним из баллов 1,2,...,10 по правилу: чем больше нравится, тем выше балл. Для получения социологических выводов необходимо было дать единую оценку привлекательности определенной профессии для совокупности выпускников школ. В качестве такой оценки в работе [4] использовалось среднее арифметическое баллов, выставленных профессии опрошенными школьниками. В частности, физика получила средний балл 7,69, а математика - 7,50. Поскольку 7,69 больше, чем 7,50, был сделан вывод, что физика более предпочтительна для школьников, чем математика.

Однако этот вывод противоречит данным работы [5], согласно которым ленинградские школьники средних классов больше любят математику, чем физику. Обсудим одно из возможных объяснений этого противоречия, которое сводится к указанию на неадекватность (с точки зрения теории измерений) методики обработки эконометрических данных, примененной в работе [4].

Дело в том, что баллы 1,2,...,10 введены конкретными исследователями, т.е. субъективно. Если одна профессия оценена в 10 баллов, а вторая - в 2, то из этого нельзя заключить, что первая ровно в 5 раз привлекательней второй. Другой коллектив социологов мог бы принять иную систему баллов, например 1,4,9,16,...,100. Естественно предположить , что упорядочивание профессий по привлекательности, присущее школьникам, не зависит от того, какой системой баллов им предложит пользоваться маркетолог. Раз так, то распределение профессий по градациям десятибалльной системы не изменится, если перейти к другой системе баллов с помощью любого допустимого преобразования в порядковой шкале, т.е. с помощью строго возрастающей функции $g : R^1 \to R^1$. Если , $Y_1, Y_2,...,Y_n$ -ответы $n$ выпускников школ, касающихся математики, а $Z_1, Z_2,...,Z_n$ -физики, то после перехода к новой системе баллов ответы относительно математики будут иметь вид $g(Y_1), g(Y_2),...,g(Y_n)$, а относительно физики - $g(Z_1), g(Z_2),...,g(Z_n)$.

Пусть единая оценка привлекательности профессии вычисляется с помощью функции $f(X_1, X_2,...,X_n)$. Какие требования естественно наложить на функцию $f : R^n \to R^1$, чтобы полученные с ее помощью выводы не зависели от того, какой именно системой баллов пользовался специалист по маркетингу образовательных услуг?

***Замечание.*** Обсуждение можно вести в терминах экспертных оценок. Тогда вместо сравнения математики и физики $n$ экспертов (а не выпускников школ) оценивают по конкурентоспособности на мировом рынке, например, две марки стали. Однако в настоящее время маркетинговые и социологические исследования более привычны, чем экспертные.

Единая оценка вычислялась для того, чтобы сравнивать профессии по привлекательности. Пусть $f(X_1, X_2,...,X_n)$ - среднее по Коши. Пусть среднее по первой совокупности меньше среднего по второй совокупности:

$$f(Y_1, Y_2,...,Y_n) < f(Z_1, Z_2,...,Z_n).$$

Тогда согласно теории измерений необходимо потребовать, чтобы для любого допустимого преобразования $g$ из группы допустимых преобразований в порядковой шкале было справедливо также неравенство

$$f(g(Y_1), g(Y_2),...,g(Y_n)) < f(g(Z_1), g(Z_2),...,g(Z_n)).$$

т.е. среднее преобразованных значений из первой совокупности также было меньше среднего преобразованных значений для второй совокупности. Причем сформулированное условие

должно быть верно для любых двух совокупностей $Y_1, Y_2,...,Y_n$ и $Z_1, Z_2,...,Z_n$ и, напомним, любого допустимого преобразования. Средние величины, удовлетворяющие сформулированному условию, называют допустимыми (в порядковой шкале). Согласно теории измерений только такими средними можно пользоваться при анализе мнений выпускников школ, экспертов и иных данных, измеренных в порядковой шкале.

Какие единые оценки привлекательности профессий $f(X_1, X_2,...,X_n)$ устойчивы относительно сравнения? Ответ на этот вопрос дается ниже в главе 2.1. В частности, оказалось, что средним арифметическим, как в работе [4] новосибирских специалистов по маркетингу образовательных услуг, пользоваться нельзя, а порядковыми статистиками, т.е. членами вариационного ряда (и только ими) - можно.

Методы анализа конкретных экономических данных, измеренных в шкалах, отличных от абсолютной, являются предметом изучения в статистике нечисловых данных как части эконометрики. Как известно, основные шкалы измерения делятся на качественные (шкалы наименований и порядка) и количественные (шкалы интервалов, отношений, разностей, абсолютная). Методы анализа статистических данных в количественных шкалах сравнительно мало отличаются от таковых в абсолютной шкале. Добавляется только требование инвариантности относительно преобразований сдвига и/или масштаба. Методы анализа качественных данных - принципиально иные.

Напомним, что исходным понятием теории измерений является совокупность $\Phi = \{\varphi\}$ допустимых преобразований шкалы (обычно Ф- группа), $\varphi : R^1 \to R^1$. Алгоритм обработки данных $W$, т.е. функция $W : R^n \to A$ (здесь $A$-множество возможных результатов работы алгоритма) называется адекватным в шкале с совокупностью допустимых преобразований Ф, если

$$W(x_1, x_2,..., x_n) = W(\varphi(x_1), \varphi(x_2),..., \phi(x_n))$$

для всех $x_i \in R^1, i = 1,2,...,n,$ и всех $\varphi \in \Phi.$ Таким образом, теорию измерений рассматриваем как теорию инвариантов относительно различных совокупностей допустимых преобразований Ф. Интерес вызывают две задачи:

а) дана группа допустимых преобразований Ф (т.е. задана шкала); какие алгоритмы анализа данных $W$ из определенного класса являются адекватными?

б) дан алгоритм анализа данных $W$; для каких шкал (т.е. групп допустимых преобразований Ф) он является адекватным?

В главе 2.1 первая задача рассматривается для алгоритмов расчета средних величин. Информацию о других результатах решения задач указанных типов можно найти в работах [2,6,7].

**Бинарные отношения.** Пусть $W : R^n \to A$ - адекватный алгоритм в шкале наименований. Можно показать, что этот алгоритм задается некоторой функцией от матрицы $B = \| b_{ij} \| = B(x_1, x_2,..., x_n),$ где

$$b_{ij} = \begin{cases} 1, & x_i = x_j, i, j = 1,2,...,n, \\ 0, & x_i \neq x_j, i, j = 1,2,...,n. \end{cases}$$

Если $W : R^n \to A$ - адекватный алгоритм в шкале порядка, то этот алгоритм задается некоторой функцией от матрицы $C = \| c_{ij} \| = C(x_1, x_2,..., x_n)$ порядка $n \times n$, где

$$c_{ij} = \begin{cases} 1, & x_i \le x_j, i, j = 1,2,...,n, \\ 0, & x_i > x_j, i, j = 1,2,...,n. \end{cases}$$

Матрицы $B$ и $C$ можно проинтерпретировать в терминах бинарных отношений. Пусть некоторая характеристика измеряется у $n$ объектов $q_1, q_2, ..., q_n$, причем $x_i$ - результат ее измерения у объекта $q_i$ Тогда матрицы $B$ и $C$ задают бинарные отношения на множестве объектов $Q = \{q_1, q_2, ..., q_n\}$. Поскольку бинарное отношение можно рассматривать как подмножество декартова квадрата $Q \times Q$, то любой матрице $D = ||d_{ij}||$ порядка $n \times n$ из 0 и 1

соответствует бинарное отношение *R(D),* определяемое следующим образом: $(q_i, q_j) \in R(D)$ тогда и только тогда, когда $d_{ij} = 1$.

Бинарное отношение *R(B)* - отношение эквивалентности, т.е. симметричное рефлексивное транзитивное отношение. Оно задает разбиение *Q* на классы эквивалентности. Два объекта $q_i$ и $q_j$ входят в один класс эквивалентности тогда и только тогда, когда $x_i = x_j, b_{ij} = 1$.

Выше показано, как разбиения возникают в результате измерений в шкале наименований. Разбиения могут появляться и непосредственно. Так, при оценке качества промышленной продукции эксперты дают разбиение показателей качества на группы. Для изучения психологического состояния людей их просят разбить предъявленные рисунки на группы сходных между собой. Аналогичная методика применяется и в иных экспериментальных психологических исследованиях, необходимых для оптимизации управления персоналом.

Во многих эконометрических задачах разбиения получаются "на выходе" (например, в кластерном анализе) или же используются на промежуточных этапах анализа данных (например, сначала проводят классификацию с целью выделения однородных групп, а затем в каждой группе строят регрессионную зависимость).

Бинарное отношение *R(C)* задает разбиение *Q* на классы эквивалентности, между которыми введено отношение строгого порядка. Два объекта $q_i$ и $q_j$ входят в один класс тогда и только тогда, когда $c_{ij} = 1$ и $c_{ji} = 1$, т.е. $x_i = x_j$. Класс эквивалентности $Q_1$ предшествует классу эквивалентности $Q_2$ тогда и только тогда, когда для любых $q_i \in Q_1, q_j \in Q_2$ имеем $c_{ij} = 1$, $c_{ji} = 0$, т.е. $x_i < x_j$. Такое бинарное отношение в статистике часто называют ранжировкой со связями; связанными считаются объекты, входящие в один класс эквивалентности. В литературе встречаются и другие названия: линейный квазипорядок, упорядочение, квазисерия, ранжирование. Если каждый из классов эквивалентности состоит только из одного элемента, то имеем обычную ранжировку (другими словами, линейный порядок).

Как известно, ранжировки возникают в результате измерений в порядковой шкале. Так, при описанном выше опросе ответ выпускника школы - это ранжировка (со связями) профессий по привлекательности. Ранжировки часто возникают и непосредственно, без промежуточного этапа - приписывания объектам квазичисловых оценок - баллов. Многочисленные примеры тому даны английским статистиком М. Кендэлом [8]. При оценке качества промышленной продукции широко применяемые нормативные и методические документы предусматривают использование ранжировок.

Для прикладных областей, кроме ранжировок и разбиений, представляют интерес толерантности, т.е. рефлексивные симметричные отношения. Толерантность - математическая модель для выражения представлений о сходстве (похожести, близости). Разбиения - частный вид толерантностей. Толерантность, обладающая свойством транзитивности - это разбиение. Однако в общем случае толерантность не обязана быть транзитивной. Толерантности появляются во многих постановках теории экспертных оценок, например, как результат парных сравнений (см. ниже).

Напомним, что любое бинарное отношение на конечном множестве может быть описано матрицей из 0 и 1.

**Дихотомические (бинарные) данные.** Это данные, которые могут принимать одно из двух значений (0 или 1), т.е. результаты измерений значений альтернативного признака. Как уже было показано, измерения в шкале наименований и порядковой шкале приводят к бинарным отношениям, а те могут быть выражены как результаты измерений по нескольким альтернативным признакам, соответствующим элементам матриц, описывающих отношения. Дихотомические данные возникают в прикладных исследованиях и многими иными путями.

В настоящее время в большинстве стандартов, технических условий, технических регламентов, договоров на поставку конкретной продукции предусмотрен контроль по альтернативному признаку. Это означает, что единица продукции относится к одной из двух категорий - "годных" или "дефектных", т.е. соответствующих или не соответствующих требованиям стандарта. Отечественными специалистами проведены обширные теоретические

исследования проблем статистического приемочного контроля по альтернативному признаку. Основополагающими в этой области являются работы академика А.Н.Колмогорова. Подход советской вероятностно-статистической школы к проблемам контроля качества продукции отражен в монографиях [9,10] (см. также главу 3.4).

Дихотомические данные - давний объект математической статистики. Особенно большое применение они имеют в экономических и социологических исследованиях, в которых большинство переменных, интересующих специалистов, измеряется по качественным шкалам. При этом дихотомические данные зачастую являются более адекватными, чем результаты измерений по методикам, использующим большее число градаций. В частности, психологические тесты типа MMPI используют только дихотомические данные. На них опираются и популярные в технико-экономическом анализе методы парных сравнений [11].

Элементарным актом в методе парных сравнений является предъявление эксперту для сравнения двух объектов (сравнение может проводиться также прибором). В одних постановках эксперт должен выбрать из двух объектов лучший по качеству, в других - ответить, похожи объекты или нет. В обоих случаях ответ эксперта можно выразить одной из двух цифр (меток)- 0 или 1. В первой постановке: 0, если лучшим объявлен первый объект; 1 - если второй. Во второй постановке: 0, если объекты похожи, схожи, близки; 1 - в противном случае.

Подводя итоги изложенному, можно сказать, что рассмотренные выше данные представимы в виде векторов из 0 и 1 (при этом матрицы, очевидно, могут быть записаны в виде векторов). Поскольку все результаты наблюдений имеют лишь несколько значащих цифр, то, используя двоичную систему счисления, любые виды анализируемых статистическими методами данных можно записать в виде векторов конечной длины (размерности) из 0 и 1. Представляется, что эта возможность в большинстве случаев имеет лишь академический интерес, но во всяком случае можно констатировать, что анализ дихотомических данных необходим во многих прикладных постановках.

**Множества.** Совокупность $X^n$ векторов $X = (x_1, x_2,...,x_n)$ из 0 и 1 размерности $n$ находится во взаимно-однозначном соответствии с совокупностью $2^n$ всех подмножеств множества $N = \{1, 2, ..., n\}$. При этом вектору $X = (x_1, x_2,...,x_n)$ соответствует подмножество $N(X) \subseteq N$, состоящее из тех и только из тех $i$, для которых $x_i = 1$. Это объясняет, почему изложение вероятностных и статистических результатов, относящихся к анализу данных, являющихся объектами нечисловой природы перечисленных выше видов, можно вести на языке конечных случайных множеств, как это было сделано в монографии [2].

Множества как исходные данные появляются и в иных постановках. Из геологических задач исходил Ж. Матерон, из электротехнических - Н.Н. Ляшенко и др. Случайные множества применялись для описания процесса случайного распространения, например распространения информации, слухов, эпидемии или пожара, а также в математической экономике. В монографии [2] рассмотрены приложения случайных множеств в теории экспертных оценок и в теории управления запасами и ресурсами (логистике).

Отметим, что с точки зрения математики реальные объекты можно моделировать случайными множествами как из конечного числа элементов, так и из бесконечного, однако при расчетах на ЭВМ неизбежна дискретизация, т.е. переход к первой из названных возможностей.

**Объекты нечисловой природы как статистические данные.** В эконометрике и прикладной математической статистике наиболее распространенный объект изучения - выборка $x_1, x_2,...,x_n$, т.е. совокупность результатов $n$ наблюдений. В различных областях статистики результат наблюдения - это или число, или конечномерный вектор, или функция... Соответственно проводится, как уже отмечалось, деление прикладной математической статистики: одномерная статистика, многомерный статистический анализ, статистика временных рядов и случайных процессов... В статистике нечисловых данных в качестве результатов наблюдений рассматриваются объекты нечисловой природы, в частности, перечисленных выше видов - измерения в шкалах, отличных от абсолютной, бинарные отношения, вектора из 0 и 1, множества, нечеткие множества. Выборка может состоять из $n$ ранжировок или $n$ толерантностей, или $n$ множеств, или $n$ нечетких множеств и т.д.

Отметим необходимость развития методов статистической обработка "разнотипных данных", обусловленную большой ролью в прикладных исследованиях "признаков смешанной природы". Речь идет о том, что результат наблюдения состояния объекта зачастую представляет собой вектор, у которого часть координат измерена по шкале наименований, часть - по порядковой шкале, часть - по шкале интервалов и т.д. Статистические методы ориентированы обычно либо на абсолютную шкалу, либо на шкалу наименований (анализ таблиц сопряженности), а потому зачастую непригодны для обработки разнотипных данных. Есть и более сложные модели разнотипных данных, например, когда некоторые координаты вектора наблюдений описываются нечеткими множествами.

Для обозначения подобных неклассических результатов наблюдений в 1979 г. в монографии [2] предложен собирательный термин - объекты нечисловой природы. Термин "нечисловой" означает, что структура пространства, в котором лежат результаты наблюдений, не является структурой действительных чисел, векторов или функций, она вообще не является структурой линейного (векторного) пространства. При расчетах объекты числовой природы, разумеется, изображаются с помощью чисел, но эти числа нельзя складывать и умножать.

С целью "стандартизации математических орудий" (выражение группы французских математиков Н.Бурбаки) целесообразно разрабатывать методы статистического анализа данных, пригодные одновременно для всех перечисленных выше видов результатов наблюдений. Кроме того, в процессе развития прикладных исследований выявляется необходимость использования новых видов объектов нечисловой природы, отличных от рассмотренных выше, например, в связи с развитием статистических методов обработки текстовой информации. Поэтому целесообразно ввести еще один вид объектов нечисловой природы - объекты произвольной природы, т.е. элементы множества, на которые не наложено никаких условий (кроме "условий регулярности", необходимых для справедливости доказываемых теорем). Другими словами, в этом случае предполагается, что результаты наблюдений (элементы выборки) лежат в произвольном пространстве $X$. Для получения теорем необходимо потребовать, чтобы $X$ удовлетворяло некоторым условиям, например, было так называемым топологическим пространством. Как известно, ряд результатов классической математической статистики получен именно в такой постановке. Так, при изучении оценок максимального правдоподобия элементы выборки могут лежать в пространстве произвольной природы. Это не влияет на рассуждения, поскольку в них рассматривается лишь зависимость плотности вероятности от параметра. Методы классификации, использующие лишь расстояние между классифицируемыми объектами, могут применяться к совокупностям объектов произвольной природы, лишь бы в пространстве, где они лежат, была задана метрика. Цель статистики нечисловых данных (в некоторых литературных источниках используется термин "статистика объектов нечисловой природы") состоит в том, чтобы систематически рассматривать методы статистической обработки данных как произвольной природы, так и относящихся к указанным выше конкретным видам объектов нечисловой природы, т.е. методы описания данных, оценивания и проверки гипотез. Взгляд с общей точки зрения позволяет получить новые результаты и в других областях прикладной статистики.

**Использование объектов нечисловой природы при формировании статистической или математической модели реального явления.** Использование объектов нечисловой природы часто порождено желанием обрабатывать более объективную, более освобожденную от погрешностей информацию. Как показали многочисленные опыты, человек более правильно (и с меньшими затруднениями) отвечает на вопросы качественного например, сравнительного, характера, чем количественного. Так, ему легче сказать, какая из двух гирь тяжелее, чем указать их примерный вес в граммах. Другими словами, использование объектов нечисловой природы - средство повысить устойчивость эконометрических и экономико-математических моделей реальных явлений. Сначала конкретные области статистики объектов нечисловой природы (а именно, прикладная теория измерений, нечеткие и случайные множества) были рассмотрены в монографии [2] как частные постановки проблемы устойчивости математических моделей социально-экономических явлений и процессов к допустимым отклонениям исходных данных и предпосылок модели, а затем была понята

необходимость проведения работ по развитию статистики объектов нечисловой природы как самостоятельного научного направления.

Обсуждение начнем со шкал измерения. Науку о единстве мер и точности измерений называют метрологией. Таким образом, репрезентативная теория измерений - часть метрологии. Методы обработки данных должны быть адекватны относительно допустимых преобразований шкал измерения в смысле репрезентативной теории измерений. Однако установление типа шкалы, т.е. задание группы преобразований $\Phi$ - дело специалиста соответствующей прикладной области. Так, оценки привлекательности профессий мы считали измеренными в порядковой шкале. Однако отдельные социологи не соглашались с этим, считая, что выпускники школ пользуются шкалой с более узкой группой допустимых преобразований, например, интервальной шкалой. Очевидно, эта проблема относится не к математике, а к наукам о человеке. Для ее решения может быть поставлен достаточно трудоемкий эксперимент. Пока же он не поставлен, целесообразно принимать порядковую шкалу, так как это гарантирует от возможных ошибок.

Порядковые шкалы широко распространены не только в социально-экономических исследованиях. Они применяются в медицине - шкала стадий гипертонической болезни по Мясникову, шкала степеней сердечной недостаточности по Стражеско-Василенко-Лангу, шкала степени выраженности коронарной недостаточности по Фогельсону; в минералогии - шкала Мооса (тальк - 1, гипс - 2, кальций - 3, флюорит - 4, апатит - 5, ортоклаз - 6, кварц - 7, топаз - 8, корунд - 9, алмаз - 10), по которому минералы классифицируются согласно критерию твердости; в географии - бофортова шкала ветров ("штиль", "слабый ветер", "умеренный ветер" и др.) и т.д. Напомним, что по шкале интервалов измеряют величину потенциальной энергии или координату точки на прямой, на которой не отмечены ни начало, ни единица измерения; по шкале отношений - большинство физических единиц: массу тела, длину, заряд, а также цены в экономике. Время измеряется по шкале разностей, если год принимаем естественной единицей измерения, и по шкале интервалов в общем случае. В процессе развития соответствующей области знания тип шкалы может меняться. Так, сначала температура измерялась по порядковой шкале (холоднее - теплее), затем - по интервальной (шкалы Цельсия, Фаренгейта, Реомюра) и, наконец, после открытия абсолютного нуля температур - по шкале отношений (шкала Кельвина). Следует отметить, что среди специалистов иногда имеются разногласия по поводу того, по каким шкалам следует считать измеренными те или иные реальные величины.

Отметим, что термин "репрезентативная" использовался, чтобы отличить рассматриваемый подход к теории измерений от классической метрологии, а также от работ А.Н.Колмогорова и А. Лебега, связанных с измерением геометрических величин, от "алгоритмической теории измерения" и др.

Необходимость использования в математических моделях реальных явлений таких объектов нечисловой природы, как бинарные отношения, множества, нечеткие множества, кратко была показана выше. Здесь же обратим внимание, что используемые в классической статистике результаты наблюдений также "не совсем числа". А именно, любая величина $X$ измеряется всегда с некоторой погрешностью $\Delta X$ и результатом наблюдения является

$$Y = X + \Delta X.$$

Как уже отмечалось, погрешностями измерений занимается метрология. Отметим справедливость следующих фактов:

а) для большинства реальных измерений невозможно полностью исключить систематическую ошибку, т.е. $M(\Delta X) \neq 0$;

б) распределение $\Delta X$ в подавляющем большинстве случаев не является нормальным (см. главу 2.1);

в) измеряемую величину $X$ и погрешность ее измерения $\Delta X$ обычно нельзя считать независимыми случайными величинами;

г) распределение погрешностей оценивается по результатам специальных наблюдений, следовательно, полностью известным считать его нельзя; зачастую исследователь располагает лишь границами для систематической погрешности и оценками таких характеристик для случайной погрешности, как дисперсия или размах.

Приведенные факты показывают ограниченность области применимости распространенной модели погрешностей, в которой $X$ и $\Delta X$ рассматриваются как независимые случайные величины, причем $\Delta X$ имеет нормальное распределение с нулевым математическим ожиданием.

Строго говоря, результаты наблюдения всегда имеют дискретное распределение, поскольку описываются числами с небольшими (1 - 5) числом значащих цифр. Возникает дилемма: либо признать, что непрерывные распределения - фикция, и прекратить ими пользоваться, либо считать, что непрерывные распределения имеют "реальные" величины $X$, которые мы наблюдаем с принципиально неустранимой погрешностью $\Delta X$. Первый выход в настоящее время нецелесообразен, так как потребует отказаться от большей части разработанного математического аппарата. Из второго следует необходимость изучения влияния неустранимых погрешностей на статистические выводы.

Погрешности $\Delta X$ можно учитывать либо с помощью вероятностной модели ($\Delta X$ - случайная величина, имеющая функцию распределения, вообще говоря, зависящую от $X$), либо с помощью нечетких множеств. Во втором случае приходим к теории нечетких чисел и к ее частному случаю - статистике интервальных данных (см. главу 3.5).

Другой источник появления погрешности $\Delta X$ связан с принятой в конструкторской и технологической документации системой допусков на контролируемые параметры изделий и деталей, с использованием шаблонов при проверке контроля качества продукции [12]. В этих случаях характеристики $\Delta X$ определяются не свойствами средств измерения, а применяемой технологией проектирования и производства. В терминах прикладной статистики сказанному соответствует группировка данных, при которой мы знаем, какому из заданных интервалов принадлежит наблюдение, но не знаем точного значения результата наблюдения. Применение группировки может дать экономический эффект, поскольку зачастую легче (в среднем) установить, к какому интервалу относится результат наблюдения, чем точно измерить его.

**Объекты нечисловой природы как результат статистической обработки данных.** Объекты нечисловой природы появляются не только на "входе" статистической процедуры, но и в процессе обработки данных, и на "выходе" в качестве итога статистического анализа.

Рассмотрим простейшую прикладную постановку задачи регрессии (см. также главу 3.2). Исходные данные имеют вид $(x_i, y_i) \in R^2, i = 1,2,...,n.$. Цель состоит в том, чтобы с достаточной точностью описать $y$ как полином от $x$, т.е. модель имеет вид

$$y_i = \sum_{k=0}^{m} a_j x_i^k + \varepsilon_i, \qquad (2)$$

где $m$ - неизвестная степень полинома; $a_0, a_1, a_2, ..., a_m$ - неизвестные коэффициенты многочлена; $\varepsilon_i, i = 1,2,...,n$, - погрешности, которые для простоты примем независимыми и имеющими одно и то же нормальное распределение. (Здесь наглядно проявляется одна из причин живучести статистических моделей на основе нормального распределения. Такие модели, хотя и, как правило, неадекватны реальной ситуации (см. главу 2.1), с математической точки зрения позволяет проникнуть глубже в суть изучаемого явления. Поэтому они пригодны для первоначального анализа ситуации, как и в рассматриваемом случае. Дальнейшие научные исследования должны быть направлены на снятие нереалистического предположения нормальности и перехода к непараметрическим моделям погрешности.) Распространенная процедура такова: сначала пытаются применить модель (2) для линейной функции ($m$ = 1), при неудаче (неадекватности модели) переходят к многочлену второго порядка ($m$ = 2), если снова неудача, то берут модель (2) с $m$= 3 и т.д. (адекватность модели проверяют по $F$-критерию Фишера).

Обсудим свойства этой процедуры в терминах прикладной статистики. Если степень полинома задана ($m = m_0$), то его коэффициенты оценивают методом наименьших квадратов, свойства этих оценок хорошо известны (см., например, главу 3.2 или монографию [13, гл.26]). Однако в описанной выше реальной постановке $m$ тоже является неизвестным параметром и подлежит оценке. Таким образом, требуется оценить объект ($m$, $a_0$, $a_1$, $a_2$, …, $a_m$), множество

значений которого можно описать как $R^1 \cup R^2 \cup R^3 \cup \ldots$ Это - объект нечисловой природы, обычные методы оценивания для него неприменимы, так как *m* - дискретный параметр. В рассматриваемой постановке разработанные к настоящему времени методы оценивания степени полинома носят в основном эвристический характер (см., например, гл. 12 монографии [14]). Свойства описанной выше распространенной процедуры рассмотрены в главе 3.2. Там показано, что степень полинома *m* при этом оценивается несостоятельно, и найдено предельное распределение оценки этого параметра, оказавшееся геометрическим.

В более общем случае линейной регрессии данные имеют вид $(y_i, X_i), i = 1,2,...,n,$ где $X_i = (x_{i1}, x_{i2}, ..., x_{iN}) \in R^N$ - вектор предикторов (факторов, объясняющих переменных), а модель такова:

$$y_i = \sum_{j \in K} a_j x_{ij} + \varepsilon_i, i = 1,2,...,n \qquad (3)$$

(здесь *K* - некоторое подмножество множества {1,2,…,n}; $\varepsilon_i$ - те же, что и в модели (2); $a_j$ - неизвестные коэффициенты при предикторах с номерами из *K*). Модель (2) сводится к модели (3), если

$$x_{i1} = 1, x_{i1} = x_i, x_{i2} = x_i^2, x_{i3} = x_i^3, ..., x_{ij} = x_i^{j-1}, ...$$

В модели (2) есть естественный порядок ввода предикторов в рассмотрение - в соответствии с возрастанием степени, а в модели (3) естественного порядка нет, поэтому здесь стоит произвольное подмножество множества предикторов. Есть только частичный порядок - чем мощность подмножества меньше, тем лучше. Модель (3) особенно актуальна в технических исследованиях (см. многочисленные примеры в журнале «Заводская лаборатория»). Она применяется в задачах управления качеством продукции и других технико-экономических исследованиях, в экономике, маркетинге и социологии, когда из большого числа факторов, предположительно влияющих на изучаемую переменную, надо отобрать по возможности наименьшее число значимых факторов и с их помощью сконструировать прогнозирующую формулу (3).

Задача оценивания модели (3) разбивается на две последовательные задачи: оценивание множества *K* - подмножества множества всех предикторов, а затем - неизвестных параметров $a_j$. Методы решения второй задачи хорошо известны и подробно изучены. Гораздо хуже обстоит дело с оцениванием объекта нечисловой природы *K*. Как уже отмечалось, существующие методы - в основном эвристические, они зачастую не являются даже состоятельными. Даже само понятие состоятельности в данном случае требует специального определения. Пусть $K_0$ - истинное подмножество предикторов, т.е. подмножество, для которого справедлива модель (3), а подмножество предикторов $K_n$ - его оценка. Оценка $K_n$ называется состоятельной, если

$$\lim_{n \to \infty} Card(K_n \Delta K_0) = 0,$$

где $\Delta$ - символ симметрической разности множеств; *Card(K)* означает число элементов в множестве *K*, а предел понимается в смысле сходимости по вероятности.

Задача оценивания в моделях регрессии, таким образом, разбивается на две - оценивание структуры модели и оценивание параметров при заданной структуре. В модели (2) структура описывается неотрицательным целым числом *m*, в модели (3) - множеством *K*. Структура - объект нечисловой природы. Задача ее оценивания сложна, в то время как задача оценивания численных параметров при заданной структуре хорошо изучена, разработаны эффективные (в смысле прикладной математической статистики) методы.

Такова же ситуация и в других методах многомерного статистического анализа - в факторном анализе (включая метод главных компонент) и в многомерном шкалировании, в иных оптимизационных постановках проблем прикладного многомерного статистического анализа.

Перейдем к объектам нечисловой природы на "выходе" статистической процедуры. Примеры многочисленны. Разбиения - итог работы многих алгоритмов классификации, в частности, алгоритмов кластер-анализа. Ранжировки - результат упорядочения профессий по

привлекательности или автоматизированной обработки мнений экспертов - членов комиссии по подведению итогов конкурса научных работ. (В последнем случае используются ранжировки со связями; так, в одну группу, наиболее многочисленную, попадают работы, не получившие наград.) Из всех объектов нечисловой природы, видимо, наиболее часты на "выходе" дихотомические данные - принять или не принять гипотезу, в частности, принять или забраковать партию продукции. Результатом статистической обработка данных может быть множество, например зона наибольшего поражения при аварии, или последовательность множеств, например, "среднемерное" описание распространения пожара (см. главу 4 в монографии [2]). Нечетким множеством Э. Борель [15] еще в начале XX в. предлагал описывать представление людей о числе зерен, образующем "кучу". С помощью нечетких множеств формализуются значения лингвистических переменных, выступающих как итоговая оценка качества систем автоматизированного проектирования, сельскохозяйственных машин, бытовых газовых плит, надежности программного обеспечения или систем управления. Можно констатировать, что все виды объектов нечисловой природы могут появляться " на выходе" статистического исследования.

### 1.1.4. Нечеткие множества – частный случай нечисловых данных

**Нечеткие множества.** Пусть $A$ - некоторое множество. Подмножество $B$ множества $A$ характеризуется своей характеристической функцией

$$\mu_B(x) = \begin{cases} 1, & x \in B, \\ 0, & x \notin B. \end{cases} \qquad (1)$$

Что такое нечеткое множество? Обычно говорят, что нечеткое подмножество $C$ множества $A$ характеризуется своей функцией принадлежности $\mu_C : A \to [0,1].$ Значение функции принадлежности в точке $x$ показывает степень принадлежности этой точки нечеткому множеству. Нечеткое множество описывает неопределенность, соответствующую точке $x$ – она одновременно и входит, и не входит в нечеткое множество $C$. За вхождение - $\mu_c(x)$ шансов, за второе – (1- $\mu_c(x)$) шансов.

Если функция принадлежности $\mu_C(x)$ имеет вид (1) при некотором $B$, то $C$ есть обычное (четкое) подмножество $A$. Таким образом, теория нечетких множество является не менее общей математической дисциплиной, чем обычная теория множеств, поскольку обычные множества – частный случай нечетких. Соответственно можно ожидать, что теория нечеткости как целое обобщает классическую математику. Однако позже мы увидим, что теория нечеткости в определенном смысле сводится к теории случайных множеств и тем самым является частью классической математики. Другими словами, по степени общности обычная математика и нечеткая математика эквивалентны. Однако для практического применения в теории принятия решений описание и анализ неопределенностей с помощью теории нечетких множеств весьма плодотворны.

Обычное подмножество можно было бы отождествить с его характеристической функцией. Этого математики не делают, поскольку для задания функции (в ныне принятом подходе) необходимо сначала задать множество. Нечеткое же подмножество с формальной точки зрения можно отождествить с его функцией принадлежности. Однако термин "нечеткое подмножество" предпочтительнее при построении математических моделей реальных явлений.

Теория нечеткости является обобщением интервальной математики. Действительно, функция принадлежности

$$\mu_B(x) = \begin{cases} 1, & x \in [a,b], \\ 0, & x \notin [a,b] \end{cases}$$

задает интервальную неопределенность – про рассматриваемую величину известно лишь, что она лежит в заданном интервале [*a*,*b*]. Тем самым описание неопределенностей с помощью нечетких множеств является более общим, чем с помощью интервалов.

Начало современной теории нечеткости положено работой 1965 г. американского ученого азербайджанского происхождения Л.А.Заде. К настоящему времени по этой теории опубликованы тысячи книг и статей, издается несколько международных журналов, выполнено достаточно много как теоретических, так и прикладных работ. Первая книга российского автора по теории нечеткости вышла в 1980 г. [16].

Л.А. Заде рассматривал теорию нечетких множеств как аппарат анализа и моделирования гуманистических систем, т.е. систем, в которых участвует человек. Его подход опирается на предпосылку о том, что элементами мышления человека являются не числа, а элементы некоторых нечетких множеств или классов объектов, для которых переход от "принадлежности" к "непринадлежности" не скачкообразен, а непрерывен. В настоящее время методы теории нечеткости используются почти во всех прикладных областях, в том числе при управлении предприятиями, качеством продукции и технологическими процессами, при описании предпочтений потребителей и варки стали.

Л.А. Заде использовал термин "fuzzy set" (нечеткое множество). На русский язык термин "fuzzy" переводили как нечеткий, размытый, расплывчатый, и даже как пушистый и туманный.

Аппарат теории нечеткости громоздок. В качестве примера дадим определения теоретико-множественных операций над нечеткими множествами. Пусть *C* и *D*- два нечетких подмножества *A* с функциями принадлежности $\mu_C(x)$ и $\mu_D(x)$ соответственно. Пересечением $C \cap D$, произведением *CD*, объединением $C \cup D$, отрицанием $\overline{C}$, суммой *C*+*D* называются нечеткие подмножества *A* с функциями принадлежности

$$\mu_{C\cap D}(x) = \min(\mu_C(x), \mu_D(x)), \quad \mu_{CD}(x) = \mu_C(x)\mu_D(x), \quad \mu_{\overline{C}}(x) = 1 - \mu_C(x),$$

$$\mu_{C\cup D}(x) = \max(\mu_C(x), \mu_D(x)), \quad \mu_{C+D}(x) = \mu_C(x) + \mu_D(x) - \mu_C(x)\mu_D(x), \quad x \in A,$$

соответственно.

Как уже отмечалось, теория нечетких множеств в определенном смысле сводится к теории вероятностей, а именно, к теории случайных множеств. Соответствующий цикл теорем приведен ниже в главе 1.4. Однако при решении прикладных задач вероятностно-статистические методы и методы теории нечеткости обычно рассматриваются как различные.

Для знакомства со спецификой нечетких множеств рассмотрим некоторые их свойства.

В дальнейшем считаем, что все рассматриваемые нечеткие множества являются подмножествами одного и того же множества *Y*.

**Законы де Моргана для нечетких множеств.** Как известно, законами же Моргана называются следующие тождества алгебры множеств

$$\overline{A \cup B} = \overline{A} \cap \overline{B}, \quad \overline{A \cap B} = \overline{A} \cup \overline{B}. \tag{2}$$

**Теорема 1.** Для нечетких множеств справедливы тождества

$$\overline{A \cup B} = \overline{A} \cap \overline{B}, \quad \overline{A \cap B} = \overline{A} \cup \overline{B}, \tag{3}$$

$$\overline{A + B} = \overline{A}\,\overline{B}, \quad \overline{AB} = \overline{A} + \overline{B}. \tag{4}$$

Доказательство теоремы 1 состоит в непосредственной проверке справедливости соотношений (3) и (4) путем вычисления значений функций принадлежности участвующих в этих соотношениях нечетких множеств на основе определений, данных выше.

Тождества (3) и (4) назовем *законами де Моргана для нечетких множеств*. В отличие от классического случая соотношений (2), они состоят из четырех тождеств, одна пара которых относится к операциям объединения и пересечения, а вторая - к операциям произведения и суммы. Как и соотношение (2) в алгебре множеств, законы де Моргана в алгебре нечетких множеств позволяют преобразовывать выражения и формулы, в состав которых входят операции отрицания.

**Дистрибутивный закон для нечетких множеств.** Некоторые свойства операций над множествами не выполнены для нечетких множеств. Так, $A + A \neq A,$ за исключением случая, когда *A* - "четкое" множество (т.е. функция принадлежности принимает только значения 0 и 1).

Верен ли дистрибутивный закон для нечетких множеств? В литературе иногда расплывчато утверждается, что "не всегда". Внесем полную ясность.

**Теорема 2.** Для любых нечетких множеств *A, B* и *C*

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C). \qquad (5)$$

В то же время равенство

$$A(B + C) = AB + AC \qquad (6)$$

справедливо тогда и только тогда, когда при всех $y \in Y$

$$(\mu_A^2(y) - \mu_A(y))\mu_B(y)\mu_C(y) = 0.$$

*Доказательство*. Фиксируем произвольный элемент $y \in Y$. Для сокращения записи обозначим $a = \mu_A(y), b = \mu_B(y), c = \mu_C(y).$ Для доказательства тождества (5) необходимо показать, что

$$\min(a, \max(b, c)) = \max(\min(a, b), \min(a, c)). \qquad (7)$$

Рассмотрим различные упорядочения трех чисел *a, b, c.* Пусть сначала $a \le b \le c.$ Тогда левая часть соотношения (7) есть $\min(a, c) = a,$ а правая $\max(a, a) = a,$ т.е. равенство (7) справедливо.

Пусть $b \le a \le c.$ Тогда в соотношении (7) слева стоит $\min(a, c) = a,$ а справа $\max(b, a) = a,$ т.е. соотношение (7) опять является равенством.

Если $b \le c \le a,$ то в соотношении (7) слева стоит $\min(a, c) = c,$ а справа $\max(b, c) = c,$ т.е. обе части снова совпадают.

Три остальные упорядочения чисел *a, b, c* разбирать нет необходимости, поскольку в соотношение (6) числа *b* и *c* входят симметрично. Тождество (5) доказано.

Второе утверждение теоремы 2 вытекает из того, что в соответствии с определениями операций над нечеткими множествами

$$\mu_{A(B+C)}(y) = a(b + c - bc) = ab + ac - abc$$

и

$$\mu_{AB+AC}(y) = ab + ac - (ab)(ac) = ab + ac - a^2bc.$$

Эти два выражения совпадают тогда и только тогда, когда, когда $a^2bc = abc,$ что и требовалось доказать.

**Определение 1.** Носителем нечеткого множества *A* называется совокупность всех точек $y \in Y$, для которых $\mu_A(y) > 0.$

**Следствие теоремы 2.** Если носители нечетких множеств *B* и *C* совпадают с *У,* то равенство *(6) имеет место тогда и только тогда, когда A* - "четкое" (т.е. обычное, классическое, не нечеткое) множество.

*Доказательство.* По условию $\mu_B(y)\mu_C(y) \neq 0$ при всех $y \in Y$. Тогда из теоремы 2 следует, что $\mu_A^2(y) - \mu_A(y) = 0,$ т.е. $\mu_A(y) = 1$ или $\mu_A(y) = 0$, что и означает, что *A* - четкое множество.

**Пример описания неопределенности с помощью нечеткого множества.** Понятие «богатый» часто используется при обсуждении социально-экономических проблем, в том числе и в связи с подготовкой и принятием решений. Однако очевидно, что разные лица вкладывают в это понятие различное содержание. Сотрудники Института высоких статистических технологий и эконометрики провели в 1996 г. небольше пилотное социологическое исследование представления различных слоёв населения о понятии "богатый человек".

Мини-анкета опроса выглядела так:

1. При каком месячном доходе (в млн. руб. на одного человека) Вы считали бы себя богатым человеком?
2. Оценив свой сегодняшний доход, к какой из категорий Вы себя относите:
   а) богатые;
   б) достаток выше среднего;
   в) достаток ниже среднего;
   г) бедные;
   д) за чертой бедности?

(В дальнейшем вместо полного наименования категорий будем оперировать буквами, например "в" - категория, "б" - категория и т.д.)

3. Ваша профессия, специальность.

Всего было опрошено 74 человека, из них 40 - научные работники и преподаватели, 34 человека - не занятых в сфере науки и образования, в том числе 5 рабочих и 5 пенсионеров. Из всех опрошенных только один (!) считает себя богатым. Несколько типичных ответов научных работников и преподавателей приведено в табл.1, а аналогичные сведения для работников коммерческой сферы – в табл.2.

Таблица 1.
Типичные ответы научных работников и преподавателей

| Ответы на вопрос 3 | Ответы на вопрос 1, млн. руб./чел. | Ответы на вопрос 2 | Пол |
|---|---|---|---|
| Кандидат наук | 1 | д | ж |
| Преподаватель | 1 | в | ж |
| Доцент | 1 | б | ж |
| Учитель | 10 | в | м |
| Старший. научный сотрудник | 10 | д | м |
| Инженер-физик | 24 | д | ж |
| Программист | 25 | г | м |
| научный работник | 45 | г | м |

Таблица 2
Типичные ответы работников коммерческой сферы.

| Ответы на вопрос 3 | Ответы на вопрос 1 | Ответы на вопрос 2 | Пол |
|---|---|---|---|
| Вице-президент банка | 100 | а | ж |
| Зам. директора банка | 50 | б | ж |
| Начальник. кредитного отдела | 50 | б | м |
| Начальник отдела ценных бумаг | 10 | б | м |
| Главный бухгалтер | 20 | д | ж |
| Бухгалтер | 15 | в | ж |
| Менеджер банка | 11 | б | м |
| Начальник отдела проектирования | 10 | в | ж |

Разброс ответов на первый вопрос – от 1 до 100 млн. руб. в месяц на человека. Результаты опроса показывают, что критерий богатства у финансовых работников в целом несколько выше, чем у научных (см. гистограммы на рис.1 и рис.2 ниже).

Опрос показал, что выявить какое-нибудь конкретное значение суммы, которая необходима "для полного счастья", пусть даже с небольшим разбросом, нельзя, что вполне естественно. Как видно из таблиц 1 и 2, денежный эквивалент богатства колеблется от 1 до 100 миллионов рублей в месяц. Подтвердилось мнение, что работники сферы образования в

подавляющем большинстве причисляют свой достаток к категории "в" и ниже (81% опрошенных), в том числе к категории "д" отнесли свой достаток 57%.

Со служащими коммерческих структур и бюджетных организаций иная картина: "г" - категория 1 человек (4%), "д" - категория 4 человека (17%), "б" - категория - 46% и 1 человек "а" - категория.

Пенсионеры, что не вызывает удивления, отнесли свой доход к категории "д" (4 человека), и лишь один человек указал "г" - категорию. Рабочие же ответили так: 4 человека - "в", и один человек - "б".

Для представления общей картины в табл.3 приведены данные об ответах работников других профессий.

Таблица 3.

Типичные ответы работников различных профессий.

| Ответы на вопрос 3 | Ответы на вопрос 1 | Ответы на вопрос 2 | Пол |
|---|---|---|---|
| Работник торговли | 1 | б | ж |
| Дворник | 2 | в | ж |
| Водитель | 10 | в | м |
| Военнослужащий | 10 | в | м |
| Владелец бензоколонки | 20 | б | ж |
| Пенсионер | 6 | д | ж |
| Начальник фабрики | 20 | б | м |
| Хирург | 5 | в | м |
| Домохозяйка | 10 | в | ж |
| Слесарь-механик | 25 | в | м |
| Юрист | 10 | б | м |
| Оператор ЭВМ | 20 | д | м |
| Работник собеса | 3 | д | ж |
| Архитектор | 25 | б | ж |

Прослеживается интересное явление: чем выше планка богатства для человека, тем к более низкой категории относительно этой планки он себя относит.

Для сводки данных естественно использовать гистограммы. Для этого необходимо сгруппировать ответы. Использовались 7 классов (интервалов):

1 – до 5 миллионов рублей в месяц на человека (включительно);
2 – от 5 до 10 миллионов;
3- от 10 до 15 миллионов;
4 – от 15 до 20 миллионов;
5 – от 20 до 25 миллионов;
6 – от 25 до 30 миллионов;
7 – более 30 миллионов.

(Во всех интервалах левая граница исключена, а правая, наоборот – включена.)

Сводная информация представлена на рис.1 (для научных работников и преподавателей) и рис.2 (для всех остальных, т.е. для лиц, не занятых в сфере науки и образования - служащих иных бюджетных организаций, коммерческих структур, рабочих, пенсионеров).

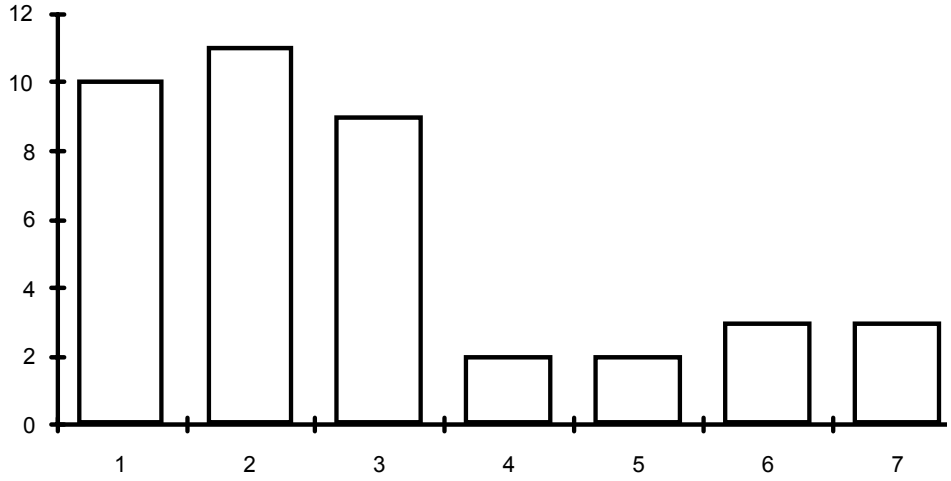



Рис.1. Гистограмма ответов на вопрос 1 для научных работников и преподавателей (40 человек).

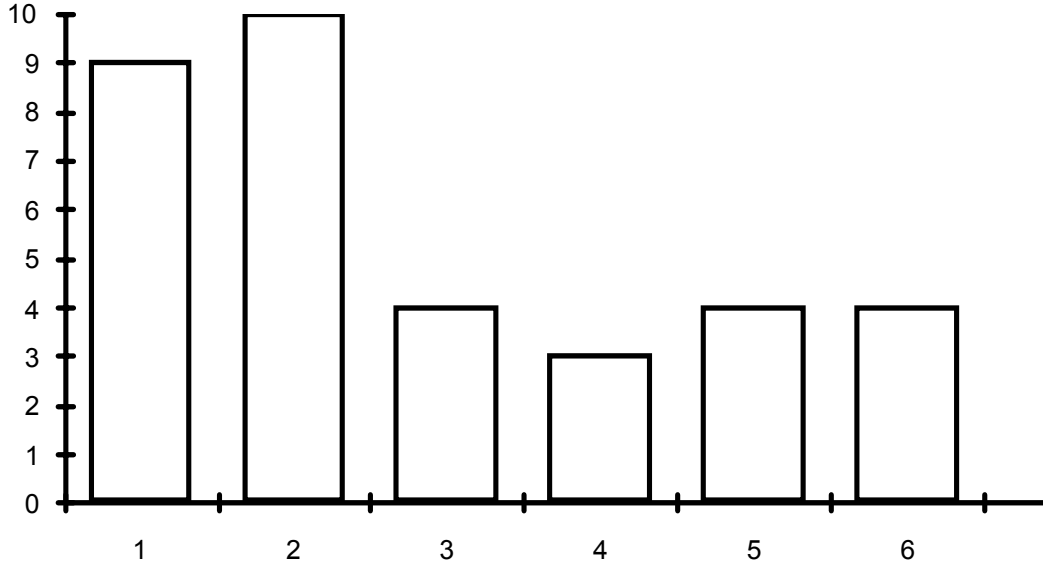



Рис.2. Гистограмма ответов на вопрос 1 для лиц, не занятых в сфере науки и образования (34 человека).

Для двух выделенных групп, а также для некоторых подгрупп второй группы рассчитаны сводные средние характеристики – выборочные средние арифметические, медианы, моды. При этом медиана группы - количество млн. руб., названное центральным по порядковому номеру опрашиваемым в возрастающем ряду ответов на вопрос 1, а мода группы - интервал, на котором столбик гистограммы - самый высокий, т.е. в него "попало" максимальное количество опрашиваемых. Результаты приведены в табл. 4.

Таблица 4.
Сводные средние характеристики ответов на вопрос 1 для различных групп (в млн. руб. в мес. на чел.).

| Группа опрошенных | Среднее арифметическое | медиана | мода |
|---|---|---|---|
| Научные работники и преподаватели | 11,66 | 7,25 | (5; 10) |
| Лиц, не занятых в сфере науки и образования | 14,4 | 20 | (5; 10) |
| Служащие коммерческих структур и | 17,91 | 10 | (5; 10) |

| | | | |
|---|---|---|---|
| бюджетных организаций | | | |
| Рабочие | 15 | 13 | - |
| Пенсионеры | 10,3 | 10 | - |

Построим нечеткое множество, описывающее понятие «богатый человек» в соответствии с представлениями опрошенных. Для этого составим табл.5 на основе рис.1 и рис.2 с учетом размаха ответов на первый вопрос.

Таблица 5.

Число ответов, попавших в интервалы

| № | Номер интервала | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| 1 | Интервал, млн. руб. в месяц | (0;1) | [1;5] | (5;10] | (10;15] | (15;20] |
| 2 | Число ответов в интервале | 0 | 19 | 21 | 13 | 5 |
| 3 | Доля ответов в интервале | 0 | 0,257 | 0,284 | 0,176 | 0,068 |
| 4 | Накопленное число ответов | 0 | 19 | 40 | 53 | 58 |
| 5 | Накопленная доля ответов | 0 | 0,257 | 0,541 | 0,716 | 0,784 |

Продолжение табл.5.

| № | Номер интервала | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|
| 1 | Интервал, млн. руб. в месяц | (20;25] | (25;30] | (30;100) | [100;+∞) |
| 2 | Число ответов в интервале | 6 | 7 | 2 | 1 |
| 3 | Доля ответов в интервале | 0,081 | 0,095 | 0,027 | 0,013 |
| 4 | Накопленное Нечетким множеством число ответов | 64 | 71 | 73 | 74 |
| 5 | Накопленная доля ответов | 0,865 | 0,960 | 0,987 | 1,000 |

Пятая строка табл.5 задает функцию принадлежности нечеткого множества, выражающего понятие "богатый человек" в терминах его ежемесячного дохода. Это нечеткое множество является подмножеством множества из 9 интервалов, заданных в строке 2 табл.5. Или множества из 9 условных номеров {0, 1, 2, …, 8}. Эмпирическая функция распределения, построенная по выборке из ответов 74 опрошенных на первый вопрос мини-анкеты, описывает понятие "богатый человек" как нечеткое подмножество положительной полуоси.

**О разработке методики ценообразования на основе теории нечетких множеств.** Для оценки значений показателей, не имеющих количественной оценки, можно использовать методы нечетких множеств. Например, в диссертации П.В. Битюкова [17] нечеткие множества применялись при моделировании задач ценообразования на электронные обучающие курсы, используемые при дистанционном обучении. Им было проведено исследование значений фактора «Уровень качества курса» с использованием нечетких множеств. В ходе практического использования предложенной П.В. Битюковым методики ценообразования значения ряда других факторов могут также определяться с использованием теории нечетких множеств. Например, ее можно использовать для определения прогноза рейтинга специальности в вузе с помощью экспертов, а также значений других факторов, относящихся к группе «Особенности курса». Опишем подход П.В. Битюкова как пример практического использования теории нечетким множеств.

Значение оценки, присваиваемой каждому интервалу для фактора «Уровень качества курса», определяется на универсальной шкале [0,1], где необходимо разместить значения лингвистической переменной «Уровень качества курса»: НИЗКИЙ, СРЕДНИЙ, ВЫСОКИЙ. Степень принадлежности некоторого значения вычисляется как отношение числа ответов, в которых оно встречалось в определенном интервале шкалы, к максимальному (для этого значения) числу ответов по всем интервалам.

Был проведен опрос экспертов о степени влияния уровня качества электронных курсов на их потребительную ценность. Каждому эксперту в процессе опроса предлагалось оценить с позиции потребителя ценность того или иного класса курсов в зависимости от уровня

качества. Эксперты давали свою оценку для каждого класса курсов по 10-ти балльной шкале (где 1 - min, 10 - max). Для перехода к универсальной шкале [0,1], все значения 10-ти балльной шкалы оценки ценности были разделены на максимальную оценку, т.е. на 10.

Используя свойства функции принадлежности, необходимо предварительно обработать данные с тем, чтобы уменьшить искажения, вносимые опросом. Естественными свойствами функций принадлежности являются наличие одного максимума и гладкие, затухающие до нуля фронты. Для обработки статистических данных можно воспользоваться так называемой матрицей подсказок. Предварительно удаляются явно ошибочные элементы. Критерием удаления служит наличие нескольких нулей в строке вокруг этого элемента.

Элементы матрицы подсказок вычисляются по формуле: $k_j = \sum_{i=1}^{n} b_{ij}, j = \overline{1,n}$,

где $b_{ij}$ - элемент таблицы с результатами анкетирования, сгруппированными по интервалам. Матрица подсказок представляет собой строку, в которой выбирается максимальный элемент: $k_{\max} = \max_j k_j$, и далее все ее элементы преобразуются по формуле:

$$c_{ij} = \frac{b_{ij} k_{\max}}{k_j}, i = \overline{1,m}, j = \overline{1,n}.$$

Для столбцов, где $k_j = 0$, применяется линейная аппроксимация:

$$c_{ij} = \frac{c_{ij-1} + c_{ij+1}}{2}, i = \overline{1,m}, j = \overline{1,n}.$$

Результаты расчетов сводятся в таблицу, на основании которой строятся функции принадлежности. Для этого находятся максимальные элементы по строкам: $c_{i\max} = \max_j c_{ij}, i = \overline{1,m}, j = \overline{1,n}$. Функция принадлежности вычисляется по формуле: $м_{ij} = c_{ij} / c_{i\max}$. Результаты расчетов приведены в табл. 6.

Таблица 6

Значения функции принадлежности лингвистической переменной

| $м_i$ | Интервал на универсальной шкале | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0,1 | 0,2 | 0,3 | 0,4 | 0,5 | 0,6 | 0,7 | 0,8 | 0,9 | 1,0 |
| $м_1$ | 0 | 0,2 | 1 | 1 | 0,89 | 0,67 | 0 | 0 | 0 | 0 |
| **$м_2$** | **0** | **0** | **0** | **0** | **0** | **0,33** | **1** | **1** | **0** | **0** |
| **$м_3$** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **1** | **1** |

На рис.3 сплошными линиями показаны функции принадлежности значений лингвистической переменной «Уровень качества курса» после обработки таблицы, содержащей результаты опроса. Как видно из графика, функции принадлежности удовлетворяют описанным выше свойствам. Для сравнения пунктирной линией показана функция принадлежности лингвистической переменной для значения НИЗКИЙ без обработки данных.

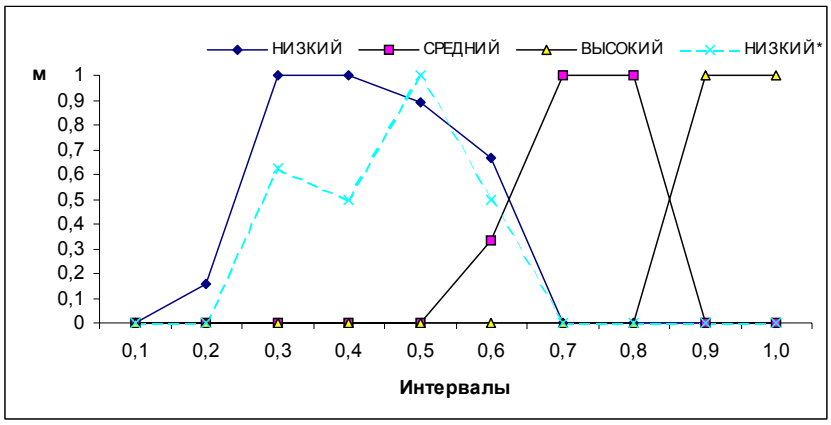



Рис. 3. График функций принадлежности значений лингвистической переменной «Уровень качества курса»

### 1.1.5. Данные и расстояния в пространствах произвольной природы

Как показано выше, исходные статистические данные могут иметь разнообразную математическую природу, являться элементами разнообразных пространств – конечномерных, функциональных, бинарных отношений, множеств, нечетких множеств и т.д. Следовательно, центральной частью прикладной статистики является статистика в пространствах произвольной природы. Эта область прикладной статистики сама по себе не используется при анализе конкретных данных. Это очевидно, поскольку конкретные данные всегда имеют вполне определенную природу. Однако общие подходы, методы, результаты статистики в пространствах произвольной природы представляют собой научный инструментарий, готовый для использования в каждой конкретной области.

**Статистика в пространствах произвольной природы.** Много ли общего у статистических методов анализа данных различной природы? На этот естественный вопрос можно сразу же однозначно ответить – да, очень много. Такой ответ будет постоянно подтверждаться и конкретизироваться на протяжении всего учебника. Несколько примеров приведем сразу же.

Прежде всего отметим, что понятия случайного события, вероятности, независимости событий и случайных величин являются общими для любых конечных вероятностных пространств и любых конечных областей значений случайных величин (см. главы 1.2 и 2.1). Поскольку все реальные явления и процессы описываются с помощью математических объектов из конечных множеств, сказанное выше означает, что конечных вероятностных пространств и дискретных случайных величин (точнее, величин, принимающих значения в конечном множестве) достаточно для всех практических применений. Переход к непрерывным моделям реальных явлений и процессов оправдан только тогда, когда этот переход облегчает проведение рассуждений и выкладок. Например, находить определенные интегралы зачастую проще, чем вычислять значения сумм. Не могу не отметить, что приведенные соображения о взаимосоотнесении дискретных и непрерывных математических моделей автор услышал более 30 лет назад от академика А.Н.Колмогорова (ясно, что за конкретную формулировку несет ответственность автор настоящего учебника).

Основные проблемы прикладной статистики – описание данных, оценивание, проверка гипотез – также в своей существенной части могут быть рассмотрены в рамках статистики в пространствах произвольной природы. Например, для описания данных могут быть использованы эмпирические и теоретические средние, плотности вероятностей и их непараметрические оценки, регрессионные зависимости. Правда, для этого пространства произвольной природы должны быть снабжены соответствующим математическим инструментарием – расстояниями (показателями близости, мерами различия) между элементами рассматриваемых пространств.

Популярный в настоящее время метод оценивания параметров распределений – метод максимального правдоподобия – не накладывает каких-либо ограничений на конкретный вид

элементов выборки. Они могут лежать в пространстве произвольной природы. Математические условия касаются только свойств плотностей вероятности и их производных по параметрам. Аналогично положение с методом одношаговых оценок, идущим на смену методу максимального правдоподобия (см. главу 2.2). Асимптотику решений экстремальных статистических задач достаточно изучить для пространств произвольной природы, а затем применять в каждом конкретном случае [18], когда задачу прикладной статистики удается представить в оптимизационном виде. Общая теория проверки статистических гипотез также не требует конкретизации математической природы рассматриваемых элементов выборок. Это относится, например, к лемме Неймана-Пирсона или теории статистических решений. Более того, естественная область построения теории статистик интегрального типа – это пространства произвольной природы (см. главу 2.3).

Совершенно ясно, что в конкретных областях прикладной статистики накоплено большое число результатов, относящимся именно к этим областям. Особенно это касается областей, исследования в которых ведутся сотни лет, в частности, статистики случайных величин (одномерной статистики). Однако принципиально важно указать на «ядро» прикладной статистики – статистику в пространствах произвольной природы. Если постоянно «держать в уме» это ядро, то становится ясно, что, например, многие методы непараметрической оценки плотности вероятности или кластер-анализа, использующие только расстояния между объектами и элементами выборки, относятся именно к статистике объектов произвольной природы, а не к статистике случайных величин или многомерному статистическому анализу. Следовательно, и применяться они могут во всех областях прикладной статистики, а не только в тех, в которых «родились».

**Расстояния (метрики).** В пространствах произвольной природы нет операции сложения, поэтому статистические процедуры не могут быть основаны на использовании сумм. Поэтому используется другой математический инструментарий, использующий понятия типа расстояния.

Как известно, расстоянием в пространстве $X$ называется числовая функция двух переменных $d(x,y)$, $x \in X$, $y \in X$, определенная на этом пространстве, т.е. в стандартных обозначениях $d$: $X^2 \to R^1$, где $R^1$ – прямая, т.е. множество всех действительных чисел. Эта функция должна удовлетворять трем условиям (иногда их называют аксиомами):

1) неотрицательности: $d(x,y) \geq 0$, причем $d(x,x) = 0$, для любых значений $x \in X$, $y \in X$;

2) симметричности: $d(x,y) = d(y,x)$ для любых $x \in X$, $y \in X$;

3) неравенства треугольника: $d(x,y) + d(y,z) \geq d(x,z)$ для любых значений $x \in X$, $y \in X$, $z \in X$.

Для термина «расстояние» часто используется синоним – «метрика».

*Пример 1.* Если $d(x,x) = 0$ и $d(x,y) = 1$ при $x \neq y$ для любых значений $x \in X$, $y \in X$, то, как легко проверить, функция $d(x,y)$ – расстояние (метрика). Такое расстояние естественно использовать в пространстве $X$ значений номинального признака: если два значения (например, названные двумя экспертами) совпадают, то расстояние равно 0, а если различны – то 1.

*Пример 2.* Расстояние, используемое в геометрии, очевидно, удовлетворяет трем приведенным выше аксиомам. Если $X$ – это плоскость, а $x(1)$ и $x(2)$ – координаты точки $x \in X$ в некоторой прямоугольной системе координат, то эту точку естественно отождествить с двумерным вектором $(x(1), x(2))$. Тогда расстояние между точками $x = (x(1), x(2))$ и $y = (y(1), y(2))$ согласно известной формуле аналитической геометрии равно

$$d(x,y) = \sqrt{(x(1) - y(1))^2 + (x(2) - y(2))^2}.$$

*Пример 3*. Евклидовым расстоянием в пространстве $R^k$ векторов вида $x = (x(1), x(2), \ldots, x(k))$ и $y = (y(1), y(2), \ldots, y(k))$ размерности $k$ называется

$$d(x,y) = \left( \sum_{j=1}^{k} (x(j) - y(j))^2 \right)^{1/2}.$$

В примере 2 рассмотрен частный случай примера 3 с $k = 2$.

*Пример 4.* В пространстве $R^k$ векторов размерности $k$ используют также так называемое «блочное расстояние», имеющее вид

$$d(x,y)=\sum_{j=1}^{k}\left|x(j)-y(j)\right|.$$

Блочное расстояние соответствует передвижению по городу, разбитому на кварталы горизонтальными и вертикальными улицами. В результате можно передвигаться только параллельно одной из осей координат.

*Пример 5.* В пространстве функций, элементами которого являются функции $x = x(t)$, $y = y(t)$, $0 \le t \le 1$, часто используют расстояние Колмогорова

$$d(x,y)=\sup_{0\le t\le 1}\left|x(t)-y(t)\right|.$$

*Пример 6.* Пространство функций, элементами которого являются функции $x = x(t)$, $y = y(t)$, $0 \le t \le 1$, превращают в метрическое пространство (т.е. в пространство с метрикой), вводя расстояние

$$d_p(x,y)=\left(\int_0^1 (x(t)-y(t))^p\,dt\right)^{1/p}.$$

Это пространство обычно обозначают $L^p$, где параметр $p \ge 1$ (при $p < 1$ не выполняются аксиомы метрического пространства, в частности, аксиома треугольника).

*Пример 7.* Рассмотрим пространство квадратных матриц порядка $k$. Как ввести расстояние между матрицами $A = \|a(i,j)\|$ и $B = \|b(i,j)\|$? Можно сложить расстояния между соответствующими элементами матриц:

$$d(A,B)=\sum_{i=1}^{k}\sum_{j=1}^{k}\left|a(i,j)-b(i,j)\right|.$$

*Пример 8.* Предыдущий пример наводит на мысль о следующем полезном свойстве расстояний. Если на некотором пространстве определены два или больше расстояний, то их сумма – также расстояние.

*Пример 9*. Пусть $A$ и $B$ – множества. Расстояние между множествами можно определить формулой

$$d(A,B)=\mu(A\Delta B).$$

Здесь м – мера на рассматриваемом пространстве множеств, Д – символ симметрической разности множеств,

$$A\Delta B=(A\setminus B)\bigcup(B\setminus A).$$

Если мера – так называемая считающая, т.е. приписывающая единичный вес каждому элементу множества, то введенное расстояние есть число несовпадающих элементов в множествах $A$ и $B$.

*Пример 10.* Между множествами можно ввести и другое расстояние:

$$d_1(A,B)=\frac{\mu(A\Delta B)}{\mu(A\cup B)}.$$

В ряде задач прикладной статистики используются функции двух переменных, для которых выполнены не все три аксиомы расстояния, а только некоторые. Их обычно называют показателями различия, поскольку чем больше различаются объекты, тем больше значение функции. Иногда в том же смысле используют термин «мера близости». Он менее удачен, поскольку большее значение функции соответствует меньшей близости.

Чаще всего отказываются от аксиомы, требующей выполнения неравенства треугольника, поскольку это требование не всегда находит обоснование в конкретной прикладной ситуации.

*Пример 11.* В конечномерном векторном пространстве показателем различия является

$$d(x,y)=\sum_{j=1}^{k}(x(j)-y(j))^2$$

(сравните с примером 3).

Показателями различия, но не расстояниями являются такие популярные в прикладной статистике показатели, как дисперсия или средний квадрат ошибки при оценивании.

Иногда отказываются также и от аксиомы симметричности.

*Пример 12.* Показателем различия чисел $x$ и $y$ является

$$d(x,y)=\left|\frac{x}{y}-1\right|.$$

Такой показатель различия используют в ряде процедур экспертного оценивания.

Что же касается первой аксиомы расстояния, то в различных постановках прикладной статистики ее обычно принимают. Вполне естественно, что наименьший показатель различия должен достигаться, причем именно на совпадающих объектах. Имеет ли смысл это наименьшее значение делать отличным от 0? Вряд ли, поскольку всегда можно добавить одну и ту же константу ко всем значениям показателя различия и тем самым добиться выполнения первой аксиомы.

В прикладной статистике используются самые разные расстояния и показатели различия, о них пойдет речь в соответствующих разделах учебника.

### 1.1.6. Аксиоматическое введение расстояний

В прикладной статистике используют большое количество метрик и показателей различия (см. примеры в предыдущем пункте). Как обоснованно выбрать то или иное расстояние для использования в конкретной задаче? В 1959 г. американский статистик Джон Кемени предложил использовать аксиоматический подход, согласно которому следует сформулировать естественные для конкретной задачи аксиомы и вывести из них вид метрики. Этот подход получил большую популярность в нашей стране после выхода в 1972 г. перевода на русский язык книги Дж. Кемени и Дж. Снелла [19], в которой дана система аксиом для расстояния Кемени между упорядочениями. (Упорядочения, как и иные бинарные отношения, естественно представить в виде квадратных матриц из 0 и 1; тогда расстояние Кемени – это расстояние из примера 7 предыдущего пункта.) Последовала большая серия работ, в которых из тех или иных систем аксиом выводился вил метрики или показателя различия для различных видов данных, прежде всего для объектов нечисловой природы. Многие полученные результаты описаны в обзоре [20], содержащем 161 ссылку, в том числе 69 на русском языке. Рассмотрим некоторые из них.

**Аксиоматическое введение расстояния между толерантностями.** Толерантность - это бинарное отношение, являющееся рефлексивным и симметричным. Его обычно используют для описания отношения сходства между реальными объектами, отношений знакомства или дружбы между людьми. От отношения эквивалентности отличается тем, что свойство транзитивности не предполагается обязательно выполненным. Действительно, Иванов может быть знаком с Петровым, Петров – с Сидоровым, но при этом ничего необычного нет в том, что Иванов и Сидоров не знакомы между собой.

Пусть множество $X$, на котором определено отношение толерантности, состоит из конечного числа элементов: $X = \{x_1, x_2,\ldots, x_k\}$. Тогда толерантность описывается квадратной матрицей $A = \|a(i,j)\|$, $i,j = 1, 2,\ldots, k$, такой, что $a(i,j) = 1$, если $x_i$ и $x_j$ связаны отношением толерантности, и $a(i,j) = 0$ в противном случае. Матрица $A$ симметрична: $a(i,j) = a(j,i)$, на главной диагонали стоят единицы: $a(i,i) = 1$. Любая матрица, удовлетворяющая приведенным в предыдущей фразе условиям, является матрицей, соответствующей некоторому отношению толерантности. Матрице $A$ можно сопоставить неориентированный граф с вершинами в точках $X$: вершины $x_i$ и $x_j$ соединены ребром тогда и только тогда, когда $a(i,j) = 1$. Толерантности используются, в частности, при проведении экспертных исследований (см. пункт 3.4.7 ниже).

Будем говорить, что толерантность $A_3$ лежит между толерантностями $A_1$ и $A_2$, если при всех $i$, $j$ число $a_3(i,j)$ лежит между числами $a_1(i,j)$ и $a_2(i,j)$, т.е. выполнены либо неравенства $a_1(i,j) \le a_3(i,j) \le a_2(i,j)$, либо неравенства $a_1(i,j) \ge a_3(i,j) \ge a_2(i,j)$.

*Теорема 1* [2]. Пусть

(I) $d(A_1, A_2)$ – метрика в пространстве толерантностей, определенных на конечном множестве $X = \{x_1, x_2,\ldots, x_k\}$;

(II) $d(A_1, A_3) + d(A_3, A_2) = d(A_1, A_2)$ тогда и только тогда, когда $A_3$ лежит между $A_1$ и $A_2$;

(III) если отношения толерантности $A_1$ и $A_2$ отличаются только на одной паре элементов, т.е. $a_1(i,j) = a_2(i,j)$ при $(i,j) \ne (i_0,j_0)$, $i<j$, $i_0 < j_0$, и $a_1(i_0, j_0) \ne a_2(i_0, j_0)$, то $d(A_1, A_2) = 1$.

Тогда

$$d(A_1, A_2) = \sum_{1 \le i < j \le k} | a_1(i,j) - a_2(i,j) | = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} | a_1(i,j) - a_2(i,j) |.$$

Таким образом, расстояние $d(A_1, A_2)$ только постоянным множителем S отличается от расстояния Кемени, введенного в пространстве всех бинарных отношений как расстояние Хемминга между описывающими отношения матрицами из 0 и 1 (см. пример 7 предыдущего пункта). Теорема 1 дает аксиоматическое введение расстояния в пространстве толерантностей. Оказалось, что оно является сужением расстояния Кемени на это пространство. Сам Дж. Кемени дал аналогичную систему аксиом для сужения на пространство упорядочений. Доказательство теоремы 1 вытекает из рассмотрений, связанных с аксиоматическим введением расстояний между множествами, и приводится ниже.

**Мера симметрической разности как расстояние между множествами**. Как известно, бинарное отношение можно рассматривать как подмножество декартова квадрата $X^2$ того множества $X$, на котором оно определено. Поэтому теорему 1 можно рассматривать как аксиоматическое введение расстояния между множествами специального вида. Укажем систему аксиом для расстояния между множествами общего вида, описанного в примере 9 предыдущего пункта.

*Определение 1.* Множество $B$ находится между множествами $A$ и $C$, если $(A \bigcap C) \subseteq B \subseteq (A \bigcup C)$.

С помощью определения 1 в совокупности множеств вводятся геометрические соотношения, использование которых полезно для восприятия рассматриваемых ситуаций.

Расстояние между двумя точками в евклидовом пространстве не изменится, если обе точки сдвинуть на один и тот же вектор. Аналогичное свойство расстояния между множествами сформулируем в виде аксиомы 1. Оно соответствует аксиоме 3 Кемени и Снелла [19, с.22] для расстояний между упорядочениями.

*Аксиома 1.* Если $A \bigcap C = B \bigcap C = \varnothing$, то $d(A,B) = d(A \bigcup C, B \bigcup C)$.

*Определение 2.* Непустая система множеств называется кольцом, если для любых двух входящих в нее множеств в эту систему входят их объединение, пересечение и разность. Множество $X$ называется единицей системы множеств, если оно входит в эту систему, а все остальные множества системы являются подмножествами $X$. Кольцо множеств, содержащее единицу, называется алгеброй множеств [21, с.38].

*Теорема 2.* Пусть $W$ - алгебра множеств, $d$: $W^2 \to R^1$. Тогда аксиома 1 эквивалентна следующему условию: $d(A,B) = d(A \setminus B, B \setminus A)$ для любых $A$, $B$ є $W$.

*Доказательство.* Поскольку

$$(A \setminus B) \bigcap (A \bigcap B) = \varnothing, \ (B \setminus A) \notin \bigcap (A \bigcap B) = \varnothing,$$

то равенство $d(A,B) = d(A \setminus B, B \setminus A)$ следует из аксиомы 1. Обратное утверждение вытекает из того, что в условиях аксиомы 1

$$(A \bigcup C) \setminus (B \bigcup C) = A \setminus B, \ (B \bigcup C) \setminus (A \bigcup C) = B \setminus A.$$

Теорема 2 доказана.

С целью внести в алгебру множеств $W$ отношение «находиться между», аналогичное используемому при аксиоматическом введении расстояний в пространствах бинарных отношений (см. условие (II) в теореме 1), примем следующую аксиому.

*Аксиома 2.* Если $B$ лежит между $A$ и $C$, то $d(A,B)+d(B,C)= d(A,C)$.

*Определение 3.* Неотрицательная функция м, определенная на алгебре множеств $W$, называется мерой, если для любых двух непересекающихся множеств $A$ и $B$ из $W$ справедливо соотношение

$$\mu(A \bigcup B) = \mu(A) + \mu(B).$$

Понятие меры – это обобщение понятий длины линии, площади фигуры, объема тела.

*Теорема 3.* Пусть $W$ – алгебра множеств, аксиомы 1 и 2 выполнены для функции $d$: $W^2 \to [0;+\infty]$. Функция $d$ симметрична: $d(A,B) = d(B,A)$ для любых $A$ и $B$ из $W$. Тогда существует, и притом единственная, мера м на $W$ такая, что

$$d(A,B) = \mu(A \Delta B) \qquad (1)$$

при всех $A$ и $B$ из $W$, где $A Д B$ – симметрическая разность множеств $A$ и $B$, т.е. $A \Delta B = (A \setminus B) \bigcup (B \setminus A)$.

*Доказательство.* Положим

$$\mu(B) = d(\varnothing, B), \ B \in W. \qquad (2)$$

Покажем, что определенная формулой (2) функция множества м является мерой. Неотрицательность м следует из неотрицательности $d$. Остается доказать аддитивность, т.е. что из $A \bigcap B = \varnothing$ следует, что

$$\mu(A \bigcup B) = \mu(A) + \mu(B), \ A \in W, B \in W. \qquad (3)$$

Поскольку $A$ всегда лежит между $\varnothing$ и $A \bigcup B$, то по аксиоме 2

$$\mu(A \bigcup B) = d(\varnothing, A \bigcup B) = d(\varnothing, A) + d(A, A \bigcup B) = \mu(A) + d(A, A \bigcup B). \quad (4)$$

Если $A \cap B = \varnothing$, то по аксиоме 1 $d(\varnothing, B) = d(A, A \bigcup B)$, откуда с учетом (4) и следует (3).

Докажем соотношение (1). Поскольку $A \setminus B$ и $B \setminus A$ имеют пустое пересечение, то согласно определению 1 пустое множество $\varnothing$ лежит между $A \setminus B$ и $B \setminus A$. Поэтому по аксиоме 2

$$d(A \setminus B, B \setminus A) = d(A \setminus B, \varnothing) + d(\varnothing, B \setminus A).$$

Из симметричности и соотношения (2) следует, что

$$d(A \setminus B, \varnothing) = d(\varnothing, A \setminus B) = \mu(A \setminus B),$$

откуда $d(A \setminus B, B \setminus A) =$ м$(A \setminus B)$ + м$(B \setminus A)$. Из соотношения (3) следует, что м$(A \setminus B)$ + м$(B \setminus A)$ = м$(A Д B)$. С другой стороны, по аксиоме 1

$$d(A \setminus B, B \setminus A) = d((A \setminus B) \bigcup (A \bigcap B), (B \setminus A) \bigcup (A \bigcap B)) = d(A, B).$$

Из трех последних равенств вытекает справедливость равенства (1).

Остается доказать единственность меры м в соотношении (1). Поскольку $A Д B = B$ при $A = \varnothing$, то из (1) следует (2), т.е. однозначность определения меры м = м($d$) по расстоянию $d$. Теорема 3 доказана.

*Теорема 4* (обратная). Пусть м – мера определенная на алгебре множеств W. Тогда функция $d(A,B)$ = м$(A Д B)$ является псевдометрикой, для нее выполнены аксиомы 1 и 2.

*Доказательство.* То, что функция $d$(A,$B$) из (1) задает псевдометрику, хорошо известно (см., например, [22, с.79]). Доказательство аксиомы 2 содержится в [23, с.181-183]. Аксиома 1 следует из того, что условия $A \bigcap C = B \bigcap C = \varnothing$ обеспечивают справедливость соотношений

$$(A \bigcup C) \Delta (B \bigcup C) = ((A \bigcup C) \setminus (B \bigcup C)) \bigcup ((B \bigcup C) \setminus (A \bigcup C)) = (A \setminus B) \bigcup (B \setminus A) = A \Delta B.$$

*Замечание.* Полагая в аксиоме 2 $A = B = C$, получаем, что $d(A,A) + d(A,A) = d(A,A)$, т.е. $d(A,A) = 0$. Согласно теоремам 3 и 4, из условий теоремы 3 следует неравенство треугольника. Таким образом, в теореме 3 действительно приведена система аксиом, определяющая семейство псевдометрик в пространстве множеств.

Обсудим независимость (друг от друга) условий теоремы 3. Отбрасывание неотрицательности функции $d$ приводит к тому, что слово «мера» в теоремах 3 и 4 необходимо заменить на «заряд» [21, с.328]. Этот термин обозначает аддитивную функцию множеств, не обладающую свойством неотрицательности. Заряд можно представить как разность двух мер.

Функция $d_1(A,B) = \sqrt{\mu(A \Delta B)}$ является псевдометрикой, для нее выполнена аксиома 1, но не выполнена аксиома 2, следовательно, ее нельзя представить в виде (1).

Приведем пример системы множеств $W$ и метрики в ней, для которых верна аксиома 2, но не верна аксиома 1, а потому эту метрику нельзя представить в виде (1). Пусть $W$ состоит из множеств $\varnothing, A, B, A \bigcup B,$ причем $A \cap B = \varnothing$, а расстояния таковы:

$$d(\varnothing, A) = d(\varnothing, B) = 1, \ d(A, A \bigcup B) = d(B, A \bigcup B) = d(A, B) = 2, \ d(\varnothing, A \bigcup B) = 3.$$

Если единица $X$ алгебры множеств $W$ конечна, т.е. $X = \{x_1, x_2, \ldots, x_k\}$, то расстояние (1) принимает вид

$$d(A,B) = \sum_{i=1}^{k} \mu_i \mid \chi_A(x_i) - \chi_B(x_i) \mid, \quad (5)$$

где ч$_C$ – индикатор (индикаторная функция) множества $C$, т.е. ч$_C(x)$ = 1, если $x$ є $C$, и ч$_C(x)$ = 0 в противном случае. Как следует из теоремы 3, неотрицательный коэффициент м$_i$ – это мера

одноэлементного множества $\{x_i\}$, а также расстояние этого множества от пустого множества, т.е.

$$\mu_i = \mu(\{x_i\}) = d(\varnothing, \{x_i\}).$$

Если все коэффициенты м$_i$ положительны, то формула (5) определяет метрику, если хотя бы один равен 0, то – псевдометрику, поскольку в таком случае найдутся два различающиеся между собой множества $A$ и B такие, что $d(A,B) = 0$.

Расстояние определяется однозначно, если априори известны коэффициенты м$_i$. В частности, равноправность объектов (элементов единицы алгебры множеств $X$) приводит к м$_i \equiv 1$. Требование равноправности содержится в аксиомах 2 и 4 Кемени [19, с.21-22].

Применим полученные результаты к толерантностям и докажем теорему 1. Совокупность всех толерантностей, определенных на конечном множестве $Y$, естественным образом ассоциируется с совокупностью всех подмножеств множества $X = \{(y_i, y_j), 1 \le i < j \le k\}$. Именно, пара $(y_i, y_j)$ входит в подмножество тогда и только тогда, когда $y_i$ и $y_j$ связаны отношением толерантности. Указанная совокупность подмножеств является алгеброй множеств с единицей $X$. Определение 1 понятия «находиться между» для множеств полностью соответствует ранее данному определению понятия «находиться между» для толерантностей.

*Теорема 5*. Пусть выполнены условия (I) и (II) теоремы 1 и аксиома 1. Тогда существуют числа м$_{ij} > 0$ такие, что

$$d(A,B) = \sum_{1 \le ii < j \le k} \mu_{ij} \mid a(i,j) - b(i,j) \mid. \qquad (6)$$

Для доказательства достаточно сослаться на теорему 3. Поскольку в условии (I) требуется, чтобы функция $d(A,B)$ являлась метрикой, то необходимо м$_{ij} > 0$.

*Теорема 6*. Пусть выполнены условия теоремы 1 и, кроме того, аксиома 1. Тогда верно заключение теоремы 1.

*Доказательство.* Рассмотрим толерантность $A$, для которой $a(i,j) = 1$ при $(i,j) = (i_0,j_0)$ и $a(i,j) = 0$ в противном случае. Согласно условию (III) теоремы 1 $d(\varnothing, A) = 1$, а согласно (6) имеем $d(\varnothing, A) = \mu_{i_0 j_0}$. Следовательно, коэффициент $\mu_{i_0 j_0} = 1$, что и требовалось доказать.

Для окончательного доказательства теоремы 1 осталось избавиться от требования справедливости аксиомы 1.

*Доказательство теоремы 1.* Рассмотрим две толерантности $A$ и $B$ такие, что при представлении их в виде множеств $A \subseteq B$. Это означает, что $a(i, j) \le b(i, j)$ при всех $i, j$. Поскольку $X$ – конечное множество, то существует конечная последовательность толерантностей $A_1, A_2, \ldots, A_m, \ldots, A_t$ такая, что $A_1 = A$, $A_t = B$, $A_1 \subseteq A_2 \subseteq \ldots \subseteq A_m \subseteq \ldots \subseteq A_t$, причем $A_{m+1}$ получается из $A_m$ заменой ровно одного значения $a_m(i_m, j_m) = 0$ на $a_{m+1}(i_m, j_m) = 1$, для $(i, j) \ne (i_m, j_m)$ при этом $a_m(i, j) = a_{m+1}(i, j)$. Тогда $A_m$ находится между $A_{m-1}$ и $A_{m+1}$, следовательно, по условию (II)

$$d(A,B) = d(A_1, A_2) + d(A_2, A_3) + \ldots + d(A_m, A_{m+1}) + \ldots + d(A_{t-1}, A_t).$$

По условию (III) $d(A_m, A_{m+1}) = 1$ при всех $m$, а потому заключение теоремы 1 верно для любых $A$ и $B$ таких, что $A \subseteq B$.

Поскольку $A \cap B$ лежит между $A$ и $B$, то по условию (II)

$$d(A, B) = d(A \cap B, A) + d(A \cap B, B).$$

При этом $A \cap B \subseteq A$ и $A \cap B \subseteq B$. Применяя результат предыдущего абзаца, получаем, заключение теоремы 1 верно всегда.

*Замечание 1*. Таким образом, условие (III) не только дает нормировку, но и заменяет аксиому 1.

*Замечание 2.* Условие (I) теоремы 1 не использовалось в доказательстве, но было приведено в первоначальной публикации [24], чтобы подчеркнуть цель рассуждения. По той же причине оно сохранено в формулировке теоремы 1, хотя в доказательстве удалось без него обойтись. Понадобилась только симметричность функции $d$.

**Аксиоматическое введение метрики в пространстве неотрицательных суммируемых функций.** Рассмотрим пространство $L(E,$ м$)$ неотрицательных суммируемых функций на множестве $E$ с мерой м. Далее в настоящем пункте будем рассматривать только

функции из $L(E, м)$. Интегрирование всюду проводится по пространству $E$ и по мере м. Будем писать $g = h$ или $g \le h$, если указанные соотношения справедливы почти всюду по м на Е (т.е. могут нарушаться лишь на множестве нулевой меры).

Аксиоматически введем расстояние в пространстве $L(E, м)$ (изложение следует работе [25]). Обозначим $M(g, h) = \max(g, h)$ и $m(g, h) = \min(g, h)$. Пусть $D$: $L(E, м)$ Ч $L(E, м) \to R^1$ – тот основной объект изучения, аксиомы для которого будут сейчас сформулированы.

*Аксиома 1*. Если $gh = 0$, $g + h \neq 0$, то $D(g, h) = 1$.

*Аксиома 2*. Если $h \le g$, то $D(g, h) = C \int (g - h)\, dм$, где множитель $C$ не зависит от $h$, т.е. $C = C(g)$.

*Лемма.* Из аксиом 1,2 следует, что для $h \le g \neq 0$

$$D(g,h) = \frac{\int (g-h)d\mu}{\int g d\mu}.$$

Для доказательства заметим, что по аксиоме 1 $D(g, 0) = 1$, а по аксиоме 2 $D(g, 0) = C \int g\, dм$, откуда $C = (\int g\, dм)^{-1}$. Подставляя это соотношение в аксиому 2, получаем заключение леммы.

Требование согласованности расстояния в пространстве $L(E, м)$ с отношением «находиться между» приводит, как и ранее для расстояния $d(A, B)$, к следующей аксиоме.

*Аксиома 3.* Для любых $g$ и $h$ справедливо равенство $D(g, h) = D(M(g, h), g) + D(M(g, h), h)$.

Замечание. В ряде реальных ситуаций естественно считать, что наибольшее расстояние между элементами пространства множеств (которое без ограничения общности можно положить равным 1), т.е. наибольшее несходство, соответствует множествам, не имеющим общих элементов. Расстояние, введенное в теореме 3 (формула (1)), этому условию не удовлетворяет. Поэтому в пространстве множеств была аксиоматически введена [20] так называемая $D$-метрика (от *dissimilarity* (англ.) – несходство), для которого это условие выполнено. Она имеет вид:

$$D(A,B) = \begin{cases} \dfrac{\mu(A \Delta B)}{\mu(A \cup B)}, & \mu(A \cup B) > 0, \\ 0, & \mu(A) = \mu(B) = 0. \end{cases} \qquad (7)$$

Приведенные выше аксиомы являются обобщениями соответствующих аксиом для $D$-метрики в пространстве множеств.

*Теорема 7*. Из аксиом 1-3 следует, что

$$D(g,h) = \begin{cases} \dfrac{\int |g-h|\, d\mu}{\int M(g,h) d\mu}, & g + h \neq 0, \\ 0, & g = h = 0. \end{cases} \qquad (8)$$

*Доказательство.* Поскольку

$$(M(g, h) - g) + (M(g, h) - h) = |g - h|,$$

то заключение теоремы 7 при $g + h \neq 0$ вытекает из леммы и аксиомы 3. Из аксиомы 2 при $g = 0$ следует, что $D(0, 0) = 0$. Легко видеть, что функция $D$, заданная формулой (8), удовлетворяет аксиомам 1-3 и, кроме того, $D(g, h) \le 1$ при любых $g$ и $h$.

*Замечание.* Если $g$ и $h$ – индикаторные функции множеств, то формула (8) переходит в формулу (7). Если $g$ и $h$ – функции принадлежности нечетких множеств, то формула (8) задает метрику в пространстве нечетких множеств, а именно, $D$-метрику в этом пространстве [20].

*Теорема 8.* Функция $D(g,h)$, определенная формулой (8), является метрикой в $L(E, м)$ (при отождествлении функций, отличающихся лишь на множестве нулевой меры), причем $D(g, f) + D(f, h) = D(g, h)$ тогда и только тогда, когда $f = g$, $f = h$ или $f = M(g, h)$.

*Доказательство.* Обратимся к определению метрики. Для рассматриваемой функции непосредственно очевидна справедливость условий неотрицательности и симметричности. Очевидна и эквивалентность условия $D(g, h) = 0$ равенству $g = h$. Остается доказать неравенство треугольника и установить, когда оно обращается в равенство.

Без ограничения общности можно считать, что рассматриваемые расстояния задаются верхней строкой формулы (8) и, кроме того,

$$R = \int M(g, f)\, d\text{м} - \int M(f, h)\, d\text{м} \geq 0$$

(частные случаи с использованием нижней строки формулы (8) рассматриваются элементарно, а справедливости последнего неравенства можно добиться заменой обозначений функций – элементов пространства $L(E, \text{м})$). Тогда

$$D(g, f) + D(f, h) \geq \frac{\int (|g - f| + |f - h|)d\mu}{\int M(g, f)d\mu}, \quad (9)$$

причем равенство имеет место тогда и только тогда, когда $R = 0$ или $f = h$. Положим

$$P = \int (|g - f| + |f - h| - |g - h|)\, d\text{м}, \quad Q = \int (M(g, f) - M(g, h))\, d\text{м}.$$

Ясно, что $P \geq 0$ и

$$\frac{\int (|g - f| + |f - h|)d\mu}{\int M9g, f)d\mu} = \frac{\int |g - h|\, d\mu + P}{\int M(g, h)d\mu + Q}. \quad (10)$$

Если $Q < 0$, то, очевидно, неравенство треугольника выполнено, причем неравенство является строгим. Рассмотрим случай $Q > 0$.

Воспользуемся следующим элементарным фактом: если $y \geq x$, $y > 0$, $P > Q > 0$, то

$$\frac{x + P}{y + Q} > \frac{x}{y}. \quad (11)$$

Из соотношений (10) и (11) вытекает, что для доказательства неравенства треугольника достаточно показать, что $P - Q > 0$.

Рассмотрим

$$k = \{|g - f| + |f - h| - |g - h|\} - M(g, f) + M(g, h).$$

Применяя равенство $(M(g, h) - g) + (M(g, h) - h) = |g - h|$ к слагаемым, заключенным в фигурные скобки, получаем, что

$$k = M(f, h) + [M(g, f) + M(f, h) - M(g, h) - 2f].$$

Применяя соотношение

$$M(g, h) = g + h - m(g, h) \quad (12)$$

К слагаемым, заключенным в квадратные скобки, получаем, что

$$k = M(f, h) - m(f, h) - m(g, f) + m(g, h).$$

Так как $M(f, h) - m(f, h) = |f - h|$, то

$$k = |f - h| - (m(g, f) - m(g, h)) \geq (f - h) - (m(g, f) - m(g, h)). \quad (13)$$

В соответствии с (12) правая часть (13) есть $M(g, f) - M(g, h)$, а потому

$$P - Q = \int k\, d\text{м} \geq Q > 0,$$

что завершает доказательство для случая $Q > 0$. При этом неравенство треугольника является строгим.

Осталось рассмотреть случай $Q = 0$. В силу соотношений (9) и (10) неравенство треугольника выполнено. Когда оно обращается в равенство? Тривиальные случаи: $f = g$ или $f = h$. Если же $f$ отлично от $g$ и $h$, то необходимо, чтобы $R = 0$ и $P = 0$. Как легко проверить, последнее условие эквивалентно неравенствам

$$m(g, h) \leq f \leq M(g, h). \quad (14)$$

Из правого неравенства в (14) следует, что $M(g, f) \leq M(g, M(g, h)) = M(g, h)$. Так как $Q = 0$, то $M(g, f) = M(g, h)$. Аналогичным образом из
соотношений

$$M(h, f) \leq M(h, M(g, h)) = M(g, h) = M(g, f)$$

и $R = 0$ следует, что $M(f, h) = M(g, h)$.

Рассмотрим измеримое множество $X = \{x \in E: h(x) < g(x)\}$. Тогда $M(g, h)(x) = M(f, h)(x) = g(x) > h(x)$, т.е. $h(x) < f(x) = M(g, h)(x)$ для почти всех $x \in X$. Для почти всех $y \in \{x \in E: h(x) > g(x)\}$ точно так же получаем $f(y) = M(g, h)(y)$. Для почти всех $z \in \{x \in E: h(x) = g(x)\}$ в силу (14) $f(z) = M(g, h)(z)$, что и завершает доказательство теоремы.

*Замечание.* Назовем функции $g$ и $h$ подобными, если существует число $b > 0$ такое, что $g = bh$. Тогда при $0 < b \leq 1$ имеем $D(g, h) = 1 - b$, т.е. расстояние между подобными функциями линейно зависит от коэффициента подобия. Далее, пусть $a > 0$, тогда $D(ag, ah) = D(g, h)$. Таким

образом, метрика (8) инвариантна по отношению к преобразованиям подобия, которые образуют группу допустимых преобразований в шкале отношений. Это дает основания именовать метрику (8) метрикой подобия [25].

## Литература


1. Суппес П., Зинес Дж. Основы теории измерений. - В сб.: Психологические измерения. - М.: Мир, 1967. - С. 9-110.
2. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука, 1979. - 296 с.
3. Носовский Г.В., Фоменко А.Т. Империя. Русь, Турция, Китай, Европа, Египет. Новая математическая хронология древности. - М.: Изд-во "Факториал", 1996. - 752 с.
4. Шубкин В.П. Социологические опыты. - М.: Мысль,1970.-256 с.
5. Щукина Г.И. Проблема познавательного интереса в педагогике. - М.: Педагогика, 1971.-352 с.
6. Орлов А.И. Статистика объектов нечисловой природы (Обзор). – Журнал «Заводская лаборатория». 1990. Т.56. No.3. С.76-83.
7. Орлов А.И. Объекты нечисловой природы. – Журнал «Заводская лаборатория». 1995. Т.61. No.3. С.43-52.
8. Кендэл М. Ранговые корреляции. - М.: Статистика, 1975. - 216 с.
9. Беляев Ю.К. Вероятностные методы выборочного контроля. - М.: Наука, 1975. - 408 с.
10. Лумельский Я.П. Статистические оценки результатов контроля качества. - М.: Изд-во стандартов, 1979. - 200 с.
11. Дэвид Г. Метод парных сравнений. - М.: Статистика, 1978.- 144 с.
12. Организация и планирование машиностроительного производства (производственный менеджмент): Учебник / К.А.Грачева, М.К.Захарова, Л.А.Одинцова и др. Под ред. Ю.В.Скворцова, Л.А.Некрасова. – М.: Высшая школа, 2003. – 470 с.
13. Кендалл М.Дж., Стьюарт А., Статистические выводы и связи. М.: Наука, 1973. - 900 с.
14. Себер Дж. Линейный регрессионный анализ. - М.: Мир, 1980. - 456 с.
15. Борель Э. Вероятность и достоверность. - М.: ГИФМЛ, 1961. - 120 с.
16. Орлов А.И. Задачи оптимизации и нечеткие переменные. - М.: Знание, 1980. - 64с.
17. Битюков П.В. Моделирование задач ценообразования на электронные обучающие курсы в области дистанционного обучения / Автореферат диссертации на соискание ученой степени кандидата экономических наук. – М.: Московский государственный университет экономики, статистики и информатики, 2002. – 24 с.
18. Орлов А.И. Асимптотика решений экстремальных статистических задач. – В сб.: Анализ нечисловых данных в системных исследованиях. Сборник трудов. Вып.10. - М.: Всесоюзный научно-исследовательский институт системных исследований, 1982. С. 4-12.
19. Кемени Дж., Снелл Дж. Кибернетическое моделирование. Некоторые приложения. – М.: Советское радио, 1972. – 192 с.
20. Раушенбах Г.В. Меры близости и сходства // Анализ нечисловой информации в социологических исследованиях. – М.; Наука, 1986. – С.169-203.
21. Колмогоров А.Н., Фомин С.В. Элементы теории функций и функционального анализа. – М.: Наука, 1972. – 496 с.
22. Окстоби Дж. Мера и категория. – М.: Мир, 1974. – 158 с.
23. Льюс Р., Галантер Е. Психофизические шкалы // Психологические измерения. – М.: Мир, 1967. – С.111-195.
24. Орлов А.И. Связь между нечеткими и случайными множествами: Нечеткие толерантности // Исследования по вероятностно-статистическому моделированию реальных систем. – М.: ЦЭМИ АН СССР, 1977. – С.140-148.
25. Орлов А.И., Раушенбах Г.В. Метрика подобия: аксиоматическое введение, асимптотическая нормальность // Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. - Пермь: Изд-во Пермского государственного университета, 1986, с.148-157.

## Контрольные вопросы и задачи

1. Приведите примеры практического использования количественных и категоризованных данных.
2. Как соотносятся группы допустимых преобразований для различных шкал измерения?
3. Почему анализ нечисловых данных занимает одно из центральных мест в прикладной статистике?
4. В каких случаях целесообразно применение нечетких множеств?
5. Справедливо ли для нечетких множеств равенство $(A+B)C = AC + BC$? А равенство $(AB)C = (AC)(BC)$?
6. Докажите, что для блочного расстояния (пример 4 из пункта 1.1.5) справедливо неравенство треугольника.
7. Расскажите о многообразии расстояний в различных пространствах статистических данных.
8. Докажите, что если $d(x, y)$ – расстояние в некотором пространстве, то $\sqrt{d(x,y)}$ - также расстояние в этом пространстве.

## Темы докладов, рефератов, исследовательских работ

1. Содержание первого сочинения по прикладной статистике - книге «Числа» в Библии.
2. Свойства основных шкал измерения.
3. Взаимосвязи различных классов объектов нечисловой природы между собой.
4. Опишите с помощью нечеткого подмножества временной шкалы понятие «молодой человек» (на основе опроса 10-20 экспертов).
5. Опишите с помощью теории нечеткости понятие «куча зерен» (на основе опроса 10-20 экспертов).
6. Центральная роль статистики объектов произвольной природы в прикладной статистике.
7. Расстояния в пространствах функций.
8. Докажите, что аксиоматически введенный в п.1.1.6 показатель различия между множествами $d(A, B) = м(AДB)$ удовлетворяет неравенству треугольника.

## 1.2. Основы вероятностно-статистических методов описания неопределенностей в прикладной статистике

### 1.2.1. Теория вероятностей и математическая статистика – научные основы прикладной статистики

**Как используются теория вероятностей и математическая статистика?** Эти дисциплины – основа вероятностно-статистических методов принятия решений. Чтобы воспользоваться их математическим аппаратом, необходимо задачи принятия решений выразить в терминах вероятностно-статистических моделей. Применение конкретного вероятностно-статистического метода принятия решений состоит из трех этапов:

- переход от экономической, управленческой, технологической реальности к абстрактной математико-статистической схеме, т.е. построение вероятностной модели системы управления, технологического процесса, процедуры принятия решений, в частности по результатам статистического контроля, и т.п.

- проведение расчетов и получение выводов чисто математическими средствами в рамках вероятностной модели;

- интерпретация математико-статистических выводов применительно к реальной ситуации и принятие соответствующего решения (например, о соответствии или несоответствии качества продукции установленным требованиям, необходимости наладки технологического процесса и т.п.), в частности, заключения (о доле дефектных единиц продукции в партии, о конкретном виде законов распределения контролируемых параметров технологического процесса и др.).

Математическая статистика использует понятия, методы и результаты теории вероятностей. Рассмотрим основные вопросы построения вероятностных моделей принятия решений в экономических, управленческих, технологических и иных ситуациях. Для активного и правильного использования нормативно-технических и инструктивно-методических документов по вероятностно-статистическим методам принятия решений нужны предварительные знания. Так, необходимо знать, при каких условиях следует применять тот или иной документ, какую исходную информацию необходимо иметь для его выбора и применения, какие решения должны быть приняты по результатам обработки данных и т.д.

**Примеры применения теории вероятностей и математической статистики.** Рассмотрим несколько примеров, когда вероятностно-статистические модели являются хорошим инструментом для решения управленческих, производственных, экономических, народнохозяйственных задач. Так, например, в романе А.Н.Толстого «Хождение по мукам» (т.1) говорится: «мастерская дает двадцать три процента брака, этой цифры вы и держитесь, - сказал Струков Ивану Ильичу».

Встает вопрос, как понимать эти слова в разговоре заводских менеджеров, поскольку одна единица продукции не может быть дефектна на 23%. Она может быть либо годной, либо дефектной. Наверно, Струков имел в виду, что в партии большого объема содержится примерно 23% дефектных единиц продукции. Тогда возникает вопрос, а что значит «примерно»? Пусть из 100 проверенных единиц продукции 30 окажутся дефектными, или из 1000 – 300, или из 100000 – 30000 и т.д., надо ли обвинять Струкова во лжи?

Или другой пример. Монетка, которую используют как жребий, должна быть «симметричной», т.е. при ее бросании в среднем в половине случаев должен выпадать герб, а в половине случаев – решетка (решка, цифра). Но что означает «в среднем»? Если провести много серий по 10 бросаний в каждой серии, то часто будут встречаться серии, в которых монетка 4 раза выпадает гербом. Для симметричной монеты это будет происходить в 20,5% серий. А если на 100000 бросаний окажется 40000 гербов, то можно ли считать монету симметричной? Процедура принятия решений строится на основе теории вероятностей и математической статистики.

Рассматриваемый пример может показаться недостаточно серьезным. Однако это не так. Жеребьевка широко используется при организации промышленных технико-экономических экспериментов, например, при обработке результатов измерения показателя качества (момента трения) подшипников в зависимости от различных технологических факторов (влияния

консервационной среды, методов подготовки подшипников перед измерением, влияния нагрузки подшипников в процессе измерения и т.п.). Допустим, необходимо сравнить качество подшипников в зависимости от результатов хранения их в разных консервационных маслах, т.е. в маслах состава *A* и *B*. При планировании такого эксперимента возникает вопрос, какие подшипники следует поместить в масло состава *A*, а какие – в масло состава *B*, но так, чтобы избежать субъективизма и обеспечить объективность принимаемого решения.

Ответ на этот вопрос может быть получен с помощью жребия. Аналогичный пример можно привести и с контролем качества любой продукции. Чтобы решить, соответствует или не соответствует контролируемая партия продукции установленным требованиям, из нее отбирается выборка. По результатам контроля выборки делается заключение о всей партии. В этом случае очень важно избежать субъективизма при формировании выборки, т.е необходимо, чтобы каждая единица продукции в контролируемой партии имела одинаковую вероятность быть отобранной в выборку. В производственных условиях отбор единиц продукции в выборку обычно осуществляют не с помощью жребия, а по специальным таблицам случайных чисел или с помощью компьютерных датчиков случайных чисел.

Аналогичные проблемы обеспечения объективности сравнения возникают при сопоставлении различных схем организации производства, оплаты труда, при проведении тендеров и конкурсов, подбора кандидатов на вакантные должности и т.п. Всюду нужна жеребьевка или подобные ей процедуры. Поясним на примере выявления наиболее сильной и второй по силе команды при организации турнира по олимпийской системе (проигравший выбывает). Пусть всегда более сильная команда побеждает более слабую. Ясно, что самая сильная команда однозначно станет чемпионом. Вторая по силе команда выйдет в финал тогда и только тогда, когда до финала у нее не будет игр с будущим чемпионом. Если такая игра будет запланирована, то вторая по силе команда в финал не попадет. Тот, кто планирует турнир, может либо досрочно «выбить» вторую по силе команду из турнира, сведя ее в первой же встрече с лидером, либо обеспечить ей второе место, обеспечив встречи с более слабыми командами вплоть до финала. Чтобы избежать субъективизма, проводят жеребьевку. Для турнира из 8 команд вероятность того, что в финале встретятся две самые сильные команды, равна 4/7. Соответственно с вероятностью 3/7 вторая по силе команда покинет турнир досрочно.

При любом измерении единиц продукции (с помощью штангенциркуля, микрометра, амперметра и т.п.) имеются погрешности. Чтобы выяснить, есть ли систематические погрешности, необходимо сделать многократные измерения единицы продукции, характеристики которой известны (например, стандартного образца). При этом следует помнить, что кроме систематической погрешности присутствует и случайная погрешность.

Поэтому встает вопрос, как по результатам измерений узнать, есть л систематическая погрешность. Если отмечать только, является ли полученная при очередном измерении погрешность положительной или отрицательной, то эту задачу можно свести к предыдущей. Действительно, сопоставим измерение с бросанием монеты, положительную погрешность – с выпадением герба, отрицательную – решетки (нулевая погрешность при достаточном числе делений шкалы практически никогда не встречается). Тогда проверка отсутствия систематической погрешности эквивалентна проверке симметричности монеты.

Целью этих рассуждений является сведение задачи проверки отсутствия систематической погрешности к задаче проверки симметричности монеты. Проведенные рассуждения приводят к так называемому «критерию знаков» в математической статистике.

При статистическом регулировании технологических процессов на основе методов математической статистики разрабатываются правила и планы статистического контроля процессов, направленные на своевременное обнаружение разладки технологических процессов и принятия мер к их наладке и предотвращению выпуска продукции, не соответствующей установленным требованиям. Эти меры нацелены на сокращение издержек производства и потерь от поставки некачественных единиц продукции. При статистическом приемочном контроле на основе методов математической статистики разрабатываются планы контроля качества путем анализа выборок из партий продукции. Сложность заключается в том, чтобы уметь правильно строить вероятностно-статистические модели принятия решений, на основе которых можно ответить на поставленные выше вопросы. В математической статистике для этого разработаны вероятностные модели и методы проверки гипотез, в частности, гипотез о

том, что доля дефектных единиц продукции равна определенному числу $p_0$, например, $p_0 = 0{,}23$ (вспомните слова Струкова из романа А.Н.Толстого).

**Задачи оценивания.** В ряде управленческих, производственных, экономических, народнохозяйственных ситуаций возникают задачи другого типа – задачи оценки характеристик и параметров распределений вероятностей.

Рассмотрим пример. Пусть на контроль поступила партия из $N$ электроламп. Из этой партии случайным образом отобрана выборка объемом $n$ электроламп. Возникает ряд естественных вопросов. Как по результатам испытаний элементов выборки определить средний срок службы электроламп и с какой точностью можно оценить эту характеристику? Как изменится точность, если взять выборку большего объема? При каком числе часов $T$ можно гарантировать, что не менее 90% электроламп прослужат $T$ и более часов?

Предположим, что при испытании выборки объемом $n$ электроламп дефектными оказались $X$ электроламп. Тогда возникают следующие вопросы. Какие границы можно указать для числа $D$ дефектных электроламп в партии, для уровня дефектности $D/N$ и т.п.?

Или при статистическом анализе точности и стабильности технологических процессов надлежит оценить такие показатели качества, как среднее значение контролируемого параметра и степень его разброса в рассматриваемом процессе. Согласно теории вероятностей в качестве среднего значения случайной величины целесообразно использовать ее математическое ожидание, а в качестве статистической характеристики разброса – дисперсию, среднее квадратическое отклонение или коэффициент вариации. Отсюда возникает вопрос: как оценить эти статистические характеристики по выборочным данным и с какой точностью это удается сделать? Аналогичных примеров можно привести очень много. Здесь важно было показать, как теория вероятностей и математическая статистика могут быть использованы в производственном менеджменте при принятии решений в области статистического управления качеством продукции.

**Что такое «математическая статистика»?** Под математической статистикой понимают «раздел математики, посвященный математическим методам сбора, систематизации, обработки и интерпретации статистических данных, а также использование их для научных или практических выводов. Правила и процедуры математической статистики опираются на теорию вероятностей, позволяющую оценить точность и надежность выводов, получаемых в каждой задаче на основании имеющегося статистического материала» [1, с.326]. При этом статистическими данными называются сведения о числе объектов в какой-либо более или менее обширной совокупности, обладающих теми или иными признаками.

По типу решаемых задач математическая статистика обычно делится на три раздела: описание данных, оценивание и проверка гипотез.

По виду обрабатываемых статистических данных математическая статистика делится на четыре направления:

- одномерная статистика (статистика случайных величин), в которой результат наблюдения описывается действительным числом;
- многомерный статистический анализ, где результат наблюдения над объектом описывается несколькими числами (вектором);
- статистика случайных процессов и временных рядов, где результат наблюдения – функция;
- статистика объектов нечисловой природы, в которой результат наблюдения имеет нечисловую природу, например, является множеством (геометрической фигурой), упорядочением или получен в результате измерения по качественному признаку.

Исторически первой появились некоторые области статистики объектов нечисловой природы (в частности, задачи оценивания доли брака и проверки гипотез о ней) и одномерная статистика. Математический аппарат для них проще, поэтому на их примере обычно демонстрируют основные идеи математической статистики.

Лишь те методы обработки данных, т.е. математической статистики, являются доказательными, которые опираются на вероятностные модели соответствующих реальных явлений и процессов. Речь идет о моделях поведения потребителей, возникновения рисков, функционирования технологического оборудования, получения результатов эксперимента, течения заболевания и т.п. Вероятностную модель реального явления следует считать

построенной, если рассматриваемые величины и связи между ними выражены в терминах теории вероятностей. Соответствие вероятностной модели реальности, т.е. ее адекватность, обосновывают, в частности, с помощью статистических методов проверки гипотез.

Невероятностные методы обработки данных являются поисковыми, их можно использовать лишь при предварительном анализе данных, так как они не дают возможности оценить точность и надежность выводов, полученных на основании ограниченного статистического материала.

Вероятностные и статистические методы применимы всюду, где удается построить и обосновать вероятностную модель явления или процесса. Их применение обязательно, когда сделанные на основе выборочных данных выводы переносятся на всю совокупность (например, с выборки на всю партию продукции).

В конкретных областях применений используются как вероятностно-статистические методы широкого применения, так и специфические. Например, в разделе производственного менеджмента, посвященного статистическим методам управления качеством продукции, используют прикладную математическую статистику (включая планирование экспериментов). С помощью ее методов проводится статистический анализ точности и стабильности технологических процессов и статистическая оценка качества. К специфическим методам относятся методы статистического приемочного контроля качества продукции, статистического регулирования технологических процессов, оценки и контроля надежности и др.

Широко применяются такие прикладные вероятностно-статистические дисциплины, как теория надежности и теория массового обслуживания. Содержание первой из них ясно из названия, вторая занимается изучением систем типа телефонной станции, на которую в случайные моменты времени поступают вызовы - требования абонентов, набирающих номера на своих телефонных аппаратах. Длительность обслуживания этих требований, т.е. длительность разговоров, также моделируется случайными величинами. Большой вклад в развитие этих дисциплин внесли член-корреспондент АН СССР А.Я. Хинчин (1894-1959), академик АН УССР Б.В.Гнеденко (1912-1995) и другие отечественные ученые.

**Коротко об истории математической статистики.** Математическая статистика как наука начинается с работ знаменитого немецкого математика Карла Фридриха Гаусса (1777-1855), который на основе теории вероятностей исследовал и обосновал метод наименьших квадратов, созданный им в 1795 г. и примененный для обработки астрономических данных (с целью уточнения орбиты малой планеты Церера). Его именем часто называют одно из наиболее популярных распределений вероятностей – нормальное, а в теории случайных процессов основной объект изучения – гауссовские процессы.

В конце XIX в. – начале XX в. крупный вклад в математическую статистику внесли английские исследователи, прежде всего К.Пирсон (1857-1936) и Р.А.Фишер (1890-1962). В частности, Пирсон разработал критерий «хи-квадрат» проверки статистических гипотез, а Фишер – дисперсионный анализ, теорию планирования эксперимента, метод максимального правдоподобия оценки параметров.

В 30-е годы XX в. поляк Ежи Нейман (1894-1977) и англичанин Э.Пирсон развили общую теорию проверки статистических гипотез, а советские математики академик А.Н. Колмогоров (1903-1987) и член-корреспондент АН СССР Н.В.Смирнов (1900-1966) заложили основы непараметрической статистики. В сороковые годы XX в. румын А. Вальд (1902-1950) построил теорию последовательного статистического анализа.

Математическая статистика бурно развивается и в настоящее время. Так, за последние 40 лет можно выделить четыре принципиально новых направления исследований [2]:

- разработка и внедрение математических методов планирования экспериментов;
- развитие статистики объектов нечисловой природы как самостоятельного направления в прикладной математической статистике;
- развитие статистических методов, устойчивых по отношению к малым отклонениям от используемой вероятностной модели;
- широкое развертывание работ по созданию компьютерных пакетов программ, предназначенных для проведения статистического анализа данных.

**Вероятностно-статистические методы и оптимизация.** Идея оптимизации пронизывает современную прикладную математическую статистику и иные статистические

методы. А именно, методы планирования экспериментов, статистического приемочного контроля, статистического регулирования технологических процессов и др. С другой стороны, оптимизационные постановки в теории принятия решений, например, прикладная теория оптимизации качества продукции и требований стандартов, предусматривают широкое использование вероятностно-статистических методов, прежде всего прикладной математической статистики.

В производственном менеджменте, в частности, при оптимизации качества продукции и требований стандартов особенно важно применять статистические методы на начальном этапе жизненного цикла продукции, т.е. на этапе научно-исследовательской подготовки опытно-конструкторских разработок (разработка перспективных требований к продукции, аванпроекта, технического задания на опытно-конструкторскую разработку). Это объясняется ограниченностью информации, доступной на начальном этапе жизненного цикла продукции, и необходимостью прогнозирования технических возможностей и экономической ситуации на будущее. Статистические методы должны применяться на всех этапах решения задачи оптимизации – при шкалировании переменных, разработке математических моделей функционирования изделий и систем, проведении технических и экономических экспериментов и т.д.

В задачах оптимизации, в том числе оптимизации качества продукции и требований стандартов, используют все области статистики. А именно, статистику случайных величин, многомерный статистический анализ, статистику случайных процессов и временных рядов, статистику объектов нечисловой природы. Выбор статистического метода для анализа конкретных данных целесообразно проводить согласно рекомендациям [3].

### 1.2.2. Основы теории вероятностей

Этот раздел содержит полные доказательства всех рассматриваемых утверждений.

**События и вероятности.** Исходное понятие при построении вероятностных моделей в задачах принятия решений – опыт (испытание). Примерами опытов являются проверка качества единицы продукции, бросание трех монет независимо друг от друга и т.д.

Первый шаг при построении вероятностной модели реального явления или процесса – выделение возможных исходов опыта. Их называют элементарными событиями. Обычно считают, что в первом опыте возможны два исхода – «единица продукции годная» и «единица продукции дефектная». Естественно принять, что при бросании монеты осуществляется одно из двух элементарных событий – «выпала решетка (цифра)» и «выпал герб». Таким образом, случаи «монета встала на ребро» или «монету не удалось найти» считаем невозможными.

При бросании трех монет элементарных событий значительно больше. Вот одно из них – «первая монета выпала гербом, вторая – решеткой, третья – снова гербом». Перечислим все элементарные события в этом опыте. Для этого обозначим выпадение герба буквой Г, а решетки – буквой Р. Имеется $2^3=8$ элементарных событий: ГГГ, ГГР, ГРГ, ГРР, РГГ, РГР, РРГ, РРР – в каждой тройке символов первый показывает результат бросания первой модели, второй – второй монеты, третий – третьей монеты.

Совокупность всех возможных исходов опыта, т.е. всех элементарных событий, называется пространством элементарных событий. Вначале мы ограничимся пространством элементарных событий, состоящим из конечного числа элементов.

С математической точки зрения пространство (совокупность) всех элементарных событий, возможных в опыте – это некоторое множество, а элементарные события – его элементы. Однако в теории вероятностей для обозначения используемых понятий по традиции используются свои термины, отличающиеся от терминов теории множеств. В табл. 1 установлено соответствие между терминологическими рядами этих двух математических дисциплин.

Таблица 1.

Соответствие терминов теории вероятностей и теории множеств

| Теория вероятностей | Теория множеств |
|---|---|
| Пространство элементарных событий | Множество |
| Элементарное событие | Элемент этого множества |
| Событие | Подмножество |
| Достоверное событие | Подмножество, совпадающее с множеством |
| Невозможное событие | Пустое подмножество $\varnothing$ |
| Сумма $A+B$ событий $A$ и $B$ | Объединение $A \cup B$ |
| Произведение $AB$ событий $A$ и $B$ | Пересечение $A \cap B$ |
| Событие, противоположное $A$ | Дополнение $A$ |
| События $A$ и $B$ несовместны | $A \cap B$ пусто |
| События $A$ и $B$ совместны | $A \cap B$ не пусто |

Как сложились два параллельных терминологических ряда? Основные понятия теории вероятностей и ее терминология сформировались в XVII-XVIII вв. Теория множеств возникла в конце XIX в. независимо от теории вероятностей и получила распространение в XX в.

Принятый в настоящее время аксиоматический подход к теории вероятностей, разработанный академиком АН СССР А.Н. Колмогоровым (1903-1987), дал возможность развивать эту дисциплину на базе теории множеств и теории меры. Этот подход позволил рассматривать теорию вероятностей и математическую статистику как часть математики, проводить рассуждения на математическом уровне строгости. В частности, было введено четкое различие между частотой и вероятностью, случайная величина стала рассматриваться как функция от элементарного исхода, и т.д. За основу методов статистического анализа данных стало возможным брать вероятностно-статистические модели, сформулированные в математических терминах. В результате удалось четко отделить строгие утверждения от обсуждения философских вопросов случайности, преодолеть подход на основе понятия равновозможности, имеющий ограниченное практическое значение. Наиболее существенно, что после работ А.Н.Колмогорова нет необходимости связывать вероятности тех или иных событий с пределами частот. Так называемые «субъективные вероятности» получили смысл экспертных оценок вероятностей.

После выхода (в 1933 г. на немецком языке и в 1936 г. – на русском) основополагающей монографии [4] аксиоматический подход к теории вероятностей стал общепринятым в научных исследованиях в этой области. Во многом перестроилось преподавание. Повысился научный уровень многих прикладных работ. Однако традиционный подход оказался живучим. Распространены устаревшие и во многом неверные представления о теории вероятностей и математической статистике. Поэтому в настоящей главе рассматриваем основные понятия, подходы, идеи, методы и результаты в этих областях, необходимые для их квалифицированного применения в задачах принятия решений.

В послевоенные годы А.Н.Колмогоров формализовал понятие случайности на основе теории информации [5]. Грубо говоря, числовая последовательность является случайной, если ее нельзя заметно сжать без потери информации. Однако этот подход не был предназначен для использования в прикладных работах и преподавании. Он представляет собой важное методологическое и теоретическое продвижение.

Перейдем к основному понятию теории вероятностей – понятию вероятности события. В методологических терминах можно сказать, что вероятность события является мерой возможности осуществления события. В ряде случаев естественно считать, что вероятность события $A$ – это число, к которому приближается отношение количества осуществлений события $A$ к общему числу всех опытов (т.е. частота осуществления события $A$) – при увеличении числа опытов, проводящихся независимо друг от друга. Иногда можно предсказать это число из соображений равновозможности. Так, при бросании симметричной монеты и герб,

и решетка имеют одинаковые шансы оказаться сверху, а именно, 1 шанс из 2, а потому вероятности выпадения герба и решетки равны 1/2.

Однако этих соображений недостаточно для развития теории. Методологическое определение не дает численных значений. Не все вероятности можно оценивать как пределы частот, и неясно, сколько опытов надо брать. На основе идеи равновозможности можно решить ряд задач, но в большинстве практических ситуаций применить ее нельзя. Например, для оценки вероятности дефектности единицы продукции. Поэтому перейдем к определениям в рамках аксиоматического подхода на базе математической модели, предложенной А.Н.Колмогоровым (1933).

*Определение 1.* Пусть конечное множество $\Omega = \{\omega\}$ является пространством элементарных событий, соответствующим некоторому опыту. Пусть каждому $\omega \in \Omega$ поставлено в соответствие неотрицательное число $P(\omega)$, называемое вероятностью элементарного события $\omega$, причем сумма вероятностей всех элементарных событий равна 1, т.е.

$$\sum_{\omega \in \Omega} P(\omega) = 1. \quad (1)$$

Тогда пара $\{\Omega, P\}$, состоящая из конечного множества $\Omega$ и неотрицательной функции $P$, определенной на $\Omega$ и удовлетворяющей условию (1), называется *вероятностным пространством*. Вероятность события $A$ равна сумме вероятностей элементарных событий, входящих в А, т.е. определяется равенством

$$P(A) = \sum_{\omega \in A} P(\omega). \quad (2)$$

Сконструирован математический объект, основной при построении вероятностных моделей. Рассмотрим примеры.

*Пример 1.* Бросанию монеты соответствует вероятностное пространство с $\Omega$ = {Г, Р} и $P$(Г) = $P$(Р) = S; здесь обозначено: Г – выпал герб, Р – выпала решетка.

*Пример 2.* Проверке качества одной единицы продукции (в ситуации, описанной в романе А.Н.Толстого «Хождение по мукам» - см. выше) соответствует вероятностное пространство с $\Omega$ = {Б, Г} и Р(Б) = 0,23, Р(Г) = 0,77; здесь обозначено: Б - дефектная единица продукции, Г – годная единица продукции; значение вероятности 0,23 взято из слов Струкова.

Отметим, что приведенное выше определение вероятности *P(A)* согласуется с интуитивным представлением о связи вероятностей события и входящих в него элементарных событий, а также с распространенным мнением, согласно которому «вероятность события $A$ – число от 0 до 1, которое представляет собой предел частоты реализации события $A$ при неограниченном числе повторений одного и того же комплекса условий».

Из определения вероятности события, свойств символа суммирования и равенства (1) вытекает, что

$$а) P(\Omega) = 1, \ б) P(\varnothing) = 0, \ в) P(A + B) = P(A) + P(B) - P(AB). \quad (3)$$

Для несовместных событий А и В согласно формуле (3) $P(A+B) = P(A)+P(B)$. Последнее утверждение называют также теоремой сложения вероятностей.

При практическом применении вероятностно-статистических методов принятия решений постоянно используется понятие независимости. Например, при применении статистических методов управления качеством продукции говорят о независимых измерениях значений контролируемых параметров у включенных в выборку единиц продукции, о независимости появления дефектов одного вида от появления дефектов другого вида, и т.д. Независимость случайных событий понимается в вероятностных моделях в следующем смысле.

*Определение 2.* События $A$ и $B$ называются независимыми, если $P(AB) = P(A)\ P(B)$. Несколько событий $A$, $B$, $C$,… называются независимыми, если вероятность их совместного осуществления равна произведению вероятностей осуществления каждого из них в отдельности: $P(ABC\ldots) = P(A)P(B)P(C)\ldots$

Это определение соответствует интуитивному представлению о независимости: осуществление или неосуществление одного события не должно влиять на осуществление или неосуществление другого. Иногда соотношение $P(AB) = P(A)\ P(B|A) = P(B)P(A|B)$, справедливое при $P(A)P(\text{В}) > 0$, называют также теоремой умножения вероятностей.

*Утверждение 1.* Пусть события $A$ и $B$ независимы. Тогда события $\overline{A}$ и $\overline{B}$ независимы, события $\overline{A}$ и $B$ независимы, события $A$ и $\overline{B}$ независимы (здесь $\overline{A}$ - событие, противоположное $A$, и $\overline{B}$ - событие, противоположное $B$).

Действительно, из свойства в) в (3) следует, что для событий $C$ и $D$, произведение которых пусто, $P(C+D) = P(C) + P(D)$. Поскольку пересечение $AB$ и $\overline{A}\,B$ пусто, а объединение есть $B$, то $P(AB) + P(\overline{A}\,B) = P(B)$. Так как А и В независимы, то $P(\overline{A}\,B) = P(B) - P(AB) = P(B) - P(A)P(B) = P(B)(1 - P(A))$. Заметим теперь, что из соотношений (1) и (2) следует, что $P(\overline{A}) = 1 - P(A)$. Значит, $P(\overline{A}\,B) = P(\overline{A})P(B)$.

Вывод равенства $P(A\,\overline{B}) = P(A)P(\overline{B})$ отличается от предыдущего лишь заменой всюду $A$ на $B$, а $B$ на $A$.

Для доказательства независимости $\overline{A}$ и $\overline{B}$ воспользуемся тем, что события $AB$, $\overline{A}\,B$, $A\,\overline{B}$, $\overline{A}\;\overline{B}$ не имеют попарно общих элементов, а в сумме составляют все пространство элементарных событий. Следовательно, $P(AB) + P(\overline{A}\,B) + P(A\,\overline{B}) + P(\overline{A}\;\overline{B}) = 1$. Воспользовавшись ранее доказанными соотношениями, получаем, что $P(\overline{A}\,B) = 1 - P(AB) - P(B)(1 - P(A)) - P(A)(1 - P(B)) = (1 - P(A))(1 - P(B)) = P(\overline{A})P(\overline{B})$, что и требовалось доказать.

*Пример 3.* Рассмотрим опыт, состоящий в бросании игрального кубика, на гранях которого написаны числа 1, 2, 3, 4, 5,6. Считаем, что все грани имеют одинаковые шансы оказаться наверху. Построим соответствующее вероятностное пространство. Покажем, что события «наверху – грань с четным номером» и «наверху – грань с числом, делящимся на 3» являются независимыми.

*Разбор примера.* Пространство элементарных исходов состоит из 6 элементов: «наверху – грань с 1», «наверху – грань с 2»,…, «наверху – грань с 6». Событие «наверху – грань с четным номером» состоит из трех элементарных событий – когда наверху оказывается 2, 4 или 6. Событие «наверху – грань с числом, делящимся на 3» состоит из двух элементарных событий – когда наверху оказывается 3 или 6. Поскольку все грани имеют одинаковые шансы оказаться наверху, то все элементарные события должны иметь одинаковую вероятность. Поскольку всего имеется 6 элементарных событий, то каждое из них имеет вероятность 1/6. По определению 1событие «наверху – грань с четным номером» имеет вероятность S, а событие «наверху – грань с числом, делящимся на 3» - вероятность 1/3. Произведение этих событий состоит из одного элементарного события «наверху – грань с 6», а потому имеет вероятность 1/6. Поскольку 1/6 = S x 1/3, то рассматриваемые события являются независимыми в соответствии с определением независимости.

В вероятностных моделях процедур принятия решений с помощью понятия независимости событий можно придать точный смысл понятию «независимые испытания». Для этого рассмотрим сложный опыт, состоящий в проведении двух испытаний. Эти испытания называются независимыми, если любые два события $A$ и $B$, из которых $A$ определяется по исходу первого испытания, а $B$ – по исходу второго, являются независимыми.

*Пример 4.* Опишем вероятностное пространство, соответствующее бросанию двух монет независимо друг от друга.

*Разбор примера.* Пространство элементарных событий состоит из четырех элементов: ГГ, ГР, РГ, РР (запись ГГ означает, что первая монета выпала гербом и вторая – тоже гербом; запись РГ – первая – решеткой, а вторая – гербом, и т.д.). Поскольку события «первая монета выпала решеткой» и «вторая монета выпала гербом» являются независимыми по определению независимых испытаний и вероятность каждого из них равна S, то вероятность РГ равна j. Аналогично вероятность каждого из остальных элементарных событий также равна j.

*Пример 5.* Опишем вероятностное пространство, соответствующее проверке качества двух единиц продукции независимо друг от друга, если вероятность дефектности равна $x$.

*Разбор примера.* Пространство элементарных событий состоит из четырех элементов:

$\omega_1$ - обе единицы продукции годны;

$\omega_2$ - первая единица продукции годна, а вторая – дефектна;

$\omega_3$ - первая единица продукции дефектна, а вторая – годна;

$\omega_4$ - обе единицы продукции являются дефектными.

Вероятность того, что единица продукции дефектна, есть $x$, а потому вероятность того, что имеет место противоположное событие, т.е. единица продукции годна, есть $1 - x$. Поскольку результат проверки первой единицы продукции не зависит от такового для второй, то $P(\omega_1) = (1-x)^2$, $P(\omega_2) = P(\omega_3) = x(1-x)$, $P(\omega_4) = x^2$.

**Замечание об условных вероятностях.** В некоторых задачах прикладной статистики оказывается полезным такое понятие, как условная вероятность $P(B|A)$ – вероятность осуществления события $B$ при условии, что событие $A$ произошло. При $P(A)>0$ по определению

$$P(B \mid A) = \frac{P(AB)}{P(A)}.$$

Для независимых событий $A$ и $B$, очевидно, $P(B|A)= P(B)$. Это равенство эквивалентно определению независимости. Понятия условной вероятности и независимости введены А.Муавром в 1718 г.

Необходимо иметь в виду, что для независимости в совокупности нескольких событий недостаточно их попарной независимости. Рассмотрим классический пример [6, с.46]. Пусть одна грань тетраэдра окрашена в красный цвет, вторая - в зеленый. Третья грань окрашена в синий цвет и четвертая – во все эти три цвета. Пусть событие $A$ состоит в том, что грань, на которую упал тетраэдр при бросании, окрашена красным (полностью или частично), событие $B$ – зеленым, событие $C$ – синим. Пусть при бросании все четыре грани тетраэдра имеют одинаковые шансы оказаться внизу. Поскольку граней четыре и две из них имеют в окраске красный цвет, то $P(A) = 1/2$. Легко подсчитать, что

$$P(B) = P(C) = P(A|B) = P(B|C) = P(C|A) = P(B|A) = P(C|A) = P(A|C) = \text{S}.$$

События $A$, $B$ и $C$, таким образом, попарно независимы. Однако если известно, что осуществились одновременно события $B$ и $C$, то это значит, что тетраэдр встал на грань, содержащую все три цвета, т.е. осуществилось и событие $A$. Следовательно, $P(ABC) = \text{j}$, в то время как для независимых событий должно быть $P(A)P(B)P(C) = 1/8$. Следовательно, события $A$, $B$ и $C$ в совокупности зависимы, хотя попарно независимы.

Предположим, что событие $B$ может осуществиться с одним и только с одним из n попарно несовместных событий $A_1, A_2,\ldots, A_k$. Тогда

$$B = \sum_{j=1}^{k} BA_j,$$

где события $BA_i$ и $BA_j$ с разными индексами $i$ и $j$ несовместны. По теореме сложения вероятностей

$$P(B) = \sum_{j=1}^{k} P(BA_j).$$

Воспользовавшись теоремой умножения, находим, что

$$P(B) = \sum_{j=1}^{k} P(A_j)P(B \mid A_j).$$

Получена т.н. «формула полной вероятности». Она широко использовалась математиками при конкретных расчетах еще в начале 18 века, но впервые была сформулирована как одно из основных утверждений теории вероятностей П.Лапласом лишь в конце этого века. Ниже она применяется, в частности, при нахождении среднего выходного уровня дефектности в задачах статистического обеспечения качества продукции.

Применим формулу полных вероятностей для вывода т.н. «формул Байеса», которые иногда используют при проверке статистических гипотез. Требуется найти вероятность события $A_i$, если известно, что событие $B$ произошло. Согласно теореме умножения

$$P(A_iB) = P(B)P(A_i|B) = P(A_i)\, P(B|A_i).$$

Следовательно,

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(B)}.$$

Используя формулу полной вероятности для знаменателя, находим, что

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{\sum_{j=1}^{k} P(A_j)P(B \mid A_j)}.$$

Две последние формулы и называют обычно формулами Байеса. Общая схема их использования такова. Пусть событие $B$ может протекать в различных условиях, относительно которых может быть сделано $k$ гипотез $A_1$, $A_2$,…, $A_k$. Априорные (от *a priori* (лат.) – до опыта) вероятности этих гипотез есть $P(A_1)$, $P(A_2)$,…, $P(A_k)$. Известно также, что при справедливости гипотезы $A_i$ вероятность осуществления события $B$ равна $P(B|A_i)$. Произведен опыт, в результате которого событие $B$ наступило. Естественно после этого уточнить оценки вероятностей гипотез. Апостериорные (от *a posteriori* (лат.) – на основе опыта) оценки вероятностей гипотез $P(A_1|B)$, $P(A_2|B)$,…, $P(A_k|B)$ даются формулами Байеса. В прикладной статистике существует направление «байесовская статистика», в которой, в частности, на основе априорного распределения параметров после проведения измерений, наблюдений, испытаний, опытов анализов вычисляют уточненные оценки параметров.

**Случайные величины и их математические ожидания.** Случайная величина – это величина, значение которой зависит от случая, т.е. от элементарного события $\omega$. Таким образом, случайная величина – это функция, определенная на пространстве элементарных событий $\Omega$. Примеры случайных величин: количество гербов, выпавших при независимом бросании двух монет; число, выпавшее на верхней грани игрального кубика; число дефектных единиц продукции среди проверенных.

Определение случайной величины $X$ как функции от элементарного события $\omega$, т.е. функции $X : \Omega \to H$, отображающей пространство элементарных событий $\Omega$ в некоторое множество $H$, казалось бы, содержит в себе противоречие. О чем идет речь – о величине или о функции? Дело в том, что наблюдается всегда лишь т.н. «реализация случайной величины», т.е. ее значение, соответствующее именно тому элементарному исходу опыта (элементарному событию), которое осуществилось в конкретной реальной ситуации. Т.е. наблюдается именно «величина». А функция от элементарного события – это теоретическое понятие, основа вероятностной модели реального явления или процесса.

Отметим, что элементы $H$ – это не обязательно числа. Ими могут быть и последовательности чисел (вектора), и функции, и математические объекты иной природы, в частности, нечисловой (упорядочения и другие бинарные отношения, множества, нечеткие множества и др.) [2]. Однако наиболее часто рассматриваются вероятностные модели, в которых элементы $H$ – числа, т.е. $H = R^1$. В иных случаях обычно используют термины «случайный вектор», «случайное множество», «случайное упорядочение», «случайный элемент» и др.

Рассмотрим случайную величину с числовыми значениями. Часто оказывается полезным связать с этой функцией число – ее «среднее значение» или, как говорят, «среднюю величину», «показатель центральной тенденции». По ряду причин, некоторые из которых будут ясны из дальнейшего, в качестве «среднего значения» обычно используют математическое ожидание.

*Определение 3.* Математическим ожиданием случайной величины $X$ называется число

$$M(X) = \sum_{\omega \in \Omega} X(\omega)P(\omega), \quad (4)$$

т.е. математическое ожидание случайной величины – это взвешенная сумма значений случайной величины с весами, равными вероятностям соответствующих элементарных событий.

*Пример 6.* Вычислим математическое ожидание числа, выпавшего на верхней грани игрального кубика. Непосредственно из определения 3 следует, что

$$M(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3{,}5.$$

*Утверждение 2.* Пусть случайная величина $X$ принимает значения $x_1, x_2, \ldots, x_m$. Тогда справедливо равенство

$$M(X) = \sum_{1 \le i \le m} x_i P(X = x_i), \quad (5)$$

т.е. математическое ожидание случайной величины – это взвешенная сумма значений случайной величины с весами, равными вероятностям того, что случайная величина принимает определенные значения.

В отличие от (4), где суммирование проводится непосредственно по элементарным событиям, случайное событие $\{X = x_i\} = \{\omega : X(\omega) = x_i\}$ может состоять из нескольких элементарных событий.

Иногда соотношение (5) принимают как определение математического ожидания. Однако с помощью определения 3, как показано далее, более легко установить свойства математического ожидания, нужные для построения вероятностных моделей реальных явлений, чем с помощью соотношения (5).

Для доказательства соотношения (5) сгруппируем в (4) члены с одинаковыми значениями случайной величины $X(\omega)$:

$$M(X) = \sum_{1 \le i \le m} \left( \sum_{\omega : X(\omega) = x_i} X(\omega) P(\omega) \right).$$

Поскольку постоянный множитель можно вынести за знак суммы, то

$$\sum_{\omega : X(\omega) = x_i} X(\omega) P(\omega) = \sum_{\omega : X(\omega) = x_i} x_i P(\omega) = x_i \sum_{\omega : X(\omega) = x_i} P(\omega).$$

По определению вероятности события

$$\sum_{\omega : X(\omega) = x_i} P(\omega) = P(X = x_i).$$

С помощью двух последних соотношений получаем требуемое:

$$M(X) = \sum_{1 \le i \le m} \left( x_i \sum_{\omega : X(\omega) = x_i} P(\omega) \right) = \sum_{1 \le i \le m} \left( x_i P(X = x_i) \right).$$

Понятие математического ожидания в вероятностно-статистической теории соответствует понятию центра тяжести в механике. Поместим в точки $x_1, x_2, \ldots, x_m$ на числовой оси массы $P(X=x_1), P(X=x_2), \ldots, P(X=x_m)$ соответственно. Тогда равенство (5) показывает, что центр тяжести этой системы материальных точек совпадает с математическим ожиданием, что показывает естественность определения 3.

*Утверждение 3.* Пусть $X$ – случайная величина, $M(X)$ – ее математическое ожидание, $a$ – некоторое число. Тогда

$$1)\ M(a)=a;\ 2)\ M(X-M(X))=0;\ 3)\ M[(X-a)^2]=M[(X-M(X))^2]+(a-M(X))^2.$$

Для доказательства рассмотрим сначала случайную величину, являющуюся постоянной, $X(\omega) = a$, т.е. функция $X(\omega)$ отображает пространство элементарных событий $\Omega$ в единственную точку $a$. Поскольку постоянный множитель можно выносить за знак суммы, то

$$M(X) = \sum_{\omega \in \Omega} aP(\omega) = a \sum_{\omega \in \Omega} P(\omega) = a.$$

Если каждый член суммы разбивается на два слагаемых, то и вся сумма разбивается на две суммы, из которых первая составлена из первых слагаемых, а вторая – из вторых. Следовательно, математическое ожидание суммы двух случайных величин $X+Y$, определенных на одном и том же пространстве элементарных событий, равно сумме математических ожиданий $M(X)$ и $M(Y)$ этих случайных величин:

$$M(X+Y) = M(X) + M(Y).$$

А потому $M(X-M(X)) = M(X) - M(M(X))$. Как показано выше, $M(M(X)) = M(X)$. Следовательно, $M(X-M(X)) = M(X) - M(X) = 0$.

Поскольку $(X - a)^2 = \{(X - M(X)) + (M(X) - a)\}^2 = (X - M(X))^2 + 2(X - M(X))(M(X) - a) + (M(X) - a)^2$, то $M[(X - a)^2] = M(X - M(X))^2 + M\{2(X - M(X))(M(X) - a)\} + M[(M(X) - a)^2]$. Упростим последнее равенство. Как показано в начале доказательства утверждения 3, математическое ожидание константы – сама эта константа, а потому $M[(M(X) - a)^2] = (M(X) - a)^2$. Поскольку постоянный множитель можно выносить за знак суммы, то $M\{2(X - M(X))(M(X) - a)\} = 2(M(X) - a)M(X - M(X))$. Правая часть последнего равенства равна 0, поскольку, как показано выше, $M(X-M(X))=0$. Следовательно, $M[(X-a)^2]=M[(X-M(X))^2]+(a-M(X))^2$, что и требовалось доказать.

Из сказанного вытекает, что $M[(X-a)^2]$ достигает минимума по $a$, равного $M[(X-M(X))^2]$, при $a = M(X)$, поскольку второе слагаемое в равенстве 3) всегда неотрицательно и равно 0 только при указанном значении $a$.

*Утверждение 4.* Пусть случайная величина $X$ принимает значения $x_1, x_2, \ldots, x_m$, а f – некоторая функция числового аргумента. Тогда

$$M[f(X)] = \sum_{1 \le i \le m} f(x_i) P(X = x_i).$$

Для доказательства сгруппируем в правой части равенства (4), определяющего математическое ожидание, члены с одинаковыми значениями $X(\omega)$:

$$M[f(X)] = \sum_{1 \le i \le m} \left( \sum_{\omega: X(\omega) = x_i} f(X(\omega)) P(\omega) \right).$$

Пользуясь тем, что постоянный множитель можно выносить за знак суммы, и определением вероятности случайного события (2), получаем

$$M[f(X)] = \sum_{1 \le i \le m} \left( f(x_i) \sum_{\omega: X(\omega) = x_i} P(\omega) \right) = \sum_{1 \le i \le m} f(x_i) P(X = x_i),$$

что и требовалось доказать.

*Утверждение 5.* Пусть $X$ и $Y$ – случайные величины, определенные на одном и том же пространстве элементарных событий, $a$ и $b$ – некоторые числа. Тогда $M(aX+bY)=aM(X)+bM(Y)$.

С помощью определения математического ожидания и свойств символа суммирования получаем цепочку равенств:

$$aM(X) + bM(Y) = a \sum_{\omega \in \Omega} X(\omega) P(\omega) + b \sum_{\omega \in \Omega} Y(\omega) P(\omega) =$$

$$= \sum_{\omega \in \Omega} (aX(\omega) + bY(\omega)) P(\omega) = M(aX + bY).$$

Требуемое доказано.

Выше показано, как зависит математическое ожидание от перехода к другому началу отсчета и к другой единице измерения (переход $Y=aX+b$), а также к функциям от случайных величин. Полученные результаты постоянно используются в технико-экономическом анализе, при оценке финансово-хозяйственной деятельности предприятия, при переходе от одной валюты к другой во внешнеэкономических расчетах, в нормативно-технической документации и др. Рассматриваемые результаты позволяют применять одни и те же расчетные формулы при различных параметрах масштаба и сдвига.

**Независимость случайных величин** – одно из базовых понятий теории вероятностей, лежащее в основе практических всех вероятностно-статистических методов принятия решений.

*Определение 4.* Случайные величины $X$ и $Y$, определенные на одном и том же пространстве элементарных событий, называются независимыми, если для любых чисел $a$ и $b$ независимы события $\{X=a\}$ и $\{Y=b\}$.

*Утверждение 6.* Если случайные величины $X$ и $Y$ независимы, $a$ и $b$ – некоторые числа, то случайные величины $X+a$ и $Y+b$ также независимы.

Действительно, события {X+a=c} и $\{Y+b=d\}$ совпадают с событиями $\{X=c-a\}$ и $\{Y=d-b\}$ соответственно, а потому независимы.

*Пример 7.* Случайные величины, определенные по результатам различных испытаний в схеме независимых испытаний, сами независимы. Это вытекает из того, что события, с помощью которых определяется независимость случайных величин, определяются по результатам различных испытаний, а потому независимы по определению независимых испытаний.

В вероятностно-статистических методах принятия решений постоянно используется следующий факт: если $X$ и $Y$ – независимые случайные величины, $f(X)$ и $g(Y)$ – случайные величины, полученные из $X$ и $Y$ с помощью некоторых функций $f$ и $g$, то $f(X)$ и $g(Y)$ – также независимые случайные величины. Например, если $X$ и $Y$ независимы, то $X^2$ и $2Y+3$ независимы, $\log X$ и $\log Y$ независимы. Доказательство рассматриваемого факта – тема одной из контрольных задач в конце главы.

Подавляющее большинство вероятностно-статистических моделей, используемых на практике, основывается на понятии независимых случайных величин. Так, результаты наблюдений, измерений, испытаний, анализов, опытов обычно моделируются независимыми случайными величинами. Часто считают, что наблюдения проводятся согласно схеме

независимых испытаний. Например, результаты финансово-хозяйственной деятельности предприятий, выработка рабочих, результаты (данные) измерений контролируемого параметра у изделий, отобранных в выборку при статистическом регулировании технологического процесса, ответы потребителей при маркетинговом опросе и другие типы данных, используемых при принятии решений, обычно рассматриваются как независимые случайные величины, вектора или элементы. Причина такой популярности понятия независимости случайных величин состоит в том, что к настоящему времени теория продвинута существенно дальше для независимых случайных величин, чем для зависимых.

Часто используется следующее свойство независимых случайных величин.

*Утверждение 7.* Если случайные величины *X* и *У* независимы, то математическое ожидание произведения ХУ равно произведению математических ожиданий *X* и *У*, т.е. *M(XУ)=M(X)M(У).*

*Доказательство.* Пусть X принимает значения $x_1, x_2, \ldots, x_m$, в то время как *У* принимает значения $y_1, y_2, \ldots, y_k$. Сгруппируем в задающей *M(XУ)* сумме члены, в которых X и *У* принимают фиксированные значения:

$$M(XY) = \sum_{1\le i\le m, 1\le j\le k} \left( \sum_{\omega: X(\omega)=x_i, Y(\omega)=y_j} X(\omega)Y(\omega)P(\omega) \right). \quad (6)$$

Поскольку постоянный множитель можно вынести за знак суммы, то

$$\sum_{\omega: X(\omega)=x_i, Y(\omega)=y_j} X(\omega)Y(\omega)P(\omega) = x_i y_j \sum_{\omega: X(\omega)=x_i, Y(\omega)=y_j} P(\omega).$$

Из последнего равенства и определения вероятности события заключаем, что равенство (6) можно преобразовать к виду

$$M(XY) = \sum_{1\le i\le m, 1\le j\le k} x_i y_j P(X = x_i, Y = y_j).$$

Так как *X* и *У* независимы, то $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$. Воспользовавшись этим равенством и свойством символа суммирования

$$\sum_{1\le i\le m, 1\le j\le k} c_i d_j = \left( \sum_{1\le i\le m} c_i \right)\left( \sum_{1\le j\le k} d_j \right),$$

заключаем, что

$$M(XY) = \left( \sum_{1\le i\le m} x_i P(X = x_i) \right)\left( \sum_{1\le j\le k} y_j P(Y = y_j) \right). \quad (7)$$

Из равенства (5) следует, что первый сомножитель в правой части (7) есть *M(X),* а второй – *M(У),* что и требовалось доказать.

*Пример 8.* Построим пример, показывающий, что из равенства *M(XУ)=M(X)M(У)* не следует независимость случайных величин *X* и *У*. Пусть вероятностное пространство состоит из трех равновероятных элементов $\omega_1, \omega_2, \omega_3$. Пусть

$$X(\omega_1) = 1, X(\omega_2) = 0, X(\omega_3) = -1, Y(\omega_1) = Y(\omega_3) = 1, Y(\omega_2) = 0.$$

Тогда *XУ = X, M(X) = M(XУ)* = 0, следовательно, *M(XУ) = M(X)M(У).* Однако при этом *P(X=0) = P(У=0) = P(X=0, У=0)* = $P(\omega_2) = 1/3$, в то время как вероятность события {*X*=0, *Y*=0} в случае независимых X и У должна была равняться $\frac{1}{3}\times\frac{1}{3}=\frac{1}{9}$.

Независимость нескольких случайных величин *X, Y, Z,…* означает по определению, что для любых чисел *x, y, z,…* справедливо равенство

*P(X=x, Y=y, Z=z,...) = P(X=x) P(Y=y) P(Z=z)...*

Например, если случайные величины определяются по результатам различных испытаний в схеме независимых испытаний, то они независимы.

**Дисперсия случайной величины.** Математическое ожидание показывает, вокруг какой точки группируются значения случайной величины. Необходимо также уметь измерить изменчивость случайной величины относительно математического ожидания. Выше показано, что $M[(X-a)^2]$ достигает минимума по *a* при *a = M(X).* Поэтому за показатель изменчивости случайной величины естественно взять именно $M[(X-M(X))^2]$.

*Определение 5.* Дисперсией случайной величины $X$ называется число $\sigma^2 = D(X) = M[(X - M(X))^2]$.

Установим ряд свойств дисперсии случайной величины, постоянно используемых в вероятностно-статистических методах принятия решений.

*Утверждение 8.* Пусть $X$ – случайная величина, $a$ и $b$ – некоторые числа, $Y = aX + b$. Тогда $D(Y) = a^2D(X)$.

Как следует из утверждений 3 и 5, $M(Y) = aM(X) + b$. Следовательно, $D(Y) = M[(Y - M(Y))^2] = M[(aX + b - aM(X) - b)^2] = M[a^2(X - M(X))^2]$. Поскольку постоянный множитель можно выносить за знак суммы, то $M[a^2(X - M(X))^2] = a^2 M[(X - M(X))^2] = a^2 D(X)$.

Утверждение 8 показывает, в частности, как меняется дисперсия результата наблюдений при изменении начала отсчета и единицы измерения. Оно дает правило преобразования расчетных формул при переходе к другим значениям параметров сдвига и масштаба.

*Утверждение 9.* Если случайные величины $X$ и $У$ независимы, то дисперсия их суммы $X+У$ равна сумме дисперсий: $D(X+Y) = D(X) + D(Y)$.

Для доказательства воспользуемся тождеством

$$(X+У-(M(X)+M(У))^2 = (X-M(X))^2 + 2(X-M(X))(У-M(У)) + (У-M(У))^2,$$

которое вытекает из известной формулы элементарной алгебры $(a+b)^2 = a^2 + 2ab + b^2$ при подстановке $a = X-M(X)$ и $b = Y-M(Y)$. Из утверждений 3 и 5 и определения дисперсии следует, что

$$D(X+Y) = D(X) + D(Y) + 2M\{(X-M(X))(У-M(У))\}.$$

Согласно утверждению 6 из независимости X и У вытекает независимость $X$-$M(X)$ и $У$-$M(У)$. Из утверждения 7 следует, что

$$M\{(X-M(X))(У-M(У))\} = M(X-M(X))M(У-M(У)).$$

Поскольку $M(X-M(X)) = 0$ (см. утверждение 3), то правая часть последнего равенства равна 0, откуда с учетом двух предыдущих равенств и следует заключение утверждения 9.

*Утверждение 10.* Пусть $X_1, X_2, \ldots, X_k$ – попарно независимые случайные величины (т.е. $X_i$ и $X_j$ независимы, если $i \neq j$ ). Пусть $Y_k$ – их сумма, $Y_k = X_1 + X_2 + \ldots + X_k$. Тогда математическое ожидание суммы равно сумме математических ожиданий слагаемых, $M(Y_k) = M(X_1) + M(X_2) + \ldots + M(X_k)$, дисперсия суммы равна сумме дисперсий слагаемых, $D(Y_k) = D(X_1) + D(X_2) + \ldots + D(X_k)$.

Соотношения, сформулированные в утверждении 10, являются основными при изучении выборочных характеристик, поскольку результаты наблюдений или измерений, включенные в выборку, обычно рассматриваются в математической статистике, теории принятия решений и эконометрике как реализации независимых случайных величин.

Для любого набора числовых случайных величин (не только независимых) математическое ожидание их суммы равно сумме их математических ожиданий. Это утверждение является обобщением утверждения 5. Строгое доказательство легко проводится методом математической индукции.

При выводе формулы для дисперсии $D(Y_k)$ воспользуемся следующим свойством символа суммирования:

$$\left(\sum_{1\le i\le k} a_i\right)^2 = \left(\sum_{1\le i\le k} a_i\right)\left(\sum_{1\le j\le k} a_j\right) = \sum_{1\le i\le k,\,1\le j\le k} a_i a_j.$$

Положим $a_i = X_i - M(X_i)$, получим

$$(X_1 + X_2 + \ldots + X_k - M(X_1) - M(X_2) - \ldots - M(X_k))^2 =$$
$$= \sum_{1\le i\le k,\,1\le j\le k} (X_i - M(X_i))(X_j - M(X_j)).$$

Воспользуемся теперь тем, что математическое ожидание суммы равно сумме математических ожиданий:

$$D(Y_k) = \sum_{1\le i\le k,\,1\le j\le k} M\{(X_i - M(X_i))(X_j - M(X_j))\}. \quad (8)$$

Как показано при доказательстве утверждения 9, из попарной независимости рассматриваемых случайных величин следует, что $M\{(X_i - M(X_i))(X_j - M(X_j))\} = 0$ при $i \neq j$ . Следовательно, в сумме (8) остаются только члены с $i=j$, а они равны как раз $D(X_i)$.

Полученные в утверждениях 8-10 фундаментальные свойства таких характеристик случайных величин, как математическое ожидание и дисперсия, постоянно используются практически во всех вероятностно-статистических моделях реальных явлений и процессов.

*Пример 9.* Рассмотрим событие $A$ и случайную величину $X$ такую, что $X(\omega)=1$, если $\omega \in A$, и $X(\omega)=0$ в противном случае, т.е. если $\omega \in \Omega \setminus A$. Покажем, что $M(X) = P(A)$, $D(X) = P(A)(1 - P(A))$.

Воспользуемся формулой (5) для математического ожидания. Случайная величина $X$ принимает два значения – 0 и 1, значение 1 с вероятностью $P(A)$ и значение 0 с вероятностью $1 - P(A)$, а потому $M(X) = 1 \cdot P(A) + 0 \cdot (1 - P(A)) = P(A)$. Аналогично $(X - M(X))^2 = (1 - P(A))^2$ с вероятностью $P(A)$ и $(X - M(X))^2 = (0 - P(A))^2$ с вероятностью $1 - P(A)$, а потому $D(A) = (1 - P(A))^2 P(A) + (P(A))^2(1 - P(A))$. Вынося общий множитель, получаем, что $D(A) = P(A)(1 - P(A))$.

*Пример 10.* Рассмотрим $k$ независимых испытаний, в каждом из которых некоторое событие $A$ может наступить, а может и не наступить. Введем случайные величины $X_1, X_2, \ldots, X_k$ следующим образом: $X_i(\omega) = 1$, если в $i$-ом испытании событие $A$ наступило, и $X_i(\omega) = 0$ в противном случае. Тогда случайные величины $X_1, X_2, \ldots, X_k$ попарно независимы (см. пример 7). Как показано в примере 9, $M(X_i) = p$, $D(X_i) = p(1 - p)$, где $p = P(A)$. Иногда $p$ называют «вероятностью успеха» - в случае, если наступление события $A$ рассматривается как «успех».

Случайная величина $B = X_1 + X_2 + \ldots + X_k$ называется биномиальной. Ясно, что $0 \le B \le k$ при всех возможных исходах опытов. Чтобы найти распределение $B$, т.е. вероятности $P(B = a)$ при $a = 0, 1, \ldots, k$, достаточно знать $p$ – вероятность наступления рассматриваемого события в каждом из опытов. Действительно, случайное событие $B = a$ осуществляется тогда и только тогда, когда событие $A$ наступает ровно при $a$ испытаниях. Если известны номера всех этих испытаний (т.е. номера в последовательности испытаний), то вероятность одновременного осуществления в $a$ опытах события $A$ и в $k$-$a$ опытах противоположного ему – это вероятность произведения $k$ независимых событий. Вероятность произведения равна произведению вероятностей, т.е. $p^a(1 - p)^{k-a}$. Сколькими способами можно задать номера $a$ испытаний из $k$? Это $\binom{k}{a}$ - число сочетаний из $k$ элементов по $a$, рассматриваемое в комбинаторике. Как известно,

$$\binom{k}{a} = \frac{k!}{a!(k-a)!},$$

где символом $k!$ обозначено произведение всех натуральных чисел от 1 до $k$, т.е. $k! = 1 \cdot 2 \cdot \ldots \cdot k$ (дополнительно принимают, что $0! = 1$). Из сказанного следует, что биномиальное распределение, т.е. распределение биномиальной случайной величины, имеет вид

$$P(B = a) = \binom{k}{a} p^a (1 - p)^{k-a}.$$

Название «биномиальное распределение» основано на том, что $P(B = a)$ является членом с номером $(a+1)$ в разложении по биному Ньютона

$$(A + C)^k = \sum_{0 \le j \le k} \binom{k}{j} A^{k-j} C^j,$$

если положить $\mathbf{A} = 1 - p$, $C = p$. Тогда при $j = a$ получим

$$\binom{k}{j} A^{k-j} C^j = P(B = a).$$

Для числа сочетаний из $k$ элементов по $a$, кроме $\binom{k}{a}$, используют обозначение $C_k^a$.

Из утверждения 10 и расчетов примера 9 следует, что для случайной величины $B$, имеющей биномиальное распределение, математическое ожидание и дисперсия выражаются формулами

$$M(B) = kp, \quad D(B) = kp(1 - p),$$

поскольку $B$ является суммой $k$ независимых случайных величин с одинаковыми математическими ожиданиями и дисперсиями, найденными в примере 9.

**Неравенства Чебышёва.** Во введении к разделу обсуждалась задача проверки того, что доля дефектной продукции в партии равна определенному числу. Для демонстрации вероятностно-статистического подхода к проверке подобных утверждений являются полезными неравенства, впервые примененные в теории вероятностей великим русским математиком Пафнутием Львовичем Чебышёвым (1821-1894) и потому носящие его имя. Эти неравенства широко используются в теории математической статистики, а также непосредственно применяются в ряде практических задач принятия решении. Например, в задачах статистического анализа технологических процессов и качества продукции в случаях, когда явный вид функции распределения результатов наблюдений не известен (см. ниже, где, в частности, они применяются в задаче исключения резко отклоняющихся результатов наблюдений).

*Первое неравенство Чебышева.* Пусть X – неотрицательная случайная величина (т.е. $X(\omega) \geq 0$ для любого $\omega \in \Omega$). Тогда для любого положительного числа *a* справедливо неравенство

$$P(X \geq a) \leq \frac{M(X)}{a}.$$

*Доказательство.* Все слагаемые в правой части формулы (4), определяющей математическое ожидание, в рассматриваемом случае неотрицательны. Поэтому при отбрасывании некоторых слагаемых сумма не увеличивается. Оставим в сумме только те члены, для которых $X(\omega) \geq a$. Получим, что

$$M(X) \geq \sum_{\omega: X(\omega) \geq a} X(\omega) P(\omega). \quad (9)$$

Для всех слагаемых в правой части (9) $X(\omega) \geq a$, поэтому

$$\sum_{\omega: X(\omega) \geq a} X(\omega) P(\omega) \geq a \sum_{\omega: X(\omega) \geq a} P(\omega) = aP(X \geq a). \quad (10)$$

Из (9) и (10) следует требуемое.

*Второе неравенство Чебышева.* Пусть X – случайная величина. Для любого положительного числа *a* справедливо неравенство

$$P(|X - M(X)| \geq a) \leq \frac{D(X)}{a^2}.$$

Это неравенство содержалось в работе П.Л.Чебышёва «О средних величинах», доложенной Российской академии наук 17 декабря 1866 г. и опубликованной в следующем году.

Для доказательства второго неравенства Чебышёва рассмотрим случайную величину $Y = (X - M(X))^2$. Она неотрицательна, и потому для любого положительного числа *b*, как следует из первого неравенства Чебышёва, справедливо неравенство

$$P(Y \geq b) \leq \frac{M(Y)}{b} = \frac{D(X)}{b}.$$

Положим $b = a^2$. Событие $\{Y \geq b\}$ совпадает с событием $\{|X - M(X)| \geq a\}$, а потому

$$P(|X - M(X)| \geq a) = P(Y \geq a^2) \leq \frac{D(X)}{a},$$

что и требовалось доказать.

*Пример 11*. Можно указать неотрицательную случайную величину *X* и положительное число *a* такие, что первое неравенство Чебышёва обращается в равенство.

Достаточно рассмотреть $X(\omega) = a$. Тогда $M(X) = a$, $M(X)/a = 1$ и $P(a \geq a) = 1$, т.е. $P(X \geq a) = M(X)|a = 1$.

Следовательно, первое неравенство Чебышёва в его общей формулировке не может быть усилено. Однако для подавляющего большинства случайных величин, используемых при вероятностно-статистическом моделировании процессов принятия решений, левые части неравенств Чебышёва много меньше соответствующих правых частей.

*Пример 12.* Может ли первое неравенство Чебышёва обращаться в равенство при всех *a*? Оказывается, нет. Покажем, что для любой неотрицательной случайной величины с ненулевым математическим ожиданием можно найти такое положительное число *a*, что первое неравенство Чебышёва является строгим.

Действительно, математическое ожидание неотрицательной случайной величины либо положительно, либо равно 0. В первом случае возьмем положительное $a$, меньшее положительного числа $M(X)$, например, положим $a = M(X)/2$. Тогда $M(X)/a$ больше 1, в то время как вероятность события не может превышать 1, а потому первое неравенство Чебышева является для этого $a$ строгим. Второй случай исключается условиями примера 11.

Отметим, что во втором случае равенство 0 математического ожидания влечет тождественное равенство 0 случайной величины. А для такой случайной величины при любом положительном $a$ и левая и правая части первого неравенства Чебышёва равны 0.

Можно ли в формулировке первого неравенства Чебышева отбросить требование неотрицательности случайной величины $X$? А требование положительности $a$? Легко видеть, что ни одно из двух требований не может быть отброшено, поскольку иначе правая часть первого неравенства Чебышева может стать отрицательной.

**Закон больших чисел.** Неравенство Чебышёва позволяет доказать замечательный результат, лежащий в основе математической статистики – закон больших чисел. Из него вытекает, что выборочные характеристики при возрастании числа опытов приближаются к теоретическим, а это дает возможность оценивать параметры вероятностных моделей по опытным данным. Без закона больших чисел не было бы *бо́льшей* части прикладной математической статистики.

*Теорема Чебышёва.* Пусть случайные величины $X_1, X_2, \ldots, X_k$ попарно независимы и существует число $C$ такое, что $D(X_i) \le C$ при всех $i = 1, 2, \ldots, k$. Тогда для любого положительного $\varepsilon$ выполнено неравенство

$$P\left\{\left|\frac{X_1 + X_2 + \ldots + X_k}{k} - \frac{M(X_1) + M(X_2) + \ldots + M(X_k)}{k}\right| \ge \varepsilon\right\} \le \frac{C}{k\varepsilon^2}. \quad (11)$$

*Доказательство.* Рассмотрим случайные величины $Y_k = X_1 + X_2 + \ldots + X_k$ и $Z_k = Y_k/k$. Тогда согласно утверждению 10

$$M(Y_k) = M(X_1) + M(X_2) + \ldots + M(X_k),\ D(Y_k) = D(X_1) + D(X_2) + \ldots + D(X_k).$$

Из свойств математического ожидания следует, что $M(Z_k) = M(Y_k)/k$, а из свойств дисперсии - что $D(Z_k) = D(Y_k)/k^2$. Таким образом,

$$M(Z_k) = \{M(X_1) + M(X_2) + \ldots + M(X_k)\}/k,$$
$$D(Z_k) = \{D(X_1) + D(X_2) + \ldots + D(X_k)\}/k^2.$$

Из условия теоремы Чебышёва, что

$$D(Z_k) \le \frac{Ck}{k^2} = \frac{C}{k}.$$

Применим к $Z_k$ второе неравенство Чебышёва. Получим для стоящей в левой части неравенства (11) вероятности оценку

$$P\{|Z_k - M(Z_k)| \ge \varepsilon\} \le \frac{D(Z_k)}{\varepsilon^2} \le \frac{C}{k\varepsilon^2},$$

что и требовалось доказать.

Эта теорема была получена П.Л.Чебышёвым в той же работе 1867 г. «О средних величинах», что и неравенства Чебышёва.

*Пример 13.* Пусть $C = 1$, $\varepsilon = 0{,}1$. При каких $k$ правая часть неравенства (11) не превосходит 0,1? 0,05? 0,00001?

В рассматриваемом случае правая часть неравенства (11) равно $100/k$. Она не превосходит 0,1, если $k$ не меньше 1000, не превосходит 0,05, если $k$ не меньше 2000, не превосходит 0,00001, если $k$ не меньше 10 000 000.

Правая часть неравенства (11), а вместе с ней и левая, при возрастании $k$ и фиксированных $C$ и $\varepsilon$ убывает, приближаясь к 0. Следовательно, вероятность того, что среднее арифметическое независимых случайных величин отличается от своего математического ожидания менее чем на $\varepsilon$, приближается к 1 при возрастании числа случайных величин, причем при любом $\varepsilon$. Это утверждение называют ЗАКОНОМ БОЛЬШИХ ЧИСЕЛ.

Наиболее важен для вероятностно-статистических методов принятия решений (и для математической статистики в целом) случай, когда все $X_i$, $i = 1, 2, \ldots$, имеют одно и то же математическое ожидание $M(X_1)$ и одну и ту же дисперсию $\sigma^2 = D(X_1)$. В качестве замены

(оценки) неизвестного исследователю математического ожидания используют выборочное среднее арифметическое

$$\overline{X} = \frac{X_1 + X_2 + ... + X_k}{k}.$$

Из закона больших чисел следует, что $\overline{X}$ при увеличении числа опытов (испытаний, измерений) сколь угодно близко приближается к $M(X_1)$, что записывают так:

$$\overline{X} \xrightarrow{P} M(X_1).$$

Здесь знак $\xrightarrow{P}$ означает «сходимость по вероятности». Обратим внимание, что понятие «сходимость по вероятности» отличается от понятия «переход к пределу» в математическом анализе. Напомним, что последовательность $b_n$ имеет предел $b$ при $n \to \infty$, если для любого сколь угодно малого $\delta > 0$ существует число $n(\delta)$ такое, что при любом $n > n(\delta)$ справедливо утверждение: $b_n \in (b-\delta; b+\delta)$. При использовании понятия «сходимость по вероятности» элементы последовательности предполагаются случайными, вводится еще одно сколь угодно малое число $\varepsilon > 0$ и утверждение $b_n \in (b-\delta; b+\delta)$ предполагается выполненным не наверняка, а с вероятностью не менее $1-\varepsilon$.

В начале главы отмечалось, что с точки зрения ряда естествоиспытателей вероятность события $A$ – это число, к которому приближается отношение количества осуществлений события $A$ к количеству всех опытов при безграничном увеличении числа опытов. Известный математики Якоб Бернулли (1654-1705), живший в городе Базель в Швейцарии, в самом конце XVII века доказал это утверждение в рамках математической модели (опубликовано доказательство было лишь после его смерти, в 1713 году). Современная формулировка теоремы Бернулли такова.

*Теорема Бернулли.* Пусть $m$ – число наступлений события $A$ в $k$ независимых (попарно) испытаниях, и $p$ есть вероятность наступления события $A$ в каждом из испытаний. Тогда при любом $\varepsilon > 0$ справедливо неравенство

$$P\left\{\left|\frac{m}{k} - p\right| \geq \varepsilon\right\} \leq \frac{p(1-p)}{k\varepsilon^2}. \quad (12)$$

*Доказательство.* Как показано в примере 10, случайная величина $m$ имеет биномиальное распределение с вероятностью успеха $p$ и является суммой $k$ независимых случайных величин $X_i$, $i$ = 1, 2. …, $k$, каждое из которых равно 1 с вероятностью $p$ и 0 с вероятностью 1-$p$, т.е. $m = X_1 + X_2 + ... + X_k$ .Применим к $X_1, X_2, ..., X_k$ теорему Чебышёва с $C = p(1 - p)$ и получим требуемое неравенство (12).

Теорема Бернулли дает возможность связать математическое определение вероятности (по А.Н.Колмогорову) с определением ряда естествоиспытателей (по Р.Мизесу (1883-1953)), согласно которому вероятность есть предел частоты в бесконечной последовательности испытаний. Продемонстрируем эту связь. Для этого сначала отметим, что

$$p(1-p) \leq 1/4$$

при всех $p$. Действительно,

$$1/4 - p(1-p) = (p - 1/2)^2 \geq 0.$$

Следовательно, в теореме Чебышёва можно использовать $C$ = j. Тогда при любом $p$ и фиксированном $\varepsilon$ правая часть неравенства (12) при возрастании $k$ приближается к 0, что и доказывает согласие математического определения в рамках вероятностной модели с мнением естествоиспытателей.

Есть и прямые экспериментальные подтверждения того, что частота осуществления определенных событий близка к вероятности, определенной из теоретических соображений. Рассмотрим бросания монеты. Поскольку и герб, и решетка имеют одинаковые шансы оказаться сверху, то вероятность выпадения герба равна S из соображений равновозможности. Французский естествоиспытатель XVIII века Бюффон бросил монету 4040 раз, герб выпал при этом 2048 раз. Частота появления герба в опыте Бюффона равна 0,507. Английский статистик К.Пирсон бросил монету 12000 раз и при этом наблюдал 6019 выпадений герба – частота 0,5016. В другой раз он бросил монету 24000 раз, герб выпал 12012 раз – частота 0,5005. Как видим, во

всех этих случаях частоты лишь незначительно отличаются от теоретической вероятности 0,5 [6, с.148].

**О проверке статистических гипотез.** С помощью неравенства (12) можно кое-что сказать по поводу проверки соответствия качества продукции заданным требованиям.

Пусть из 100000 единиц продукции 30000 оказались дефектными. Согласуется ли это с гипотезой о том, что вероятность дефектности равна 0,23? Прежде всего, какую вероятностную модель целесообразно использовать? Принимаем, что проводится сложный опыт, состоящий из 100000 испытаний 100000 единиц продукции на годность. Считаем, что испытания (попарно) независимы и что в каждом испытании вероятность того, что единица продукции является дефектной, равна $p$. В реальном опыте получено, что событие «единица продукции не является годной» осуществилось 30000 раз при 100000 испытаниях. Согласуется ли это с гипотезой о том, что вероятность дефектности $p = 0{,}23$?

Для проверки гипотезы воспользуемся неравенством (12). В рассматриваемом случае $k = 100000$, $m = 30000$, $m/k = 0{,}3$, $p = 0{,}23$, $m/k - p = 0{,}07$. Для проверки гипотезы поступают так. Оценим вероятность того, что $m/k$ отличается от $p$ так же, как в рассматриваемом случае, или больше, т.е. оценим вероятность выполнения неравенства $|m/k - 0{,}23| \geq 0{,}07$. Положим в неравенстве (12) $p = 0{,}23$, $\varepsilon = 0{,}07$. Тогда

$$P\left\{\left|\frac{m}{k} - 0{,}23\right| \geq 0{,}07\right\} \leq \frac{0{,}23 \cdot 0{,}77}{0{,}0049k} \approx \frac{36{,}11}{k}. \quad (13)$$

При $k = 100000$ правая часть (13) меньше 1/2500. Значит, вероятность того, что отклонение будет не меньше наблюдаемого, весьма мала. Следовательно, если исходная гипотеза верна, то в рассматриваемом опыте осуществилось событие, вероятность которого меньше 1/2500. Поскольку 1/2500 – очень маленькое число, то исходную гипотезу надо отвергнуть.

Подробнее методы проверки статистических гипотез будут рассмотрены ниже. Здесь отметим, что одна из основных характеристик метода проверки гипотезы – уровень значимости, т.е. вероятность отвергнуть проверяемую гипотезу (ее в математической статистике называют нулевой и обозначают $H_0$), когда она верна. Для проверки статистической гипотезы часто поступают так. Выбирают уровень значимости - малое число $\alpha$. Если описанная в предыдущем абзаце вероятность меньше $\alpha$, то гипотезу отвергают, как говорят, на уровне значимости $\alpha$. Если эта вероятность больше или равна $\alpha$, то гипотезу принимают. Обычно в вероятностно-статистических методах принятия решений выбирают $\alpha = 0{,}05$, значительно реже $\alpha = 0{,}01$ или $\alpha = 0{,}1$, в зависимости от конкретной практической ситуации. В рассматриваемом случае $\alpha$, напомним, та доля опытов (т.е. проверок партий по 100000 единиц продукции), в которой мы отвергаем гипотезу $H_0$: $p = 0{,}23$, хотя она верна.

Насколько результат проверки гипотезы $H_0$ зависит от числа испытаний $k$? Пусть при $k = 100$, $k = 1000$, $k = 10000$ оказалось, что $m = 30$, $m = 300$, $m = 3000$ соответственно, так что во всех случаях $m/k = 0{,}3$. Какие значения принимает вероятность

$$P_k = P\left\{\left|\frac{m}{k} - 0{,}23\right| \geq 0{,}07\right\}$$

и ее оценка – правая часть формулы (13)?

При $k = 100$ правая часть (13) равна приблизительно 0,36, что не дает оснований отвергнуть гипотезу. При $k = 1000$ правая часть (13) равна примерно 0,036. Гипотеза отвергается на уровне значимости $\alpha = 0{,}05$ (и б = 0,1), но на основе оценки вероятности с помощью правой части формулы (13) не удается отвергнуть гипотезу на уровне значимости б = 0,01. При $k = 10000$ правая часть (13) меньше 1/250, и гипотеза отвергается на всех обычно используемых уровнях значимости.

Более точные расчеты, основанные на применении центральной предельной теоремы теории вероятностей (см. ниже), дают $P_{100} = 0{,}095$, $P_{1000} = 0{,}0000005$, так что оценка (13) является в рассматриваемом случае весьма завышенной. Причина в том, что получена она из наиболее общих соображений, применительно ко всем возможным случайным величинам улучшить ее нельзя (см. пример 11 выше), но применительно к биномиальному распределению – можно.

Ясно, что без введения уровня значимости не обойтись, ибо даже очень большие отклонения $m/k$ от $p$ имеют положительную вероятность осуществления. Так, при

справедливости гипотезы $H_0$ событие «все 100000 единиц продукции являются дефектными» отнюдь не является невозможным с математической точки зрения, оно имеет положительную вероятность осуществления, равную $0{,}23^{100000}$, хотя эта вероятность и невообразимо мала.

Аналогично разберем проверку гипотезы о симметричности монеты.

*Пример 14.* Если монета симметрична, то $p$ = S, где $p$ – вероятность выпадения герба. Согласуется ли с этой гипотезой результат эксперимента, в котором при 10000 бросаниях выпало 4000 гербов?

В рассматриваемом случае $m/k = 0{,}4$. Положим в неравенстве (12) $p = 0{,}5$, $e = 0{,}1$:

$$P\left\{\left|\frac{m}{k} - 0{,}5\right| \geq 0{,}1\right\} \leq \frac{0{,}5 \cdot 0{,}5}{0{,}01k} = \frac{25}{k}.$$

При $k = 10000$ правая часть последнего неравенства равна 1/400. Значит, если исходная гипотеза верна, то в нашем единственном эксперименте осуществилось событие, вероятность которого весьма мала – меньше 1/400. Поэтому исходную гипотезу необходимо отвергнуть.

Если из 1000 бросаний монеты гербы выпали в 400 случаях, то правая часть выписанного выше неравенства равна 1/40. Гипотеза симметричности отклоняется на уровне значимости 0,05 (и 0,1), но рассматриваемые методы не дают возможности отвергнуть ее на уровне значимости 0,01.

Если $k = 100$, а $m = 40$, то правая часть неравенства равна j. Оснований для отклонения гипотезы нет. С помощью более тонких методов, основанных на центральной предельной теореме теории вероятностей, можно показать, что левая часть неравенства равна приблизительно 0,05. Это показывает, как важно правильно выбрать метод проверки гипотезы или оценивания параметров. Следовательно, целесообразна стандартизация подобных методов, позволяющая сэкономить усилия, необходимые для сравнения и выбора наилучшего метода, а также избежать устаревших, неверных или неэффективных методов.

Ясно, что даже по нескольким сотням опытов нельзя достоверно отличить абсолютно симметричную монету ($p$ = S) от несколько несимметричной монеты (для которой, скажем, $p$ = 0,49). Более того, любая реальная монета несколько несимметрична, так что монета с $p$ = S - математическая абстракция. Между тем в ряде управленческих и производственных ситуаций необходимо осуществить справедливую жеребьевку, а для этого требуется абсолютно симметричная монета. Например, речь может идти об очередности рассмотрения инвестиционных проектов комиссией экспертов, о порядке вызова для собеседования кандидатов на должность, об отборе единиц продукции из партии в выборку для контроля и т.п.

*Пример 15.* Можно ли с помощью несимметричной монеты получить последовательность испытаний с двумя исходами, каждый из которых имеет вероятность 1/2 ?

Ответ: да, можно. Приведем способ, предложенный видным польским математиком Гуго Штейнгаузом (1887-1972).

Будем бросать монету два раза подряд и записывать исходы бросаний так (Г – герб, Р – решетка, на первом месте стоит результат первого бросания, на втором – второго): ГР запишем как Г, в то время РГ запишем как Р, а ГГ и РР вообще не станем записывать. Например, если исходы бросаний окажутся такими:

ГР, РГ, ГР, РР, ГР, РГ, ГГ, РГ, РР, РГ,

то запишем их в виде:

Г, Р, Г, Г, Р, Р, Р.

Сконструированная таким образом последовательность обладает теми же свойствами, что и полученная при бросании идеально симметричной монеты, поскольку даже у несимметричной монеты последовательность ГР встречается столь же часто, как и последовательность РГ.

Применим теорему Бернулли и неравенство (12) к обработке реальных данных.

*Пример 16.* С 1871 г. по 1900 г. в Швейцарии родились 1359671 мальчик и 1285086 девочек. Совместимы ли эти данные с предположением о том, что вероятность рождения мальчика равна 0,5? А с предположением, что она равна 0,515? Другими словами, требуется проверить нулевые гипотезы $H_0$: $p = 0{,}5$ и $H_0$: $p = 0{,}515$ с помощью неравенства (12).

Число испытаний равно общему числу рождений, т.е. 1359671 + 1285086 = 2644757. Есть все основания считать испытания независимыми. Число рождений мальчиков составляет приблизительно 0,514 всех рождений. В случае $p$ = S имеем $e$ = 0,014, и правая часть неравенства (12) имеет вид

$$\frac{0{,}5\times 0{,}5}{0{,}014\times 0{,}014\times 2644757}\approx 0{,}00001.$$

Таким образом, гипотезу $p$ = 0,5 следует считать несовместимой с приведенными в условии данными. В случае $p$ = 0,515 имеем $e$ = 0,001, и правая часть (12) равна приблизительно 0,1, так что с помощью неравенства (12) отклонить гипотезу $H_0$: $p$ = 0,515 нельзя.

Итак, здесь на основе элементарной теории вероятностей (с конечным пространством элементарных событий) мы сумели построить вероятностные модели для описания проверки качества деталей (единиц продукции) и бросания монет и предложить методы проверки гипотез, относящихся к этим явлениям. В математической статистике есть более тонкие и сложные методы проверки описанных выше гипотез, которыми и пользуются в практических расчетах.

Можно спросить: «В рассмотренных выше моделях вероятности были известны заранее – со слов Струкова или же из-за того, что мы предположили симметричность монеты. А как строить модели, если вероятности неизвестны? Как оценить неизвестные вероятности?» Теорема Бернулли – результат, с помощью которого дается ответ на этот вопрос. Именно, оценкой неизвестной вероятности $p$ является число $m/k$, поскольку доказано, что при возрастании $k$ вероятность того, что $m/k$ отличается от $p$ более чем на какое-либо фиксированное число, приближается к 0. Оценка будет тем точнее, чем больше $k$. Более того, можно доказать, что с некоторой точки зрения (см. далее) оценка $m/k$ для вероятности $p$ является наилучшей из возможных (в терминах математической статистики – состоятельной, несмещенной и эффективной).

### 1.2.3. Суть вероятностно-статистических методов принятия решений

Как подходы, идеи и результаты теории вероятностей и математической статистики используются при принятии решений?

Базой является вероятностная модель реального явления или процесса, т.е. математическая модель, в которой объективные соотношения выражены в терминах теории вероятностей. Вероятности используются прежде всего для описания неопределенностей, которые необходимо учитывать при принятии решений. Имеются в виду как нежелательные возможности (риски), так и привлекательные («счастливый случай»). Иногда случайность вносится в ситуацию сознательно, например, при жеребьевке, случайном отборе единиц для контроля, проведении лотерей или опросов потребителей.

Теория вероятностей позволяет по одним вероятностям рассчитать другие, интересующие исследователя. Например, по вероятности выпадения герба можно рассчитать вероятность того, что при 10 бросаниях монет выпадет не менее 3 гербов. Подобный расчет опирается на вероятностную модель, согласно которой бросания монет описываются схемой независимых испытаний, кроме того, выпадения герба и решетки равновозможны, а потому вероятность каждого из этих событий равна S. Более сложной является модель, в которой вместо бросания монеты рассматривается проверка качества единицы продукции. Соответствующая вероятностная модель опирается на предположение о том, что контроль качества различных единиц продукции описывается схемой независимых испытаний. В отличие от модели с бросанием монет необходимо ввести новый параметр – вероятность $p$ того, что единица продукции является дефектной. Модель будет полностью описана, если принять, что все единицы продукции имеют одинаковую вероятность оказаться дефектными. Если последнее предположение неверно, то число параметров модели возрастает. Например, можно принять, что каждая единица продукции имеет свою вероятность оказаться дефектной.

Обсудим модель контроля качества с общей для всех единиц продукции вероятностью дефектности $p$. Чтобы при анализе модели «дойти до числа», необходимо заменить $p$ на некоторое конкретное значение. Для этого необходимо выйти из рамок вероятностной модели и обратиться к данным, полученным при контроле качества. Математическая статистика решает обратную задачу по отношению к теории вероятностей. Ее цель – на основе результатов наблюдений (измерений, анализов, испытаний, опытов) получить выводы о вероятностях, лежащих в основе вероятностной модели. Например, на основе частоты появления дефектных изделий при контроле можно сделать выводы о вероятности дефектности (см. теорему Бернулли

выше). На основе неравенства Чебышева делались выводы о соответствии частоты появления дефектных изделий гипотезе о том, что вероятность дефектности принимает определенное значение.

Таким образом, применение математической статистики опирается на вероятностную модель явления или процесса. Используются два параллельных ряда понятий – относящиеся к теории (вероятностной модели) и относящиеся к практике (выборке результатов наблюдений). Например, теоретической вероятности соответствует частота, найденная по выборке. Математическому ожиданию (теоретический ряд) соответствует выборочное среднее арифметическое (практический ряд). Как правило, выборочные характеристики являются оценками теоретических. При этом величины, относящиеся к теоретическому ряду, «находятся в головах исследователей», относятся к миру идей (по древнегреческому философу Платону), недоступны для непосредственного измерения. Исследователи располагают лишь выборочными данными, с помощью которых они стараются установить интересующие их свойства теоретической вероятностной модели.

Зачем же нужна вероятностная модель? Дело в том, что только с ее помощью можно перенести свойства, установленные по результатам анализа конкретной выборки, на другие выборки, а также на всю так называемую генеральную совокупность. Термин «генеральная совокупность» используется, когда речь идет о большой, но конечной совокупности изучаемых единиц. Например, о совокупности всех жителей России или совокупности всех потребителей растворимого кофе в Москве. Цель маркетинговых или социологических опросов состоит в том, чтобы утверждения, полученные по выборке из сотен или тысяч человек, перенести на генеральные совокупности в несколько миллионов человек. При контроле качества в роли генеральной совокупности выступает партия продукции.

Чтобы перенести выводы с выборки на более обширную совокупность, необходимы те или иные предположения о связи выборочных характеристик с характеристиками этой более обширной совокупности. Эти предположения основаны на соответствующей вероятностной модели.

Конечно, можно обрабатывать выборочные данные, не используя ту или иную вероятностную модель. Например, можно рассчитывать выборочное среднее арифметическое, подсчитывать частоту выполнения тех или иных условий и т.п. Однако результаты расчетов будут относиться только к конкретной выборке, перенос полученных с их помощью выводов на какую-либо иную совокупность некорректен. Иногда подобную деятельность называют «анализ данных». По сравнению с вероятностно-статистическими методами анализ данных имеет ограниченную познавательную ценность.

Итак, использование вероятностных моделей на основе оценивания и проверки гипотез с помощью выборочных характеристик – вот суть вероятностно-статистических методов принятия решений.

Подчеркнем, что логика использования выборочных характеристик для принятия решений на основе теоретических моделей предполагает одновременное использование двух параллельных рядов понятий, один из которых соответствует вероятностным моделям, а второй – выборочным данным. К сожалению, в ряде литературных источников, обычно устаревших либо написанных в рецептурном духе, не делается различия между выборочными и теоретическими характеристиками, что приводит читателей к недоумениям и ошибкам при практическом использовании статистических методов.

### 1.2.4. Случайные величины и их распределения

**Распределения случайных величин и функции распределения**. Распределение числовой случайной величины – это функция, которая однозначно определяет вероятность того, что случайная величина принимает заданное значение или принадлежит к некоторому заданному интервалу.

Первое – если случайная величина принимает конечное число значений. Тогда распределение задается функцией $P(X = x)$, ставящей каждому возможному значению $x$ случайной величины $X$ вероятность того, что $X = x$.

Второе – если случайная величина принимает бесконечно много значений. Это возможно лишь тогда, когда вероятностное пространство, на котором определена случайная величина, состоит из бесконечного числа элементарных событий. Тогда распределение задается набором вероятностей $P(a \le X < b)$ для всех пар чисел $a, b$ таких, что $a<b$. Распределение может быть задано с помощью т.н. *функции распределения* $F(x) = P(X<x)$, определяющей для всех действительных $x$ вероятность того, что случайная величина $X$ принимает значения, меньшие $x$. Ясно, что

$$P(a \le X < b) = F(b) - F(a).$$

Это соотношение показывает, что как распределение может быть рассчитано по функции распределения, так и, наоборот, функция распределения – по распределению.

Используемые в вероятностно-статистических методах принятия решений и других прикладных исследованиях функции распределения бывают либо дискретными, либо непрерывными, либо их комбинациями.

Дискретные функции распределения соответствуют дискретным случайным величинам, принимающим конечное число значений или же значения из множества, элементы которого можно перенумеровать натуральными числами (такие множества в математике называют счетными). Их график имеет вид ступенчатой лестницы (рис. 1).

*Пример 1.* Число $X$ дефектных изделий в партии принимает значение 0 с вероятностью 0,3, значение 1 с вероятностью 0,4, значение 2 с вероятностью 0,2 и значение 3 с вероятностью 0,1. График функции распределения случайной величины $X$ изображен на рис.1.

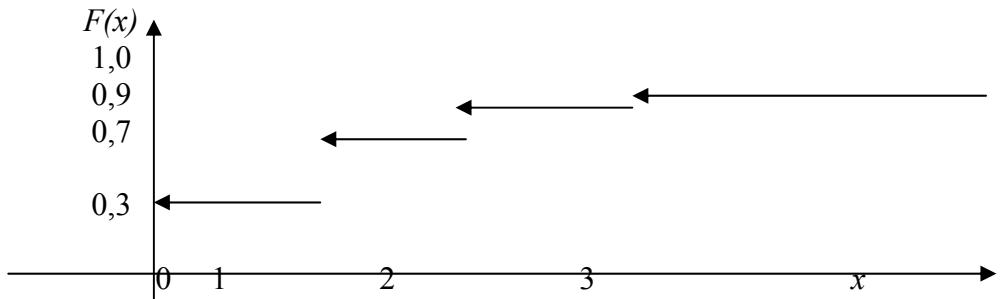



Рис.1. График функции распределения числа дефектных изделий.

Непрерывные функции распределения не имеют скачков. Они монотонно возрастают [1] при увеличении аргумента – от 0 при $x \to -\infty$ до 1 при $x \to +\infty$. Случайные величины, имеющие непрерывные функции распределения, называют непрерывными.

Непрерывные функции распределения, используемые в вероятностно-статистических методах принятия решений, имеют производные. Первая производная $f(x)$ функции распределения $F(x)$ называется плотностью вероятности,

$$f(x) = \frac{dF(x)}{dx}.$$

По плотности вероятности можно определить функцию распределения:

$$F(x) = \int_{-\infty}^{x} f(y)dy.$$

Для любой функции распределения

$$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to +\infty} F(x) = 1,$$

а потому

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

[1] В некоторых случаях, например, при изучении цен, объемов выпуска или суммарной наработки на отказ в задачах надежности, функции распределения постоянны на некоторых интервалах, в которые значения исследуемых случайных величин не могут попасть.

Перечисленные свойства функций распределения постоянно используются в вероятностно-статистических методах принятия решений. В частности, из последнего равенства вытекает конкретный вид констант в формулах для плотностей вероятностей, рассматриваемых ниже.

*Пример 2.* Часто используется следующая функция распределения:

$$F(x)=\begin{cases}0, & x<a,\\ \dfrac{x-a}{b-a}, & a\le x\le b,\\ 1, & x>b.\end{cases} \quad (1)$$

где $a$ и $b$ – некоторые числа, $a<b$. Найдем плотность вероятности этой функции распределения:

$$f(x)=\begin{cases}0, & x<a,\\ \dfrac{1}{b-a}, & a<x<b,\\ 0, & x>b\end{cases}$$

(в точках $x = a$ и $x = b$ производная функции $F(x)$ не существует).

Случайная величина с функцией распределения (1) называется «равномерно распределенной на отрезке [*a; b*]».

Смешанные функции распределения встречаются, в частности, тогда, когда наблюдения в какой-то момент прекращаются. Например, при анализе статистических данных, полученных при использовании планов испытаний на надежность, предусматривающих прекращение испытаний по истечении некоторого срока. Или при анализе данных о технических изделиях, потребовавших гарантийного ремонта.

*Пример 3.* Пусть, например, срок службы электрической лампочки – случайная величина с функцией распределения $F(t)$, а испытание проводится до выхода лампочки из строя, если это произойдет менее чем за 100 часов от начала испытаний, или до момента $t_0 = 100$ часов. Пусть $G(t)$ – функция распределения времени эксплуатации лампочки в исправном состоянии при этом испытании. Тогда

$$G(t)=\begin{cases}F(t), & t\le 100\\ 1, & t>100.\end{cases}$$

Функция $G(t)$ имеет скачок в точке $t_0$, поскольку соответствующая случайная величина принимает значение $t_0$ с вероятностью $1-F(t_0)>0$.

**Характеристики случайных величин.** В вероятностно-статистических методах принятия решений используется ряд характеристик случайных величин, выражающихся через функции распределения и плотности вероятностей.

При описании дифференциации доходов, при нахождении доверительных границ для параметров распределений случайных величин и во многих иных случаях используется такое понятие, как «квантиль порядка $p$», где $0 < p < 1$ (обозначается $x_p$). Квантиль порядка $p$ – значение случайной величины, для которого функция распределения принимает значение $p$ или имеет место «скачок» со значения меньше $p$ до значения больше $p$ (рис.2). Может случиться, что это условие выполняется для всех значений x, принадлежащих этому интервалу (т.е. функция распределения постоянна на этом интервале и равна $p$). Тогда каждое такое значение называется «квантилем порядка $p$». Для непрерывных функций распределения, как правило, существует единственный квантиль $x_p$ порядка $p$ (рис.2), причем

$$F(x_p) = p. \quad (2)$$

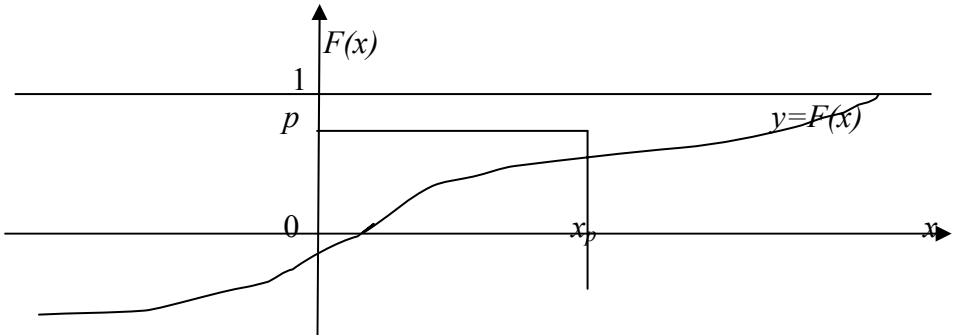


Рис.2. Определение квантиля $x_p$ порядка $p$.

*Пример 4.* Найдем квантиль $x_p$ порядка $p$ для функции распределения $F(x)$ из (1).

При $0 < p < 1$ квантиль $x_p$ находится из уравнения

$$\frac{x-a}{b-a} = p,$$

т.е. $x_p = a + p(b - a) = a(1- p) + bp$. При $p = 0$ любое $x \le a$ является квантилем порядка $p = 0$. Квантилем порядка $p = 1$ является любое число $x \ge b$.

Для дискретных распределений, как правило, не существует $x_p$, удовлетворяющих уравнению (2). Точнее, если распределение случайной величины дается табл.1, где $x_1 < x_2 < ... < x_k$, то равенство (2), рассматриваемое как уравнение относительно $x_p$, имеет решения только для $k$ значений $p$, а именно,

$p = p_1$,
$p = p_1 + p_2$,
$p = p_1 + p_2 + p_3$,
…
$p = p_1 + p_2 + ... + p_m$, $3 < m < k$,
...
$p = p_1 + p_2 + ... + p_k$.

Таблица 1.

Распределение дискретной случайной величины

| Значения $x$ случайной величины $X$ | $x_1$ | $x_2$ | ... | $x_k$ |
|---|---|---|---|---|
| Вероятности $P(X = x)$ | $p_1$ | $p_2$ | ... | $p_k$ |

Для перечисленных $k$ значений вероятности $p$ решение $x_p$ уравнения (2) неединственно, а именно,

$$F(x) = p_1 + p_2 + ... + p_m$$

для всех $x$ таких, что $x_m < x \le x_{m+1}$. Т.е. $x_p$ – любое число из интервала $(x_m; x_{m+1}]$. Для всех остальных $p$ из промежутка (0;1), не входящих в перечень (3), имеет место «скачок» со значения меньше $p$ до значения больше $p$. А именно, если

$$p_1 + p_2 + ... + p_m < p < p_1 + p_2 + ... + p_m + p_{m+1},$$

то $x_p = x_{m+1}$.

Рассмотренное свойство дискретных распределений создает значительные трудности при табулировании и использовании подобных распределений, поскольку невозможным оказывается точно выдержать типовые численные значения характеристик распределения. В частности, это так для критических значений и уровней значимости непараметрических статистических критериев (см. ниже), поскольку распределения статистик этих критериев дискретны.

Большое значение в статистике имеет квантиль порядка $p$ = S. Он называется медианой (случайной величины $X$ или ее функции распределения $F(x)$) и обозначается $Me(X)$. В геометрии есть понятие «медиана» - прямая, проходящая через вершину треугольника и делящая противоположную его сторону пополам. В математической статистике медиана делит пополам не сторону треугольника, а распределение случайной величины: равенство $F(x_{0,5}) = 0{,}5$ означает, что вероятность попасть левее $x_{0,5}$ и вероятность попасть правее $x_{0,5}$ (или непосредственно в $x_{0,5}$) равны между собой и равны S, т.е.

$$P(X < x_{0,5}) = P(X \ge x_{0,5}) = S.$$

Медиана указывает «центр» распределения. С точки зрения одной из современных концепций – теории устойчивых статистических процедур – медиана является более хорошей характеристикой случайной величины, чем математическое ожидание [2,7]. При обработке результатов измерений в порядковой шкале (см. главу о теории измерений) медианой можно пользоваться, а математическим ожиданием – нет.

Ясный смысл имеет такая характеристика случайной величины, как мода – значение (или значения) случайной величины, соответствующее локальному максимуму плотности

вероятности для непрерывной случайной величины или локальному максимуму вероятности для дискретной случайной величины.

Если $x_0$ – мода случайной величины с плотностью $f(x)$, то, как известно из дифференциального исчисления, $\frac{df(x_0)}{dx} = 0$.

У случайной величины может быть много мод. Так, для равномерного распределения (1) каждая точка $x$ такая, что $a < x < b$, является модой. Однако это исключение. Большинство случайных величин, используемых в вероятностно-статистических методах принятия решений и других прикладных исследованиях, имеют одну моду. Случайные величины, плотности, распределения, имеющие одну моду, называются унимодальными.

Математическое ожидание для дискретных случайных величин с конечным числом значений рассмотрено в главе «События и вероятности». Для непрерывной случайной величины $X$ математическое ожидание $M(X)$ удовлетворяет равенству

$$M(X) = \int_{-\infty}^{+\infty} xf(x)dx.$$

являющемуся аналогом формулы (5) из утверждения 2 главы «События и вероятности».

*Пример 5.* Математическое ожидание для равномерно распределенной случайной величины $X$ равно

$$M(X) = \int_a^b \frac{x}{b-a}dx = \frac{1}{b-a}\frac{x^2}{2}\Big|_a^b = \frac{1}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \frac{a+b}{2}.$$

Для рассматриваемых в настоящей главе случайных величин верны все те свойства математических ожиданий и дисперсий, которые были рассмотрены ранее для дискретных случайных величин с конечным числом значений. Однако доказательства этих свойств не приводим, поскольку они требуют углубления в математические тонкости, не являющегося необходимым для понимания и квалифицированного применения вероятностно-статистических методов принятия решений.

*Замечание.* В настоящем учебнике сознательно обходятся математические тонкости, связанные, в частности, с понятиями измеримых множеств и измеримых функций, $\sigma$-алгебры событий и т.п. Желающим освоить эти понятия необходимо обратиться к специальной литературе, в частности, к энциклопедии [1].

Каждая из трех характеристик – математическое ожидание, медиана, мода – описывает «центр» распределения вероятностей. Понятие «центр» можно определять разными способами – отсюда три разные характеристики. Однако для важного класса распределений – симметричных унимодальных – все три характеристики совпадают.

Плотность распределения $f(x)$ – плотность симметричного распределения, если найдется число $x_0$ такое, что

$$f(x) = f(2x_0 - x). \quad (3)$$

Равенство (3) означает, что график функции $y = f(x)$ симметричен относительно вертикальной прямой, проходящей через центр симметрии $x = x_0$. Из (3) следует, что функция симметричного распределения удовлетворяет соотношению

$$F(x) = 1 - F(2x_0 - x). \quad (4)$$

Для симметричного распределения с одной модой математическое ожидание, медиана и мода совпадают и равны $x_0$.

Наиболее важен случай симметрии относительно 0, т.е. $x_0 = 0$. Тогда (3) и (4) переходят в равенства

$$f(x) = f(-x) \quad (5)$$

и

$$F(x) = 1 - F(-x) \quad (6)$$

соответственно. Приведенные соотношения показывают, что симметричные распределения нет необходимости табулировать при всех $x$, достаточно иметь таблицы при $x \geq x_0$.

Отметим еще одно свойство симметричных распределений, постоянно используемое в вероятностно-статистических методах принятия решений и других прикладных исследованиях. Для непрерывной функции распределения

$$P(|X| \le a) = P(-a \le X \le a) = F(a) - F(-a),$$

где $F$ – функция распределения случайной величины $X$. Если функция распределения $F$ симметрична относительно 0, т.е. для нее справедлива формула (6), то

$$P(|X| \le a) = 2F(a) - 1.$$

Часто используют другую формулировку рассматриваемого утверждения: если

$$1 - F(a) = \alpha,$$

то

$$P(|X| > a) = 2\alpha.$$

Если $x_\alpha$ и $x_{1-\alpha}$ - квантили порядка $\alpha$ и $1-\alpha$ соответственно (см. (2)) функции распределения, симметричной относительно 0, то из (6) следует, что

$$x_\alpha = -x_{1-\alpha}.$$

От характеристик положения – математического ожидания, медианы, моды – перейдем к характеристикам разброса случайной величины $X$: дисперсии $D(X) = \sigma^2$, среднему квадратическому отклонению $\sigma$ и коэффициенту вариации $v$. Определение и свойства дисперсии для дискретных случайных величин рассмотрены в предыдущей главе. Для непрерывных случайных величин

$$D(X) = M[(X - M(X))^2] = \int_{-\infty}^{+\infty} (x - M(X))^2 f(x)dx.$$

Среднее квадратическое отклонение – это неотрицательное значение квадратного корня из дисперсии:

$$\sigma = +\sqrt{D(X)}.$$

Коэффициент вариации – это отношение среднего квадратического отклонения к математическому ожиданию:

$$v = \frac{\sigma}{M(X)}.$$

Коэффициент вариации применяется при $M(X)>0$. Он измеряет разброс в относительных единицах, в то время как среднее квадратическое отклонение – в абсолютных.

*Пример 6.* Для равномерно распределенной случайной величины $X$ найдем дисперсию, среднеквадратическое отклонение и коэффициент вариации. Дисперсия равна:

$$D(X) = \int_a^b \frac{1}{b-a}\left(x - \frac{a+b}{2}\right)^2 dx.$$

Замена переменной $y = x - \dfrac{a+b}{2}$ дает возможность записать:

$$D(X) = \frac{1}{b-a}\int_{-c}^{c} y^2 dy = \frac{1}{b-a}\frac{y^3}{3}\Big|_{-c}^{c} = \frac{2c^3}{3(b-a)} = \frac{(b-a)^2}{12},$$

где $c = (b - a)/2$. Следовательно, среднее квадратическое отклонение равно $\sigma = \dfrac{b-a}{2\sqrt{3}}$, а коэффициент вариации таков: $v = \dfrac{b-a}{\sqrt{3}(a+b)}$.

По каждой случайной величине $X$ определяют еще три величины – центрированную $Y$, нормированную $V$ и приведенную $U$. Центрированная случайная величина $Y$ – это разность между данной случайной величиной $X$ и ее математическим ожиданием $M(X)$, т.е. $Y = X - M(X)$. Математическое ожидание центрированной случайной величины $Y$ равно 0, а дисперсия – дисперсии данной случайной величины: $M(Y) = 0$, $D(Y) = D(X)$. Функция распределения $F_Y(x)$ центрированной случайной величины $Y$ связана с функцией распределения $F(x)$ исходной случайной величины $X$ соотношением:

$$F_Y(x) = F(x + M(X)).$$

Для плотностей этих случайных величин справедливо равенство

$$f_Y(x) = f(x + M(X)).$$

Нормированная случайная величина $V$ – это отношение данной случайной величины $X$ к ее среднему квадратическому отклонению $\sigma$, т.е. $V = X/\sigma$. Математическое ожидание и дисперсия нормированной случайной величины $V$ выражаются через характеристики $X$ так:

$$M(V) = \frac{M(X)}{\sigma} = \frac{1}{v}, \quad D(V) = 1,$$

где $v$ – коэффициент вариации исходной случайной величины $X$. Для функции распределения $F_V(x)$ и плотности $f_V(x)$ нормированной случайной величины $V$ имеем:

$$F_V(x) = F(\sigma x), \quad f_V(x) = \sigma f(\sigma x),$$

где $F(x)$ – функция распределения исходной случайной величины $X$, а $f(x)$ – ее плотность вероятности.

Приведенная случайная величина $U$ – это центрированная и нормированная случайная величина:

$$U = \frac{X - M(X)}{\sigma}.$$

Для приведенной случайной величины

$$M(U) = 0, \ D(U) = 1, \ F_U(x) = F(\sigma x + M(X)), \ f_U(x) = \sigma f(\sigma x + M(X)). \quad (7)$$

Нормированные, центрированные и приведенные случайные величины постоянно используются как в теоретических исследованиях, так и в алгоритмах, программных продуктах, нормативно-технической и инструктивно-методической документации. В частности, потому, что равенства $M(U) = 0, \ D(U) = 1$ позволяют упростить обоснования методов, формулировки теорем и расчетные формулы.

Используются преобразования случайных величин и более общего плана. Так, если $Y = aX + b$, где $a$ и $b$ – некоторые числа, то

$$M(Y) = aM(X) + b, \ D(Y) = \sigma^2 D(X), \ F_Y(x) = F\left(\frac{x-b}{a}\right), \ f_Y(x) = \frac{1}{a} f\left(\frac{x-b}{a}\right). \quad (8)$$

*Пример 7.* Если $a = 1/\sigma$, $b = -M(X)/\sigma$, то $Y$ – приведенная случайная величина, и формулы (8) переходят в формулы (7).

С каждой случайной величиной $X$ можно связать множество случайных величин $Y$, заданных формулой $Y = aX + b$ при различных $a>0$ и $b$. Это множество называют *масштабно-сдвиговым семейством*, порожденным случайной величиной $X$. Функции распределения $F_Y(x)$ составляют масштабно сдвиговое семейство распределений, порожденное функцией распределения $F(x)$. Вместо $Y = aX + b$ часто используют запись

$$Y = \frac{X - c}{d}, \quad (9)$$

где

$$d = \frac{1}{a} > 0, \ c = -\frac{b}{a}.$$

Число $c$ называют параметром сдвига, а число $d$ - параметром масштаба. Формула (9) показывает, что $X$ – результат измерения некоторой величины – переходит в $У$ – результат измерения той же величины, если начало измерения перенести в точку $c$, а затем использовать новую единицу измерения, в $d$ раз большую старой.

Для масштабно-сдвигового семейства (9) распределение X называют стандартным. В вероятностно-статистических методах принятия решений и других прикладных исследованиях используют стандартное нормальное распределение, стандартное распределение Вейбулла-Гнеденко, стандартное гамма-распределение и др. (см. ниже).

Применяют и другие преобразования случайных величин. Например, для положительной случайной величины $X$ рассматривают $Y = \lg X$, где $\lg X$ – десятичный логарифм числа $X$. Цепочка равенств

$$F_Y(x) = P(\lg X < x) = P(X < 10^x) = F(10^x)$$

связывает функции распределения $X$ и $Y$.

При обработке данных используют такие характеристики случайной величины $X$ как моменты порядка $q$, т.е. математические ожидания случайной величины $X^q$, $q$ = 1, 2, … Так, само математическое ожидание – это момент порядка 1. Для дискретной случайной величины момент порядка $q$ может быть рассчитан как

$$m_q = M(X^q) = \sum_i x_i^q P(X = x_i).$$

Для непрерывной случайной величины

$$m_q = M(X^q) = \int_{-\infty}^{+\infty} x^q f(x)dx.$$

Моменты порядка $q$ называют также начальными моментами порядка $q$, в отличие от родственных характеристик – центральных моментов порядка $q$, задаваемых формулой

$$\mu_q = M[(X - M(X))^q], \ q = 2,3,...$$

Так, дисперсия – это центральный момент порядка 2.

**Нормальное распределение и центральная предельная теорема.** В вероятностно-статистических методах принятия решений часто идет речь о нормальном распределении. Иногда его пытаются использовать для моделирования распределения исходных данных (эти попытки не всегда являются обоснованными – см. ниже). Более существенно, что многие методы обработки данных основаны на том, что расчетные величины имеют распределения, близкие к нормальному.

Пусть $X_1, X_2, ..., X_n, \ldots$– независимые одинаково распределенные случайные величины с математическими ожиданиями $M(X_i) = m$ и дисперсиями $D(X_i) = \sigma^2$, $i$ = 1, 2,…, $n$,… Как следует из результатов предыдущей главы,

$$M(X_1 + X_2 + ... + X_n) = nm, \ D(X_1 + X_2 + ... + X_n) = n\sigma^2.$$

Рассмотрим приведенную случайную величину $U_n$ для суммы $X_1 + X_2 + ... + X_n$, а именно,

$$U_n = \frac{X_1 + X_2 + ... + X_n - nm}{\sigma\sqrt{n}}.$$

Как следует из формул (7), $M(U_n) = 0$, $D(U_n) = 1$.

*Центральная предельная теорема* (для одинаково распределенных слагаемых). Пусть $X_1, X_2, ..., X_n, \ldots$– независимые одинаково распределенные случайные величины с математическими ожиданиями $M(X_i) = m$ и дисперсиями $D(X_i) = \sigma^2$, $i$ = 1, 2,…, $n$,… Тогда для любого x существует предел

$$\lim_{n\to\infty} P\left(\frac{X_1 + X_2 + ... + X_n - nm}{\sigma\sqrt{n}} < x\right) = \Phi(x),$$

где $\Phi(x)$ – функция стандартного нормального распределения.

Подробнее о функции $\Phi(x)$ – ниже (читается «фи от икс», поскольку $\Phi$ – греческая прописная буква «фи»).

Центральная предельная теорема (ЦПТ) носит свое название по той причине, что она является центральным, наиболее часто применяющимся математическим результатом теории вероятностей и математической статистики. История ЦПТ занимает около 200 лет – с 1730 г., когда английский математик А.Муавр (1667-1754) опубликовал первый результат, относящийся к ЦПТ (см. ниже о теореме Муавра-Лапласа), до двадцатых – тридцатых годов XX в., когда финн Дж.У. Линдеберг, француз Поль Леви (1886-1971), югослав В. Феллер (1906-1970), русский А.Я. Хинчин (1894-1959) и другие ученые получили необходимые и достаточные условия справедливости классической центральной предельной теоремы.

Развитие рассматриваемой тематики на этом отнюдь не прекратилось – изучали случайные величины, не имеющие дисперсии, т.е. те, для которых

$$\int_{-\infty}^{+\infty} x^2 f(x)dx = +\infty$$

(академик Б.В.Гнеденко и др.), ситуацию, когда суммируются случайные величины (точнее, случайные элементы) более сложной природы, чем числа (академики Ю.В.Прохоров, А.А.Боровков и их соратники), и т.д.

Функция распределения *Ф(x)* задается равенством

$$\Phi(x) = \int_{-\infty}^{x} \varphi(x)dx,$$

где $\varphi(y)$ - плотность стандартного нормального распределения, имеющая довольно сложное выражение:

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}.$$

Здесь $\pi$=3,1415925… - известное в геометрии число, равное отношению длины окружности к диаметру, $e$ = 2,718281828… - основание натуральных логарифмов (для запоминания этого числа обратите внимание, что 1828 – год рождения писателя Л.Н.Толстого). Как известно из математического анализа,

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n.$$

При обработке результатов наблюдений функцию нормального распределения не вычисляют по приведенным формулам, а находят с помощью специальных таблиц или компьютерных программ. Лучшие на русском языке «Таблицы математической статистики» составлены членами-корреспондентами АН СССР Л.Н. Большевым и Н.В.Смирновым [8].

Вид плотности стандартного нормального распределения $\varphi(y)$ вытекает из математической теории, которую не имеем возможности здесь рассматривать, равно как и доказательство ЦПТ.

Для иллюстрации приводим небольшие таблицы функции распределения *Ф(x)* (табл.2) и ее квантилей (табл.3). Функция *Ф(x)* симметрична относительно 0, что отражается в табл.2-3.

Таблица 2.

Функция стандартного нормального распределения.

| *x* | *Ф(x)* | *x* | *Ф(x)* | *x* | *Ф(x)* |
|---|---|---|---|---|---|
| -5,0 | 0,00000029 | -1,0 | 0,158655 | 2,0 | 0,9772499 |
| -4,0 | 0,00003167 | -0,5 | 0,308538 | 2,5 | 0,99379033 |
| -3,0 | 0,00134990 | 0,0 | 0,500000 | 3,0 | 0,99865010 |
| -2,5 | 0,00620967 | 0,5 | 0,691462 | 4,0 | 0,99996833 |
| -2,0 | 0,0227501 | 1,0 | 0,841345 | 5,0 | 0,99999971 |
| -1,5 | 0,0668072 | 1,5 | 0,9331928 | | |

Если случайная величина *X* имеет функцию распределения *Ф(x),* то *M(X)* = 0, *D(X)* = 1. Это утверждение доказывается в теории вероятностей, исходя из вида плотности вероятностей $\varphi(y)$. Оно согласуется с аналогичным утверждением для характеристик приведенной случайной величины $U_n$, что вполне естественно, поскольку ЦПТ утверждает, что при безграничном возрастании числа слагаемых функция распределения $U_n$ стремится к функции стандартного нормального распределения *Ф(x),* причем при любом *x*.

Таблица 3.

Квантили стандартного нормального распределения.

| *p* | Квантиль порядка *p* | *p* | Квантиль порядка *p* |
|---|---|---|---|
| 0,01 | -2,326348 | 0,60 | 0,253347 |
| 0,025 | -1,959964 | 0,70 | 0,524401 |
| 0,05 | -1,644854 | 0,80 | 0,841621 |
| 0,10 | -1,281552 | 0,90 | 1,281552 |
| 0,30 | -0,524401 | 0,95 | 1,644854 |

| | | | |
|---|---|---|---|
| 0,40 | -0,253347 | 0,975 | 1,959964 |
| 0,50 | 0,000000 | 0,99 | 2,326348 |

Введем понятие семейства нормальных распределений. По определению нормальным распределением называется распределение случайной величины $X$, для которой распределение приведенной случайной величины есть $\Phi(x)$. Как следует из общих свойств масштабно-сдвиговых семейств распределений (см. выше), нормальное распределение – это распределение случайной величины

$Y = \sigma X + m$,

где $X$ – случайная величина с распределением $\Phi(X)$, причем $m = M(Y)$, $\sigma^2 = D(Y)$. Нормальное распределение с параметрами сдвига $m$ и масштаба $\sigma$ обычно обозначается $N(m, \sigma)$ (иногда используется обозначение $N(m, \sigma^2)$).

Как следует из (8), плотность вероятности нормального распределения $N(m, \sigma)$ есть

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}.$$

Нормальные распределения образуют масштабно-сдвиговое семейство. При этом параметром масштаба является $d = 1/\sigma$, а параметром сдвига $c = -m/\sigma$.

Для центральных моментов третьего и четвертого порядка нормального распределения справедливы равенства

$$\mu_3 = 0, \ \mu_4 = 3\sigma^4.$$

Эти равенства лежат в основе классических методов проверки того, что результаты наблюдений подчиняются нормальному распределению. В настоящее время нормальность обычно рекомендуется проверять по критерию $W$ Шапиро – Уилка. Проблема проверки нормальности обсуждается ниже.

Если случайные величины $X_1$ и $X_2$ имеют функции распределения $N(m_1, \sigma_1)$ и $N(m_2, \sigma_2)$ соответственно, то $X_1 + X_2$ имеет распределение $N\left(m_1 + m_2; \sqrt{\sigma_1^2 + \sigma_2^2}\right)$. Следовательно, если случайные величины $X_1, X_2, \ldots, X_n$ независимы и имеют одно и тоже распределение $N(m, \sigma)$, то их среднее арифметическое

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

имеет распределение $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$. Эти свойства нормального распределения постоянно используются в различных вероятностно-статистических методах принятия решений, в частности, при статистическом регулировании технологических процессов и в статистическом приемочном контроле по количественному признаку.

С помощью нормального распределения определяются три распределения, которые в настоящее время часто используются при статистической обработке данных.

Распределение $\chi^2$ (хи - квадрат) – распределение случайной величины

$$X = X_1^2 + X_2^2 + \ldots + X_n^2,$$

где случайные величины $X_1, X_2, \ldots, X_n$ независимы и имеют одно и тоже распределение $N(0,1)$. При этом число слагаемых, т.е. $n$, называется «числом степеней свободы» распределения хи – квадрат.

Распределение $t$ Стьюдента – это распределение случайной величины

$$T = \frac{U\sqrt{n}}{\sqrt{X}},$$

где случайные величины $U$ и $X$ независимы, $U$ имеет распределение стандартное нормальное распределение $N(0,1)$, а $X$ – распределение хи – квадрат с $n$ степенями свободы. При этом $n$ называется «числом степеней свободы» распределения Стьюдента. Это распределение было введено в 1908 г. английским статистиком В. Госсетом, работавшем на фабрике, выпускающей пиво. Вероятностно-статистические методы использовались для принятия экономических и

технических решений на этой фабрике, поэтому ее руководство запрещало В. Госсету публиковать научные статьи под своим именем. Таким способом охранялась коммерческая тайна, «ноу-хау» в виде вероятностно-статистических методов, разработанных В. Госсетом. Однако он имел возможность публиковаться под псевдонимом «Стьюдент». История Госсета - Стьюдента показывает, что еще сто лет менеджерам Великобритании была очевидна большая экономическая эффективность вероятностно-статистических методов принятия решений.

Распределение Фишера – это распределение случайной величины

$$F = \frac{\frac{1}{k_1} X_1}{\frac{1}{k_2} X_2},$$

где случайные величины $X_1$ и $X_2$ независимы и имеют распределения хи – квадрат с числом степеней свободы $k_1$ и $k_2$ соответственно. При этом пара $(k_1, k_2)$ – пара «чисел степеней свободы» распределения Фишера, а именно, $k_1$ – число степеней свободы числителя, а $k_2$ – число степеней свободы знаменателя. Распределение случайной величины F названо в честь великого английского статистика Р.Фишера (1890-1962), активно использовавшего его в своих работах.

Выражения для функций распределения хи - квадрат, Стьюдента и Фишера, их плотностей и характеристик, а также таблицы можно найти в специальной литературе (см., например, [8]).

Как уже отмечалось, нормальные распределения в настоящее время часто используют в вероятностных моделях в различных прикладных областях. В чем причина такой широкой распространенности этого двухпараметрического семейства распределений? Она проясняется следующей теоремой.

*Центральная предельная теорема* (для разнораспределенных слагаемых). Пусть $X_1, X_2, \ldots, X_n, \ldots$ - независимые случайные величины с математическими ожиданиями $M(X_1), M(X_2), \ldots, M(X_n), \ldots$ и дисперсиями $D(X_1), D(X_2), \ldots, D(X_n), \ldots$ соответственно. Пусть

$$U_n = \frac{X_1 + X_2 + \ldots + X_n - M(X_1) - M(X_2) - \ldots - M(X_n)}{\sqrt{D(X_1) + D(X_2) + \ldots + D(X_n)}}.$$

Тогда при справедливости некоторых условий, обеспечивающих малость вклада любого из слагаемых в $U_n$,

$$\lim_{n \to \infty} P(U_n < x) = \Phi(x)$$

для любого $x$.

Условия, о которых идет речь, не будем здесь формулировать. Их можно найти в специальной литературе (см., например, [6]). «Выяснение условий, при которых действует ЦПТ, составляет заслугу выдающихся русских ученых А.А.Маркова (1857-1922) и, в особенности, А.М.Ляпунова (1857-1918)» [9, с.197].

Центральная предельная теорема показывает, что в случае, когда результат измерения (наблюдения) складывается под действием многих причин, причем каждая из них вносит лишь малый вклад, а совокупный итог определяется *аддитивно*, т.е. путем сложения, то распределение результата измерения (наблюдения) близко к нормальному.

Иногда считают, что для нормальности распределения достаточно того, что результат измерения (наблюдения) $X$ формируется под действием многих причин, каждая из которых оказывает малое воздействие. Это не так. Важно, как эти причины действуют. Если аддитивно – то $X$ имеет приближенно нормальное распределение. Если *мультипликативно* (т.е. действия отдельных причин перемножаются, а не складываются), то распределение $X$ близко не к нормальному, а к т.н. логарифмически нормальному, т.е. не $X$, а lg X имеет приблизительно нормальное распределение. Если же нет оснований считать, что действует один из этих двух механизмов формирования итогового результата (или какой-либо иной вполне определенный механизм), то про распределение $X$ ничего определенного сказать нельзя.

Из сказанного вытекает, что в конкретной прикладной задаче нормальность результатов измерений (наблюдений), как правило, нельзя установить из общих соображений, ее следует проверять с помощью статистических критериев. Или же использовать непараметрические статистические методы, не опирающиеся на предположения о принадлежности функций

распределения результатов измерений (наблюдений) к тому или иному параметрическому семейству.

**Непрерывные распределения, используемые в вероятностно-статистических методах принятия решений.** Кроме масштабно-сдвигового семейства нормальных распределений, широко используют ряд других семейств распределения – логарифмически нормальных, экспоненциальных, Вейбулла-Гнеденко, гамма-распределений. Рассмотрим эти семейства.

Случайная величина $X$ имеет логарифмически нормальное распределение, если случайная величина $Y = \lg X$ имеет нормальное распределение. Тогда $Z = \ln X = 2{,}3026\ldots Y$ также имеет нормальное распределение $N(a_1,y_1)$, где $\ln X$ - натуральный логарифм $X$. Плотность логарифмически нормального распределения такова:

$$f(x;a_1,\sigma_1) = \begin{cases} \dfrac{1}{\sigma_1\sqrt{2\pi}x}\exp\left[-\dfrac{(\ln x - a_1)^2}{2\sigma_1^2}\right], & x > 0, \\ 0, & x \le 0. \end{cases}$$

Из центральной предельной теоремы следует, что произведение $X = X_1X_2\ldots X_n$ независимых положительных случайных величин $X_i$, $i = 1, 2,\ldots, n$, при больших $n$ можно аппроксимировать логарифмически нормальным распределением. В частности, мультипликативная модель формирования заработной платы или дохода приводит к рекомендации приближать распределения заработной платы и дохода логарифмически нормальными законами. Для России эта рекомендация оказалась обоснованной - статистические данные подтверждают ее.

Имеются и другие вероятностные модели, приводящие к логарифмически нормальному закону. Классический пример такой модели дан А.Н.Колмогоровым [10], который из физически обоснованной системы постулатов вывел заключение о том, что размеры частиц при дроблении кусков руды, угля и т.п. на шаровых мельницах имеют логарифмически нормальное распределение.

Перейдем к другому семейству распределений, широко используемому в различных вероятностно-статистических методах принятия решений и других прикладных исследованиях, - семейству экспоненциальных распределений. Начнем с вероятностной модели, приводящей к таким распределениям. Для этого рассмотрим "поток событий", т.е. последовательность событий, происходящих одно за другим в какие-то моменты времени. Примерами могут служить: поток вызовов на телефонной станции; поток отказов оборудования в технологической цепочке; поток отказов изделий при испытаниях продукции; поток обращений клиентов в отделение банка; поток покупателей, обращающихся за товарами и услугами, и т.д. В теории потоков событий справедлива теорема, аналогичная центральной предельной теореме, но в ней речь идет не о суммировании случайных величин, а о суммировании потоков событий. Рассматривается суммарный поток, составленный из большого числа независимых потоков, ни один из которых не оказывает преобладающего влияния на суммарный поток. Например, поток вызовов, поступающих на телефонную станцию, слагается из большого числа независимых потоков вызовов, исходящих от отдельных абонентов. Доказано [6], что в случае, когда характеристики потоков не зависят от времени, суммарный поток полностью описывается одним числом $\lambda$ - интенсивностью потока. Для суммарного потока рассмотрим случайную величину $X$ - длину промежутка времени между последовательными событиями. Ее функция распределения имеет вид

$$F(x;\lambda) = P(X \le x) = \begin{cases} 1 - e^{-\lambda x}, & x \ge 0, \\ 0, & x < 0. \end{cases} \quad (10)$$

Это распределение называется экспоненциальным распределением, т.к. в формуле (10) участвует экспоненциальная функция $e^{-\lambda x}$. Величина 1/л - масштабный параметр. Иногда вводят и параметр сдвига $c$, экспоненциальным называют распределение случайной величины $X + c$, где распределение $X$ задается формулой (10).

Экспоненциальные распределения - частный случай т. н. распределений Вейбулла - Гнеденко. Они названы по фамилиям инженера В. Вейбулла, введшего эти распределения в

практику анализа результатов усталостных испытаний, и математика Б.В.Гнеденко (1912-1995), получившего такие распределения в качестве предельных при изучении максимального из результатов испытаний. Пусть $X$ - случайная величина, характеризующая длительность функционирования изделия, сложной системы, элемента (т.е. ресурс, наработку до предельного состояния и т.п.), длительность функционирования предприятия или жизни живого существа и т.д. Важную роль играет интенсивность отказа

$$\lambda(x) = \frac{f(x)}{1 - F(x)}, \quad (11)$$

где $F(x)$ и $f(x)$ - функция распределения и плотность случайной величины $X$.

Опишем типичное поведение интенсивности отказа. Весь интервал времени можно разбить на три периода. На первом из них функция $л(x)$ имеет высокие значения и явную тенденцию к убыванию (чаще всего она монотонно убывает). Это можно объяснить наличием в рассматриваемой партии единиц продукции с явными и скрытыми дефектами, которые приводят к относительно быстрому выходу из строя этих единиц продукции. Первый период называют "периодом приработки" (или "обкатки"). Именно на него обычно распространяется гарантийный срок.

Затем наступает период нормальной эксплуатации, характеризующийся приблизительно постоянной и сравнительно низкой интенсивностью отказов. Природа отказов в этот период носит внезапный характер (аварии, ошибки эксплуатационных работников и т.п.) и не зависит от длительности эксплуатации единицы продукции.

Наконец, последний период эксплуатации - период старения и износа. Природа отказов в этот период - в необратимых физико-механических и химических изменениях материалов, приводящих к прогрессирующему ухудшению качества единицы продукции и окончательному выходу ее из строя.

Каждому периоду соответствует свой вид функции $л(x)$. Рассмотрим класс степенных зависимостей

$$л(x) = л_0 b x^{b-1}, \quad (12)$$

где $л_0 > 0$ и $b > 0$ - некоторые числовые параметры. Значения $b < 1$, $b = 0$ и $b > 1$ отвечают виду интенсивности отказов в периоды приработки, нормальной эксплуатации и старения соответственно.

Соотношение (11) при заданной интенсивности отказа $л(x)$ - дифференциальное уравнение относительно функции $F(x)$. Из теории дифференциальных уравнений следует, что

$$F(x) = 1 - \exp\left\{-\int_0^x \lambda(t)dt\right\}. \quad (13)$$

Подставив (12) в (13), получим, что

$$F(x) = \begin{cases} 1 - \exp[-\lambda_0 x^b], & x \geq 0, \\ 0, & x < 0. \end{cases} \quad (14)$$

Распределение, задаваемое формулой (14) называется распределением Вейбулла - Гнеденко. Поскольку

$$\lambda_0 x^b = \left(\frac{x}{a}\right)^b,$$

где

$$a = \lambda_0^{-\frac{1}{b}}, \quad (15)$$

то из формулы (14) следует, что величина $a$, задаваемая формулой (15), является масштабным параметром. Иногда вводят и параметр сдвига, т.е. функциями распределения Вейбулла - Гнеденко называют $F(x - c)$, где $F(x)$ задается формулой (14) при некоторых $л_0$ и $b$.

Плотность распределения Вейбулла - Гнеденко имеет вид

$$f(x;a,b,c)=\begin{cases}\dfrac{b}{a}\left(\dfrac{x-c}{a}\right)^{b-1}\exp\left[-\left(\dfrac{x-c}{a}\right)^{b}\right], & x\ge c,\\ 0, & x<c,\end{cases} \qquad (16)$$

где $a > 0$ - параметр масштаба, $b > 0$ - параметр формы, $c$ - параметр сдвига. При этом параметр $a$ из формулы (16) связан с параметром $л_0$ из формулы (14) соотношением, указанным в формуле (15).

Экспоненциальное распределение - весьма частный случай распределения Вейбулла - Гнеденко, соответствующий значению параметра формы $b = 1$.

Распределение Вейбулла - Гнеденко применяется также при построении вероятностных моделей ситуаций, в которых поведение объекта определяется "наиболее слабым звеном". Подразумевается аналогия с цепью, сохранность которой определяется тем ее звеном, которое имеет наименьшую прочность. Другими словами, пусть $X_1, X_2, \ldots, X_n$ - независимые одинаково распределенные случайные величины,

$$X(1) = \min(X_1, X_2, \ldots, X_n),\ X(n) = \max(X_1, X_2, \ldots, X_n).$$

В ряде прикладных задач большую роль играют $X(1)$ и $X(n)$, в частности, при исследовании максимально возможных значений ("рекордов") тех или иных значений, например, страховых выплат или потерь из-за коммерческих рисков, при изучении пределов упругости и выносливости стали, ряда характеристик надежности и т.п. Показано, что при больших n распределения $X(1)$ и $X(n)$, как правило, хорошо описываются распределениями Вейбулла - Гнеденко. Основополагающий вклад в изучение распределений $X(1)$ и $X(n)$ внес советский математик Б.В.Гнеденко. Использованию полученных результатов в экономике, менеджменте, технике и других областях посвящены труды В. Вейбулла, Э. Гумбеля, В.Б. Невзорова, Э.М. Кудлаева и многих иных специалистов.

Перейдем к семейству гамма-распределений. Они широко применяются в экономике и менеджменте, теории и практике надежности и испытаний, в различных областях техники, метеорологии и т.д. В частности, гамма-распределению подчинены во многих ситуациях такие величины, как общий срок службы изделия, длина цепочки токопроводящих пылинок, время достижения изделием предельного состояния при коррозии, время наработки до $k$-го отказа, $k = 1, 2, \ldots$, и т.д. Продолжительность жизни больных хроническими заболеваниями, время достижения определенного эффекта при лечении в ряде случаев имеют гамма-распределение. Это распределение наиболее адекватно для описания спроса в экономико-математических моделях управления запасами (логистики).

Плотность гамма-распределения имеет вид

$$f(x;a,b,c)=\begin{cases}\dfrac{1}{\Gamma(a)}(x-c)^{a-1}b^{-a}\exp\left[-\dfrac{x-c}{b}\right], & x\ge c,\\ 0, & x<c.\end{cases} \qquad (17)$$

Плотность вероятности в формуле (17) определяется тремя параметрами $a, b, c$, где $a>0$, $b>0$. При этом $a$ является параметром формы, $b$ - параметром масштаба и $c$ - параметром сдвига. Множитель $1/Г(a)$ является нормировочным, он введен, чтобы

$$\int_{-\infty}^{+\infty} f(x;a,b,c)dx = 1.$$

Здесь $Г(a)$ - одна из используемых в математике специальных функций, так называемая "гамма-функция", по которой названо и распределение, задаваемое формулой (17),

$$\Gamma(a) = \int_{0}^{+\infty} x^{a-1}e^{-x}dx.$$

При фиксированном $a$ формула (17) задает масштабно-сдвиговое семейство распределений, порождаемое распределением с плотностью

$$f(x;a)=\begin{cases}\dfrac{1}{\Gamma(a)}x^{a-1}e^{-x}, & x\ge 0,\\ 0, & x<0.\end{cases} \qquad (18)$$

Распределение вида (18) называется стандартным гамма-распределением. Оно получается из формулы (17) при $b = 1$ и $c = 0$.

Частным случаем гамма-распределений при $a = 1$ являются экспоненциальные распределения (с $л = 1/b$). При натуральном $a$ и $c$=0 гамма-распределения называются распределениями Эрланга. С работ датского ученого К.А.Эрланга (1878-1929), сотрудника Копенгагенской телефонной компании, изучавшего в 1908-1922 гг. функционирование телефонных сетей, началось развитие теории массового обслуживания. Эта теория занимается вероятностно-статистическим моделированием систем, в которых происходит обслуживание потока заявок, с целью принятия оптимальных решений. Распределения Эрланга используют в тех же прикладных областях, в которых применяют экспоненциальные распределения. Это основано на следующем математическом факте: сумма k независимых случайных величин, экспоненциально распределенных с одинаковыми параметрами л и $c$, имеет гамма-распределение с параметром формы $a = k$, параметром масштаба $b = 1/л$ и параметром сдвига $kc$. При $c = 0$ получаем распределение Эрланга.

Если случайная величина $X$ имеет гамма-распределение с параметром формы $a$ таким, что $d = 2a$ - целое число, $b = 1$ и $c = 0$, то $2X$ имеет распределение хи-квадрат с $d$ степенями свободы.

Случайная величина $X$ с гвмма-распределением имеет следующие характеристики:

- математическое ожидание $M(X) = ab + c$,
- дисперсию $D(X) = y^2 = ab^2$,
- коэффициент вариации $v = \dfrac{b\sqrt{a}}{ab + c}$,
- асимметрию $M[(X - M(X))^3] = \dfrac{2}{\sqrt{a}}$,
- эксцесс $\dfrac{M[(X - M(X))^4]}{\sigma^4} - 3 = \dfrac{6}{a}$.

Нормальное распределение - предельный случай гамма-распределения. Точнее, пусть Z - случайная величина, имеющая стандартное гамма-распределение, заданное формулой (18). Тогда

$$\lim_{a \to \infty} P\left\{ \frac{Z - a}{\sqrt{a}} < x \right\} = \Phi(x)$$

для любого действительного числа $x$, где $Ф(x)$ - функция стандартного нормального распределения $N(0,1)$.

В прикладных исследованиях используются и другие параметрические семейства распределений, из которых наиболее известны система кривых Пирсона, ряды Эджворта и Шарлье. Здесь они не рассматриваются.

**Дискретные распределения, используемые в вероятностно-статистических методах принятия решений.** Наиболее часто используют три семейства дискретных распределений - биномиальных, гипергеометрических и Пуассона, а также некоторые другие семейства - геометрических, отрицательных биномиальных, мультиномиальных, отрицательных гипергеометрических и т.д.

Как уже говорилось, биномиальное распределение имеет место при независимых испытаниях, в каждом из которых с вероятностью $p$ появляется событие $A$. Если общее число испытаний $n$ задано, то число испытаний $Y$, в которых появилось событие $A$, имеет биномиальное распределение. Для биномиального распределения вероятность принятия случайной величиной $Y$ значения $y$ определяется формулой

$$P(Y = y \mid p, n) = \binom{n}{y} p^y (1 - p)^{n-y}, \quad y = 0,1,2,\ldots,n, \quad (19)$$

где

$$\binom{n}{y} = \frac{n!}{y!(n - y)!} = C_n^y -$$

- число сочетаний из $n$ элементов по $y$, известное из комбинаторики. Для всех $y$, кроме 0, 1, 2, …, $n$, имеем $P(Y=y)=0$. Биномиальное распределение при фиксированном объеме выборки $n$ задается параметром $p$, т.е. биномиальные распределения образуют однопараметрическое семейство. Они применяются при анализе данных выборочных исследований [2], в частности, при изучении предпочтений потребителей, выборочном контроле качества продукции по планам одноступенчатого контроля, при испытаниях совокупностей индивидуумов в демографии, социологии, медицине, биологии и др.

Если $Y_1$ и $Y_2$ - независимые биномиальные случайные величины с одним и тем же параметром $p_0$, определенные по выборкам с объемами $n_1$ и $n_2$ соответственно, то $Y_1 + Y_2$ - биномиальная случайная величина, имеющая распределение (19) с $p = p_0$ и $n = n_1 + n_2$. Это замечание расширяет область применимости биномиального распределения, позволяя объединять результаты нескольких групп испытаний, когда есть основания полагать, что всем этим группам соответствует один и тот же параметр.

Характеристики биномиального распределения вычислены ранее:

$$M(Y) = np,\ D(Y) = np(1-p).$$

В разделе "События и вероятности" для биномиальной случайной величины доказан закон больших чисел:

$$\lim_{n\to\infty} P\left\{\left|\frac{Y}{n} - p\right| \ge \varepsilon\right\} = 0$$

для любого $\varepsilon > 0$. С помощью центральной предельной теоремы закон больших чисел можно уточнить, указав, насколько $Y/n$ отличается от $p$.

*Теорема Муавра-Лапласа.* Для любых чисел a и $b$, $a<b$, имеем

$$\lim_{n\to\infty} P\left\{a \le \frac{Y - np}{\sqrt{np(1-p)}} < b\right\} = \Phi(b) - \Phi(a),$$

где $\Phi(x)$ – функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1.

Для доказательства достаточно воспользоваться представлением $Y$ в виде суммы независимых случайных величин, соответствующих исходам отдельных испытаний, формулами для $M(Y)$ и $D(Y)$ и центральной предельной теоремой.

Эта теорема для случая $p$ = S доказана английским математиком А.Муавром (1667-1754) в 1730 г. В приведенной выше формулировке она была доказана в 1810 г. французским математиком Пьером Симоном Лапласом (1749 – 1827).

Гипергеометрическое распределение имеет место при выборочном контроле конечной совокупности объектов объема N по альтернативному признаку. Каждый контролируемый объект классифицируется либо как обладающий признаком $A$, либо как не обладающий этим признаком. Гипергеометрическое распределение имеет случайная величина $Y$, равная числу объектов, обладающих признаком $A$ в случайной выборке объема $n$, где $n<N$. Например, число $Y$ дефектных единиц продукции в случайной выборке объема $n$ из партии объема $N$ имеет гипергеометрическое распределение, если $n<N$. Другой пример – лотерея. Пусть признак $A$ билета – это признак «быть выигрышным». Пусть всего билетов $N$, а некоторое лицо приобрело $n$ из них. Тогда число выигрышных билетов у этого лица имеет гипергеометрическое распределение.

Для гипергеометрического распределения вероятность принятия случайной величиной Y значения y имеет вид

$$P(Y = y \mid N, d, n) = \frac{\binom{n}{y}\binom{N-n}{D-y}}{\binom{N}{D}}, \quad (20)$$

где $D$ – число объектов, обладающих признаком $A$, в рассматриваемой совокупности объема $N$. При этом $y$ принимает значения от $\max\{0,\ n - (N - D)\}$ до $\min\{n,\ D\}$, при прочих $y$ вероятность в формуле (20) равна 0. Таким образом, гипергеометрическое распределение определяется тремя

параметрами – объемом генеральной совокупности *N,* числом объектов *D* в ней, обладающих рассматриваемым признаком *A*, и объемом выборки *n.*

Простой случайной выборкой объема *n* из совокупности объема *N* называется выборка, полученная в результате случайного отбора, при котором любой из $\binom{N}{n}$ наборов из *n* объектов имеет одну и ту же вероятность быть отобранным. Методы случайного отбора выборок респондентов (опрашиваемых) или единиц штучной продукции рассматриваются в инструктивно-методических и нормативно-технических документах. Один из методов отбора таков: объекты отбирают один из другим, причем на каждом шаге каждый из оставшихся в совокупности объектов имеет одинаковые шансы быть отобранным. В литературе для рассматриваемого типа выборок используются также термины «случайная выборка», «случайная выборка без возвращения».

Поскольку объемы генеральной совокупности (партии) *N* и выборки *n* обычно известны, то подлежащим оцениванию параметром гипергеометрического распределения является *D*. В статистических методах управления качеством продукции *D* – обычно число дефектных единиц продукции в партии. Представляет интерес также характеристика распределения *D/N* – уровень дефектности.

Для гипергеометрического распределения

$$M(Y) = n\frac{D}{N}, \quad D(Y) = n\frac{D}{N}\left(1-\frac{D}{N}\right)\left((1-\frac{n-1}{N-1}\right).$$

Последний множитель в выражении для дисперсии близок к 1, если *N>10n*. Если при этом сделать замену *p = D/N,* то выражения для математического ожидания и дисперсии гипергеометрического распределения перейдут в выражения для математического ожидания и дисперсии биномиального распределения. Это не случайно. Можно показать, что

$$P(Y = y \mid N,d,n) = \frac{\binom{n}{y}\binom{N-n}{D-y}}{\binom{N}{D}} \approx P(Y = y \mid p,n) = \binom{n}{y} p^y (1-p)^{n-y}, \;\; y = 0,1,2,...,n,$$

при *N>10n,* где *p = D/N.* Справедливо предельное соотношение

$$\lim_{N\to\infty, \frac{D}{N}\to p} P(Y = y \mid N,d,n) = P(Y = y \mid p,n), \;\; y = 0,1,2,...,n,$$

и этим предельным соотношением можно пользоваться при *N>10n.*

Третье широко используемое дискретное распределение – распределение Пуассона. Случайная величина Y имеет распределение Пуассона, если

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \;\; y = 0,1,2,...,$$

где л – параметр распределения Пуассона, и *P(Y=y)*=0 для всех прочих *y* (при у=0 обозначено 0! =1). Для распределения Пуассона

$$M(Y) = \text{л}, \; D(Y) = \text{л}.$$

Это распределение названо в честь французского математика С.Д.Пуассона (1781-1840), впервые получившего его в 1837 г. Распределение Пуассона является предельным случаем биномиального распределения, когда вероятность *p* осуществления события мала, но число испытаний *n* велико, причем *np* = л. Точнее, справедливо предельное соотношение

$$\lim_{n\to\infty, np\to\lambda} P(Y = y \mid p,n) = \frac{\lambda^y e^{-\lambda}}{y!}, \;\; y = 0,1,2,...$$

Поэтому распределение Пуассона (в старой терминологии «закон распределения») часто называют также «законом редких событий».

Распределение Пуассона возникает в теории потоков событий (см. выше). Доказано, что для простейшего потока с постоянной интенсивностью Л число событий (вызовов), происшедших за время *t*, имеет распределение Пуассона с параметром л = Л*t*. Следовательно, вероятность того, что за время *t* не произойдет ни одного события, равна $e^{-\text{Л}t}$, т.е. функция распределения длины промежутка между событиями является экспоненциальной.

Распределение Пуассона используется при анализе результатов выборочных маркетинговых обследований потребителей, расчете оперативных характеристик планов статистического приемочного контроля в случае малых значений приемочного уровня дефектности, для описания числа разладок статистически управляемого технологического процесса в единицу времени, числа «требований на обслуживание», поступающих в единицу времени в систему массового обслуживания, статистических закономерностей несчастных случаев и редких заболеваний, и т.д.

Описание иных параметрических семейств дискретных распределений и возможности их практического использования рассматриваются в литературе.

### 1.2.5. Основные проблемы прикладной статистики - описание данных, оценивание и проверка гипотез

Выделяют три основные области статистических методов обработки результатов наблюдений – описание данных, оценивание (характеристик и параметров распределений, регрессионных зависимостей и др.) и проверка статистических гипотез. Рассмотрим основные понятия, применяемые в этих областях.

**Основные понятия, используемые при описании данных.** Описание данных – предварительный этап статистической обработки. Используемые при описании данных величины применяются при дальнейших этапах статистического анализа – оценивании и проверке гипотез, а также при решении иных задач, возникающих при применении вероятностно-статистических методов принятия решений, например, при статистическом контроле качества продукции и статистическом регулировании технологических процессов.

Статистические данные – это результаты наблюдений (измерений, испытаний, опытов, анализов). Функции результатов наблюдений, используемые, в частности, для оценки параметров распределений и (или) для проверки статистических гипотез, называют «статистиками». (Для математиков надо добавить, что речь идет об измеримых функциях.) Если в вероятностной модели результаты наблюдений рассматриваются как случайные величины (или случайные элементы), то статистики, как функции случайных величин (элементов), сами являются случайными величинами (элементами). Статистики, являющиеся выборочными аналогами характеристик случайных величин (математического ожидания, медианы, дисперсии, моментов и др.) и используемые для оценивания этих характеристик, называют статистическими характеристиками.

Основополагающее понятие в вероятностно-статистических методах принятия решений – выборка. Как уже говорилось, выборка – это 1) набор наблюдаемых значений или 2) множество объектов, отобранные из изучаемой совокупности. Например, единицы продукции, отобранные из контролируемой партии или потока продукции для контроля и принятия решений. Наблюдаемые значения обозначим $x_1, x_2, \ldots, x_n$, где $n$ – объем выборки, т.е. число наблюдаемых значений, составляющих выборку. О втором виде выборок уже шла речь при рассмотрении гипергеометрического распределения, когда под выборкой понимался набор единиц продукции, отобранных из партии. Там же обсуждалась вероятностная модель случайной выборки.

В вероятностной модели выборки первого вида наблюдаемые значения обычно рассматривают как реализацию независимых одинаково распределенных случайных величин $X_1(\omega), X_2(\omega), \ldots, X_n(\omega), \omega \in \Omega$. При этом считают, что полученные при наблюдениях конкретные значения $x_1, x_2, \ldots, x_n$ соответствуют определенному элементарному событию $\omega = \omega_0$, т.е.

$$x_1 = X_1(\omega_0), x_2 = X_2(\omega_0), \ldots, x_n = X_n(\omega_0), \omega_0 \in \Omega .$$

При повторных наблюдениях будут получены иные наблюдаемые значения, соответствующие другому элементарному событию $\omega = \omega_1$. Цель обработки статистических данных состоит в том, чтобы по результатам наблюдений, соответствующим элементарному событию $\omega = \omega_0$, сделать выводы о вероятностной мере $P$ и результатах наблюдений при различных возможных $\omega = \omega_1$.

Применяют и другие, более сложные вероятностные модели выборок. Например, цензурированные выборки соответствуют испытаниям, проводящимся в течение определенного

промежутка времени. При этом для части изделий удается замерить время наработки на отказ, а для остальных лишь констатируется, что наработки на отказ для них больше времени испытания. Для выборок второго вида отбор объектов может проводиться в несколько этапов. Например, для входного контроля сигарет могут сначала отбираться коробки, в отобранных коробках – блоки, в выбранных блоках – пачки, а в пачках – сигареты. Четыре ступени отбора. Ясно, что выборка будет обладать иными свойствами, чем простая случайная выборка из совокупности сигарет.

Из приведенного выше определения математической статистики следует, что описание статистических данных дается с помощью частот. Частота – это отношение числа $X$ наблюдаемых единиц, которые принимают заданное значение или лежат в заданном интервале, к общему числу наблюдений $n$, т.е. частота – это $X/n$. (В более старой литературе иногда $X/n$ называется относительной частотой, а под частотой имеется в виду $X$. В старой терминологии можно сказать, что относительная частота – это отношение частоты к общему числу наблюдений.)

Отметим, что обсуждаемое определение приспособлено к нуждам одномерной статистики. В случае многомерного статистического анализа, статистики случайных процессов и временных рядов, статистики объектов нечисловой природы нужны несколько иные определения понятия «статистические данные». Не считая нужным давать такие определения, отметим, что в подавляющем большинстве практических постановок исходные статистические данные – это выборка или несколько выборок. А выборка – это конечная совокупность соответствующих математических объектов (чисел, векторов, функций, объектов нечисловой природы).

Число $X$ имеет биномиальное распределение, задаваемое вероятностью $p$ того, что случайная величина, с помощью которой моделируются результаты наблюдений, принимает заданное значение или лежит в заданном интервале, и общим числом наблюдений $n$. Из закона больших чисел (теорема Бернулли) следует, что

$$\frac{X}{n} \to p$$

при n→∞ (сходимость по вероятности), т.е. частота сходится к вероятности. Теорема Муавра-Лапласа позволяет уточнить скорость сходимости в этом предельном соотношении.

Чтобы от отдельных событий перейти к одновременному рассмотрению многих событий, используют накопленную частоту. Так называется отношение числа единиц, для которых результаты наблюдения меньше заданного значения, к общему числу наблюдений. (Это понятие используется, если результаты наблюдения – действительные числа, а не вектора, функции или объекты нечисловой природы.) Функция, которая выражает зависимость между значениями количественного признака и накопленной частотой, называется эмпирической функцией распределения. Итак, эмпирической функцией распределения $F_n(x)$ называется доля элементов выборки, меньших $x$. Эмпирическая функция распределения содержит всю информацию о результатах наблюдений.

Чтобы записать выражение для эмпирической функции распределения в виде формулы, введем функцию $c(x, y)$ двух переменных:

$$c(x,y) = \begin{cases} 0, & x \le y, \\ 1, & x > y. \end{cases}$$

Случайные величины, моделирующие результаты наблюдений, обозначим $X_1(\omega), X_2(\omega), ..., X_n(\omega), \omega \in \Omega$. Тогда эмпирическая функция распределения $F_n(x)$ имеет вид

$$F_n(x) = F_n(x,\omega) = \frac{1}{n} \sum_{1 \le i \le n} c(x, X_i(\omega)).$$

Из закона больших чисел следует, что для каждого действительного числа $x$ эмпирическая функция распределения $F_n(x)$ сходится к функции распределения $F(x)$ результатов наблюдений, т.е.

$$F_n(x) \to F(x) \quad (1)$$

при $n \to \infty$. Советский математик В.И. Гливенко (1897-1940) доказал в 1933 г. более сильное утверждение: сходимость в (1) равномерна по x, т.е.

$$\sup_x | F_n(x) - F(x) | \to 0 \quad (2)$$

при $n \to \infty$ (сходимость по вероятности).

В (2) использовано обозначение sup (читается как «супремум»). Для функции *g(x)* под $\sup_x g(x)$ понимают наименьшее из чисел *a* таких, что $g(x) \le a$ при всех *x*. Если функция *g(x)* достигает максимума в точке $x_0$, то $\sup_x g(x) = g(x_0)$. В таком случае вместо sup пишут max. Хорошо известно, что не все функции достигают максимума.

В том же 1933 г. А.Н.Колмогоров усилил результат В.И. Гливенко для непрерывных функций распределения *F(x)*. Рассмотрим случайную величину

$$D_n = \sqrt{n} \sup_x | F_n(x) - F_0(x) |$$

и ее функцию распределения

$$K_n(x) = P\{D_n \le x\}.$$

По теореме А.Н.Колмогорова

$$\lim_{n\to\infty} K_n(x) = K(x)$$

при каждом *x*, где *K(x)* – т.н. функция распределения Колмогорова.

Рассматриваемая работа А.Н. Колмогорова породила одно из основных направлений математической статистики – т.н. непараметрическую статистику. И в настоящее время непараметрические критерии согласия Колмогорова, Смирнова, омега-квадрат широко используются. Они были разработаны для проверки согласия с *полностью известным* теоретическим распределением, т.е. предназначены для проверки гипотезы $H_0 : F(x) \equiv F_0(x)$. Основная идея критериев Колмогорова, омега-квадрат и аналогичных им состоит в измерении расстояния между функцией эмпирического распределения и функцией теоретического распределения. Различаются эти критерии видом расстояний в пространстве функций распределения. Аналитические выражения для предельных распределений статистик, расчетные формулы, таблицы распределений и критических значений широко распространены [8], поэтому не будем их приводить.

Кроме эмпирической функции распределения, для описания данных используют и другие статистические характеристики. В качестве выборочных средних величин постоянно используют выборочное среднее арифметическое, т.е. сумму значений рассматриваемой величины, полученных по результатам испытания выборки, деленную на ее объем:

$$\bar{x} = \frac{1}{n} \sum_{1 \le i \le n} x_i,$$

где *n* – объем выборки, $x_i$ – результат измерения (испытания) *i*-ого элемента выборки.

Другой вид выборочного среднего – выборочная медиана. Она определяется через порядковые статистики.

Порядковые статистики – это члены вариационного ряда, который получается, если элементы выборки $x_1, x_2, \ldots, x_n$ расположить в порядке неубывания:

$$x(1) \le x(2) \le \ldots \le x(k) \le \ldots \le x(n).$$

*Пример 1.* Для выборки $x_1 = 1$, $x_2 = 7$, $x_3 = 4$, $x_4 = 2$, $x_5 = 8$, $x_6 = 0$, $x_7 = 5$, $x_8 = 7$ вариационный ряд имеет вид 0, 1, 2, 4, 5, 7, 7, 8, т.е. $x(1) = 0 = x_6$, $x(2) = 1 = x_1$, $x(3) = 2 = x_4$, $x(4) = 4 = x_3$, $x(5) = 5 = x_7$, $x(6) = x(7) = 7 = x_2 = x_8$, $x(8) = 8 = x_5$.

В вариационном ряду элемент *x(k)* называется *k*-той порядковой статистикой. Порядковые статистики и функции от них широко используются в вероятностно-статистических методах принятия решений, в эконометрике и в других прикладных областях [2].

Выборочная медиана $\tilde{x}$ - результат наблюдения, занимающий центральное место в вариационном ряду, построенном по выборке с нечетным числом элементов, или полусумма двух результатов наблюдений, занимающих два центральных места в вариационном ряду, построенном по выборке с четным числом элементов. Таким образом, если объем выборки *n* – нечетное число, $n = 2k+1$, то медиана $\tilde{x} = x(k+1)$, если же *n* – четное число, $n = 2k$, то медиана $\tilde{x} = [x(k) + x(k+1)]/2$, где $x(k)$ и $x(k+1)$ – порядковые статистики.

В качестве выборочных показателей рассеивания результатов наблюдений чаще всего используют выборочную дисперсию, выборочное среднее квадратическое отклонение и размах выборки.

Согласно [8] выборочная дисперсия $s^2$ – это сумма квадратов отклонений выборочных результатов наблюдений от их среднего арифметического, деленная на объем выборки:

$$s^2 = \frac{1}{n}\sum_{1\le i\le n}(x_i - \bar{x})^2.$$

Выборочное среднее квадратическое отклонение $s$ – неотрицательный квадратный корень из дисперсии, т.е. $s = +\sqrt{s^2}$.

В некоторых литературных источниках выборочной дисперсией называют другую величину:

$$s_0^2 = \frac{1}{n-1}\sum_{1\le i\le n}(x_i - \bar{x})^2.$$

Она отличается от $s^2$ постоянным множителем:

$$s^2 = \left(1 - \frac{1}{n}\right)s_0^2.$$

Соответственно выборочным средним квадратическим отклонением в этих литературных источниках называют величину $s_0 = +\sqrt{s_0^2}$. Тогда, очевидно,

$$s = \sqrt{1 - \frac{1}{n}}\,s_0.$$

Различие в определениях приводит к различию в алгоритмах расчетов, правилах принятия решений и соответствующих таблицах. Поэтому при использовании тех или иных нормативно-технических и инструктивно-методических материалов, программных продуктов, таблиц необходимо обращать внимание на способ определения выборочных характеристик.

Выбор $s_0^2$, а не $s^2$, объясняется тем, что

$$M(s_0^2) = D(X) = \sigma^2,$$

где $X$ – случайная величина, имеющая такое же распределение, как и результаты наблюдений. В терминах теории статистического оценивания это означает, что $s_0^2$ - несмещенная оценка дисперсии (см. ниже). В то же время статистика $s^2$ не является несмещенной оценкой дисперсии результатов наблюдений, поскольку

$$M(s^2) = \left(1 - \frac{1}{n}\right)\sigma^2.$$

Однако у $s^2$ есть другое свойство, оправдывающее использование этой статистики в качестве выборочного показателя рассеивания. Для известных результатов наблюдений $x_1, x_2, \ldots, x_n$ рассмотрим случайную величину $Y$ с распределением вероятностей

$$P(Y = x_i) = \frac{1}{n}, \quad i = 1,2,\ldots,n,$$

и $P(Y = x) = 0$ для всех прочих $x$. Это распределение вероятностей называется эмпирическим. Тогда функция распределения $Y$ – это эмпирическая функция распределения, построенная по результатам наблюдений $x_1, x_2, \ldots, x_n$. Вычислим математическое ожидание и дисперсию случайной величины $Y$:

$$M(Y) = \bar{x}, \quad D(Y) = s^2.$$

Второе из этих равенств и является основанием для использования $s^2$ в качестве выборочного показателя рассеивания.

Отметим, что математические ожидания выборочных средних квадратических отклонений $M(s)$ и $M(s_0)$, вообще говоря, не равняются теоретическому среднему квадратическому отклонению у. Например, если $X$ имеет нормальное распределение, объем выборки $n = 3$, то

$$M(s) = 0{,}724\sigma, \quad M(s_0) = 0{,}887\sigma.$$

Кроме перечисленных выше статистических характеристик, в качестве выборочного показателя рассеивания используют размах $R$ – разность между $n$-й и первой порядковыми статистиками в выборке объема $n$, т.е. разность между наибольшим и наименьшим значениями в выборке: $R = x(n) – x(1)$.

В ряде вероятностно-статистических методов принятия решений применяют и иные показатели рассеивания. В частности, в методах статистического регулирования процессов используют средний размах – среднее арифметическое размахов, полученных в определенном количестве выборок одинакового объема. Популярно и межквартильное расстояние, т.е. расстояние между выборочными квартилями $x([0,75n])$ и $x([0,25n])$ порядка 0,75 и 0,25 соответственно, где $[0,75n]$ – целая часть числа $0,75n$, а $[0,25n]$ –целая часть числа $0,25n$.

**Основные понятия, используемые при оценивании**. Оценивание – это определение приближенного значения неизвестной характеристики или параметра распределения (генеральной совокупности), иной оцениваемой составляющей математической модели реального (экономического, технического и др.) явления или процесса по результатам наблюдений. Иногда формулируют более коротко: оценивание – это определение приближенного значения неизвестного параметра генеральной совокупности по результатам наблюдений. При этом параметром генеральной совокупности может быть либо число, либо набор чисел (вектор), либо функция, либо множество или иной объект нечисловой природы. Например, по результатам наблюдений, распределенных согласно биномиальному закону, оценивают число – параметр $p$ (вероятность успеха). По результатам наблюдений, имеющих гамма-распределение, оценивают набор из трех чисел – параметры формы $a$, масштаба $b$ и сдвига $c$. Способ оценивания функции распределения дается теоремами В.И. Гливенко и А.Н. Колмогорова. Оценивают также плотности вероятности, функции, выражающие зависимости между переменными, включенными в вероятностные модели экономических, управленческих или технологических процессов, и т.д. Целью оценивания может быть нахождение упорядочения инвестиционных проектов по экономической эффективности или технических изделий (объектов) по качеству, формулировка правил технической или медицинской диагностики и т.д. (Упорядочения в математической статистике называют также ранжировками. Это – один из видов объектов нечисловой природы.)

Оценивание проводят с помощью оценок – статистик, являющихся основой для оценивания неизвестного параметра распределения. В ряде литературных источников термин «оценка» встречается в качестве синонима термина «оценивание». Употреблять одно и то же слово для обозначения двух разных понятий нецелесообразно: оценивание – это действие, а оценка – статистика (функция от результатов наблюдений), используемая в процессе указанного действия или являющаяся его результатом.

Оценивание бывает двух видов – точечное оценивание и оценивание с помощью доверительной области.

Точечное оценивание - способ оценивания, заключающийся в том, что значение оценки принимается как неизвестное значение параметра распределения.

*Пример 2.* Пусть результаты наблюдений $x_1, x_2, \ldots, x_n$ рассматривают в вероятностной модели как случайную выборку из нормального распределения $N(m,y)$. Т.е. считают, что результаты наблюдений моделируются как реализации $n$ независимых одинаково распределенных случайных величин, имеющих функцию нормального распределения $N(m,y)$ с некоторыми математическим ожиданием $m$ и средним квадратическим отклонением $y$, неизвестными статистику. Требуется оценить параметры $m$ и $y$ (или $y^2$) по результатам наблюдений. Оценки обозначим $m^*$ и $(y^2)^*$ соответственно. Обычно в качестве оценки $m^*$ математического ожидания $m$ используют выборочное среднее арифметическое $\bar{x}$, а в качестве оценки $(y^2)^*$ дисперсии $y^2$ используют выборочную дисперсию $s^2$, т.е.

$$m^* = \bar{x}, \ (y^2)^* = s^2.$$

Для оценивания математического ожидания $m$ могут использоваться и другие статистики, например, выборочная медиана $\tilde{x}$, полусумма минимального и максимального членов вариационного ряда

$$m^{**} = [x(1)+x(n)]/2$$

и др. Для оценивания дисперсии $y^2$ также имеется ряд оценок, в частности, $s_0^2$ (см. выше) и оценка, основанная на размахе $R$, имеющая вид

$$(y^2)^{**} = [a(n)R]^2,$$

где коэффициенты $a(n)$ берут из специальных таблиц [8]. Эти коэффициенты подобраны так, чтобы для выборок из нормального распределения

$$M[a(n)R] = y.$$

Наличие нескольких методов оценивания одних и тех же параметров приводит к необходимости выбора между этими методами.

Как сравнивать методы оценивания между собой? Сравнение проводят на основе таких показателей качества методов оценивания, как состоятельность, несмещенность, эффективность и др.

Рассмотрим оценку $и_n$ числового параметра и, определенную при $n = 1, 2, \ldots$ Оценка $и_n$ называется *состоятельной*, если она сходится по вероятности к значению оцениваемого параметра и при безграничном возрастании объема выборки. Выразим сказанное более подробно. Статистика $и_n$ является состоятельной оценкой параметра и тогда и только тогда, когда для любого положительного числа е справедливо предельное соотношение

$$\lim_{n\to\infty} P\{|\theta_n - \theta| > \varepsilon\} = 0.$$

*Пример 3.* Из закона больших чисел следует, что $и_n = \bar{x}$ является состоятельной оценкой $и = M(X)$ (в приведенной выше теореме Чебышёва предполагалось существование дисперсии $D(X)$; однако, как доказал А.Я. Хинчин [6], достаточно выполнения более слабого условия – существования математического ожидания $M(X)$).

*Пример 4.* Все указанные выше оценки параметров нормального распределения являются состоятельными.

Вообще, все (за редчайшими исключениями) оценки параметров, используемые в вероятностно-статистических методах принятия решений, являются состоятельными.

*Пример 5*. Так, согласно теореме В.И. Гливенко, эмпирическая функция распределения $F_n(x)$ является состоятельной оценкой функции распределения результатов наблюдений $F(x)$.

При разработке новых методов оценивания следует в первую очередь проверять состоятельность предлагаемых методов.

Второе важное свойство оценок – *несмещенность*. Несмещенная оценка $и_n$ – это оценка параметра и, математическое ожидание которой равно значению оцениваемого параметра: $M(и_n) = и$.

*Пример 6.* Из приведенных выше результатов следует, что $\bar{x}$ и $s_0^2$ являются несмещенными оценками параметров $m$ и $y^2$ нормального распределения. Поскольку $M(\tilde{x}) = M(m^{**}) = m$, то выборочная медиана $\tilde{x}$ и полусумма крайних членов вариационного ряда $m^{**}$ - также несмещенные оценки математического ожидания $m$ нормального распределения. Однако

$$M(s^2) \neq \sigma^2, \quad M[(\sigma^2)^{**}] \neq \sigma^2,$$

поэтому оценки $s^2$ и $(y^2)^{**}$ не являются состоятельными оценками дисперсии $y^2$ нормального распределения.

Оценки, для которых соотношение $M(и_n) = и$ неверно, называются смещенными. При этом разность между математическим ожиданием оценки $и_n$ и оцениваемым параметром и, т.е. $M(и_n) - и$, называется смещением оценки.

*Пример 7.* Для оценки $s^2$, как следует из сказанного выше, смещение равно

$$M(s^2) - y^2 = -y^2/n.$$

Смещение оценки $s^2$ стремится к 0 при $n \to \infty$.

Оценка, для которой смещение стремится к 0, когда объем выборки стремится к бесконечности, называется *асимптотически несмещенной*. В примере 7 показано, что оценка $s^2$ является асимптотически несмещенной.

Практически все оценки параметров, используемые в вероятностно-статистических методах принятия решений, являются либо несмещенными, либо асимптотически несмещенными. Для несмещенных оценок показателем точности оценки служит дисперсия – чем дисперсия меньше, тем оценка лучше. Для смещенных оценок показателем точности служит

математическое ожидание квадрата оценки $M(и_n - и)^2$. Как следует из основных свойств математического ожидания и дисперсии,

$$d_n(\theta_n) = M[(\theta_n - \theta)^2] = D(\theta_n) + (M(\theta_n) - \theta)^2, \quad (3)$$

т.е. математическое ожидание квадрата ошибки складывается из дисперсии оценки и квадрата ее смещения.

Для подавляющего большинства оценок параметров, используемых в вероятностно-статистических методах принятия решений, дисперсия имеет порядок $1/n$, а смещение – не более чем $1/n$, где $n$ – объем выборки. Для таких оценок при больших $n$ второе слагаемое в правой части (3) пренебрежимо мало по сравнению с первым, и для них справедливо приближенное равенство

$$d_n(\theta_n) = M[(\theta_n - \theta)^2] \approx D(\theta_n) \approx \frac{c}{n}, \quad c = c(\theta_n, \theta), \quad (4)$$

где $c$ – число, определяемое методом вычисления оценок $и_n$ и истинным значением оцениваемого параметра и.

С дисперсией оценки связано третье важное свойство метода оценивания – *эффективность*. Эффективная оценка – это несмещенная оценка, имеющая наименьшую дисперсию из всех возможных несмещенных оценок данного параметра.

Доказано [11], что $\bar{x}$ и $s_0^2$ являются эффективными оценками параметров $m$ и $y^2$ нормального распределения. В то же время для выборочной медианы $\tilde{x}$ справедливо предельное соотношение

$$\lim_{n\to\infty} \frac{D(\bar{x})}{D(\tilde{x})} = \frac{2}{\pi} \approx 0{,}637.$$

Другими словами, эффективность выборочной медианы, т.е. отношение дисперсии эффективной оценки $\bar{x}$ параметра $m$ к дисперсии несмещенной оценки $\tilde{x}$ этого параметра при больших n близка к 0,637. Именно из-за сравнительно низкой эффективности выборочной медианы в качестве оценки математического ожидания нормального распределения обычно используют выборочное среднее арифметическое.

Понятие эффективности вводится для несмещенных оценок, для которых $M(и_n) = и$ для всех возможных значений параметра и. Если не требовать несмещенности, то можно указать оценки, при некоторых и имеющие меньшую дисперсию и средний квадрат ошибки, чем эффективные.

*Пример 8.* Рассмотрим «оценку» математического ожидания $m_1 \equiv 0$. Тогда $D(m_1) = 0$, т.е. всегда меньше дисперсии $D(\bar{x})$ эффективной оценки $\bar{x}$. Математическое ожидание среднего квадрата ошибки $d_n(m_1) = m^2$, т.е. при $|m| < \sigma/\sqrt{n}$ имеем $d_n(m_1) < d_n(\bar{x})$. Ясно, однако, что статистику $m_1 \equiv 0$ бессмысленно рассматривать в качестве оценки математического ожидания $m$.

*Пример 9.* Более интересный пример рассмотрен американским математиком Дж. Ходжесом:

$$T_n = \begin{cases} \bar{x}, & |\bar{x}| > n^{-1/4}, \\ 0{,}5\bar{x}, & |\bar{x}| \le n^{-1/4}. \end{cases}$$

Ясно, что $T_n$ – состоятельная, асимптотически несмещенная оценка математического ожидания $m$, при этом, как нетрудно вычислить,

$$\lim_{n\to\infty} n d_n(T_n) = \begin{cases} \sigma^2, & m \neq 0, \\ \dfrac{\sigma^2}{4}, & m = 0. \end{cases}$$

Последняя формула показывает, что при $m \neq 0$ оценка $T_n$ не хуже $\bar{x}$ (при сравнении по среднему квадрату ошибки $d_n$), а при $m = 0$ – в четыре раза лучше.

Подавляющее большинство оценок $и_n$, используемых в вероятностно-статистических методах принятия решений, являются асимптотически нормальными, т.е. для них справедливы предельные соотношения:

$$\lim_{n\to\infty} P\left\{ \frac{\theta_n - M(\theta_n)}{\sqrt{D(\theta_n)}} < x \right\} = \Phi(x)$$

для любого $x$, где $\Phi(x)$ – функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Это означает, что для больших объемов выборок (практически - несколько десятков или сотен наблюдений) распределения оценок полностью описываются их математическими ожиданиями и дисперсиями, а качество оценок – значениями средних квадратов ошибок $d_n(\text{и}_n)$.

Наилучшими асимптотически нормальными оценками, сокращенно НАН-оценками, называются те, для которых средний квадрат ошибки $d_n(\text{и}_n)$ принимает при больших объемах выборки наименьшее возможное значение, т.е. величина $c = c(\text{и}_n,\text{и})$ в формуле (4) минимальна. Ряд видов оценок – так называемые одношаговые оценки и оценки максимального правдоподобия – являются НАН-оценками, именно они обычно используются в вероятностно-статистических методах принятия решений.

Какова точность оценки параметра? В каких границах он может лежать? В нормативно-технической и инструктивно-методической документации, в таблицах и программных продуктах наряду с алгоритмами расчетов точечных оценок даются правила нахождения доверительных границ. Они и указывают точность точечной оценки. При этом используются такие термины, как доверительная вероятность, доверительный интервал. Если речь идет об оценивании нескольких числовых параметров, или же функции, упорядочения и т.п., то говорят об оценивании с помощью доверительной области.

*Доверительная область* – это область в пространстве параметров, в которую с заданной вероятностью входит неизвестное значение оцениваемого параметра распределения. «Заданная вероятность» называется *доверительной вероятностью* и обычно обозначается г. Пусть И – пространство параметров. Рассмотрим статистику $\text{И}_1 = \text{И}_1(x_1, x_2, \ldots, x_n)$ – функцию от результатов наблюдений $x_1, x_2, \ldots, x_n$, значениями которой являются подмножества пространства параметров И. Так как результаты наблюдений – случайные величины, то $\text{И}_1$ – также случайная величина, значения которой – подмножества множества И, т.е. $\text{И}_1$ – случайное множество. Напомним, что множество – один из видов объектов нечисловой природы, случайные множества изучают в теории вероятностей и статистике объектов нечисловой природы.

В ряде литературных источников, к настоящему времени во многом устаревших, под случайными величинами понимают только те из них, которые в качестве значений принимают действительные числа. Согласно справочнику академика РАН Ю.В.Прохорова и проф. Ю.А.Розанова [12] случайные величины могут принимать значения из любого множества. Так, случайные вектора, случайные функции, случайные множества, случайные ранжировки (упорядочения) – это отдельные виды случайных величин. Используется и иная терминология: термин «случайная величина» сохраняется только за числовыми функциями, определенными на пространстве элементарных событий, а в случае иных областей значений используется термин «случайный элемент». (Замечание для математиков: все рассматриваемые функции, определенные на пространстве элементарных событий, предполагаются измеримыми.)

Статистика $\text{И}_1$ называется *доверительной областью*, соответствующей доверительной вероятности г, если

$$P\{\theta \in \Theta_1(x_1, x_2, \ldots, x_n)\} = \gamma. \quad (5)$$

Ясно, что этому условию удовлетворяет, как правило, не одна, а много доверительных областей. Из них выбирают для практического применения какую-либо одну, исходя из дополнительных соображений, например, из соображений симметрии или минимизируя объем доверительной области, т.е. меру множества $\text{И}_1$.

При оценке одного числового параметра в качестве доверительных областей обычно применяют доверительные интервалы (в том числе лучи), а не иные типа подмножеств прямой. Более того, для многих двухпараметрических и трехпараметрических распределений (нормальных, логарифмически нормальных, Вейбулла-Гнеденко, гамма-распределений и др.) обычно используют точечные оценки и построенные на их основе доверительные границы для каждого из двух или трех параметров отдельно. Это делают для удобства пользования результатами расчетов: доверительные интервалы легче применять, чем фигуры на плоскости или тела в трехмерном пространстве.

Как следует из сказанного выше, *доверительный интервал* – это интервал, который с заданной вероятностью накроет неизвестное значение оцениваемого параметра распределения. Границы доверительного интервала называют *доверительными границами*. Доверительная

вероятность г – вероятность того, что доверительный интервал накроет действительное значение параметра, оцениваемого по выборочным данным. Оцениванием с помощью доверительного интервала называют способ оценки, при котором с заданной доверительной вероятностью устанавливают границы доверительного интервала.

Для числового параметра и рассматривают верхнюю доверительную границу $и_В$, нижнюю доверительную границу $и_Н$ и двусторонние доверительные границы – верхнюю $и_{1В}$ и нижнюю $и_{1Н}$. Все четыре доверительные границы – функции от результатов наблюдений $x_1, x_2, \ldots, x_n$ и доверительной вероятности г.

Верхняя доверительная граница $и_В$ – случайная величина $и_В = и_В(x_1, x_2, \ldots, x_n; г)$, для которой $Р(и \le и_В) = г$, где и – истинное значение оцениваемого параметра. Доверительный интервал в этом случае имеет вид $(-\infty; и_В]$.

Нижняя доверительная граница $и_Н$ – случайная величина $и_Н = и_Н(x_1, x_2, \ldots, x_n; г)$, для которой $Р(и \ge и_Н) = г$, где и – истинное значение оцениваемого параметра. Доверительный интервал в этом случае имеет вид $[и_Н; +\infty)$.

Двусторонние доверительные границы - верхняя $и_{1В}$ и нижняя $и_{1Н}$ - это случайные величины $и_{1В} = и_{1В}(x_1, x_2, \ldots, x_n; г)$ и $и_{1Н} = и_{1Н}(x_1, x_2, \ldots, x_n; г)$ такие, что $Р(и_{1Н} \le и \le и_{1В}) = г$, где и – истинное значение оцениваемого параметра. Доверительный интервал в этом случае имеет вид $[и_{1Н}; и_{1В}]$.

Вероятности, связанные с доверительными границами, можно записать в виде частных случаев формулы (5):

$$P\{\theta \in (-\infty; \theta_B]\} = \gamma, \;\; P\{\theta \in [\theta_H; +\infty)\} = \gamma, \;\; P\{\theta \in [\theta_H; \theta_B]\} = \gamma.$$

В нормативно-технической и инструктивно-методической документации, научной и учебной литературе используют два типа правил определения доверительных границ – построенных на основе точного распределения и построенных на основе асимптотического распределения некоторой точечной оценки $и_n$ параметра и. Рассмотрим примеры.

*Пример 10*. Пусть $x_1, x_2, \ldots, x_n$ – выборка из нормального закона $N(m, у)$, параметры $m$ и у неизвестны. Укажем доверительные границы для $m$.

Известно [11], что случайная величина

$$Y = \sqrt{n}\,\frac{\bar{x} - m}{s_0}$$

имеет распределение Стьюдента с (т-1) степенью свободы, где $\bar{x}$ - выборочное среднее арифметическое и $s_0$ – выборочное среднее квадратическое отклонение. Пусть $t_\gamma(n-1)$ и $t_{1-\gamma}(n-1)$ - квантили указанного распределения порядка г и 1-г соответственно. Тогда

$$P\{Y \le t_г(n\text{-}1)\} = г,\; P\{Y \ge t_{1\text{-}г}(n\text{-}1)\} = г.$$

Следовательно,

$$P\left\{m \ge \bar{x} - t_\gamma(n-1)\frac{s_0}{\sqrt{n}}\right\} = \gamma,$$

т.е. в качестве нижней доверительной границы $и_Н$, соответствующей доверительной вероятности г, следует взять

$$\theta_H(x_1, x_2, \ldots, x_n; \gamma) = \bar{x} - t_\gamma(n-1)\frac{s_0}{\sqrt{n}}. \quad (6)$$

Аналогично получаем, что

$$P\left\{m \le \bar{x} - t_{1-\gamma}(n-1)\frac{s_0}{\sqrt{n}}\right\} = \gamma.$$

Поскольку распределение Стьюдента симметрично относительно 0, то $t_{1-\gamma}(n-1) = -t_{1-\gamma}(n-1)\,t_\gamma(n-1)$. Следовательно, в качестве верхней доверительной границы $и_В$ для $m$, соответствующей доверительной вероятности г, следует взять

$$\theta_B(x_1, x_2, \ldots, x_n; \gamma) = \bar{x} + t_\gamma(n-1)\frac{s_0}{\sqrt{n}}. \quad (7)$$

Как построить двусторонние доверительные границы? Положим

$$\theta_{1H} = \theta_H(x_1, x_2, ..., x_n; \gamma_1),\ \theta_{1B} = \theta_B(x_1, x_2, ..., x_n; \gamma_2),$$

где $и_{1H}$ и $и_{1B}$ заданы формулами (6) и (7) соответственно. Поскольку неравенство $и_{1H} \le m \le и_{1B}$ выполнено тогда и только тогда, когда

$$t_{\gamma_2}(n-1) \ge Y \ge t_{1-\gamma_1}(n-1),$$

то

$$P\{и_{1H} \le m \le и_{1B}\} = г_1 + г_2 - 1,$$

(в предположении, что $г_1 > 0{,}5$; $г_2 > 0{,}5$). Следовательно, если $г = г_1 + г_2 - 1$, то $и_{1H}$ и $и_{1B}$ – двусторонние доверительные границы для $m$, соответствующие доверительной вероятности г. Обычно полагают $г_1 = г_2$, т.е. в качестве двусторонних доверительных границ $и_{1H}$ и $и_{1B}$, соответствующих доверительной вероятности г, используют односторонние доверительные границы $и_H$ и $и_B$, соответствующие доверительной вероятности $(1+г)/2$.

Другой вид правил построения доверительных границ для параметра и основан на асимптотической нормальности некоторой точечной оценки $и_n$ этого параметра. В вероятностно-статистических методах принятия решений используют, как уже отмечалось, несмещенные или асимптотически несмещенные оценки $и_n$, для которых смещение либо равно 0, либо при больших объемах выборки пренебрежимо мало по сравнению со средним квадратическим отклонением оценки $и_n$. Для таких оценок при всех $x$

$$\lim_{n\to\infty} P\left\{\frac{\theta_n - \theta}{\sqrt{D(\theta_n)}} \le x\right\} = \Phi(x),$$

где $\Phi(x)$ – функция нормального распределения $N(0;1)$. Пусть $u_г$ – квантиль порядка г распределения $N(0;1)$. Тогда

$$\lim_{n\to\infty} P\left\{\frac{\theta_n - \theta}{\sqrt{D(\theta_n)}} \le u_\gamma\right\} = \gamma \quad (8)$$

Поскольку неравенство

$$\frac{\theta_n - \theta}{\sqrt{D(\theta_n)}} \le u_\gamma$$

равносильно неравенству

$$\theta_n - u_\gamma \sqrt{D(\theta_n)} \le \theta,$$

то в качестве $и_H$ можно было бы взять левую часть последнего неравенства. Однако точное значение дисперсии $D(и_n)$ обычно неизвестно. Зато часто удается доказать, что дисперсия оценки имеет вид

$$D(\theta_n) = \frac{h(\theta)}{n}$$

(с точностью до пренебрежимо малых при росте $n$ слагаемых), где $h(и)$ – некоторая функция от неизвестного параметра и. Справедлива теорема о наследовании сходимости [7, §2.4], согласно которой при подстановке в $h(и)$ оценки $и_n$ вместо и соотношение (8) остается справедливым, т.е.

$$\lim_{n\to\infty} P\left\{\theta_n - u_\gamma \frac{\sqrt{h(\theta_n)}}{\sqrt{n}} \le \theta\right\} = \gamma.$$

Следовательно, в качестве приближенной нижней доверительной границы следует взять

$$\theta_H = \theta_n - u_\gamma \frac{\sqrt{h(\theta_n)}}{\sqrt{n}},$$

а в качестве приближенной верхней доверительной границы -

$$\theta_B = \theta_n + u_\gamma \frac{\sqrt{h(\theta_n)}}{\sqrt{n}}.$$

С ростом объема выборки качество приближенных доверительных границ улучшается, т.к. вероятности событий $\{и \ge и_H\}$ и $\{и \le и_B\}$ стремятся к г. Для построения двусторонних доверительных границ поступают аналогично правилу, указанному выше в примере 10 для

интервального оценивания параметра *m* нормального распределения. А именно, используют односторонние доверительные границы, соответствующие доверительной вероятности (1+г)/2.

При обработке экономических, управленческих или технических статистических данных обычно используют значение доверительной вероятности г = 0,95. Применяют также значения г = 0,99 или г = 0,90. Иногда встречаются значения г = 0,80, г = 0,975, г = 0,98 и др.

Для дискретных распределений, таких, как биномиальное, гипергеометрическое или распределение Пуассона (а также распределения статистики Колмогорова

$$D_n = \sqrt{n} \sup_x | F_n(x) - F_0(x) |$$

и других непараметрических статистик), функции распределения имеют скачки. Поэтому для заданного заранее значения г, например, г= 0,95, нельзя указать доверительные границы, поскольку уравнения, с помощью которых вводятся доверительные границы, не имеют ни одного решения. Так, рассмотрим биномиальное распределение

$$P(Y = y \mid p, n) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0,1,2,...,n,$$

где *Y* – число осуществлений события, *n* – объем выборки. Для него нельзя указать статистику *K(Y, n)* такую, что

$$P\{p \le K(Y, n)\} = г,$$

поскольку *K(Y, n)* – функция от *Y* и может принимать не больше значений, чем принимает *Y*, т.е. *n* + 1, а для г имеется бесконечно много возможных значений – столько, сколько точек на отрезке. Сказанная означает, что верхней доверительной границы в случае биномиального распределения не существует.

Для дискретных распределений приходится изменить определения доверительных границ. Покажем изменения на примере биномиального распределения. Так, в качестве верхней доверительной границы $и_В$ используют наименьшее *K(Y, n)* такое, что

$$P\{p \le K(Y, n)\} \ge г.$$

Аналогичным образом поступают для других доверительных границ и других распределений. Необходимо иметь в виду, что при небольших *n* и *p* истинная доверительная вероятность $P\{p \le K(Y, n)\}$ может существенно отличаться от номинальной г, как это подробно продемонстрировано в работе [13]. Поэтому наряду с величинами типа *K(Y, n)* (т.е. доверительных границ) при разработке таблиц и компьютерных программ необходимо предусматривать возможность получения и величин типа $P\{p \le K(Y, n)\}$ (т.е. достигаемых доверительных вероятностей).

**Основные понятия, используемые при проверке гипотез.** Статистическая гипотеза – любое предположение, касающееся неизвестного распределения случайных величин (элементов). Приведем формулировки нескольких статистических гипотез:

1. Результаты наблюдений имеют нормальное распределение с нулевым математическим ожиданием.

2. Результаты наблюдений имеют функцию распределения *N*(0,1).

3. Результаты наблюдений имеют нормальное распределение.

4. Результаты наблюдений в двух независимых выборках имеют одно и то же нормальное распределение.

5. Результаты наблюдений в двух независимых выборках имеют одно и то же распределение.

Различают нулевую и альтернативную гипотезы. Нулевая гипотеза – гипотеза, подлежащая проверке. Альтернативная гипотеза – каждая допустимая гипотеза, отличная от нулевой. Нулевую гипотезу обозначают $H_0$, альтернативную – $H_1$ (от Hypothesis – «гипотеза» (англ.)).

Выбор тех или иных нулевых или альтернативных гипотез определяется стоящими перед менеджером, экономистом, инженером, исследователем прикладными задачами. Рассмотрим примеры.

*Пример 11.* Пусть нулевая гипотеза – гипотеза 2 из приведенного выше списка, а альтернативная – гипотеза 1. Сказанное означает, то реальная ситуация описывается вероятностной моделью, согласно которой результаты наблюдений рассматриваются как реализации независимых одинаково распределенных случайных величин с функцией

распределения $N(0,\text{у})$, где параметр у неизвестен статистику. В рамках этой модели нулевую гипотезу записывают так:

$$H_0: \text{у} = 1,$$

а альтернативную так:

$$H_1: \text{у} \neq 1.$$

*Пример 12.* Пусть нулевая гипотеза – по-прежнему гипотеза 2 из приведенного выше списка, а альтернативная – гипотеза 3 из того же списка. Тогда в вероятностной модели управленческой, экономической или производственной ситуации предполагается, что результаты наблюдений образуют выборку из нормального распределения $N(m, \text{у})$ при некоторых значениях $m$ и у. Гипотезы записываются так:

$$H_0: m = 0, \text{у} = 1$$

(оба параметра принимают фиксированные значения);

$$H_1: m \neq 0 \text{ и/или } \text{у} \neq 1$$

(т.е. либо $m \neq 0$, либо $\text{у} \neq 1$, либо и $m \neq 0$, и $\text{у} \neq 1$).

*Пример 13.* Пусть $H_0$ – гипотеза 1 из приведенного выше списка, а $H_1$ – гипотеза 3 из того же списка. Тогда вероятностная модель – та же, что в примере 12,

$$H_0: m = 0, \text{у произвольно};$$

$$H_1: m \neq 0, \text{у произвольно}.$$

*Пример 14.* Пусть $H_0$ – гипотеза 2 из приведенного выше списка, а согласно $H_1$ результаты наблюдений имеют функцию распределения $F(x)$, не совпадающую с функцией стандартного нормального распределения $\Phi(x)$. Тогда

$$H_0: F(x) = \Phi(x) \text{ при всех } x \text{ (записывается как } F(x) \equiv \Phi(x));$$

$$H_1: F(x_0) \neq \Phi(x_0) \text{ при некотором } x_0 \text{ (т.е. неверно, что } F(x) \equiv \Phi(x)).$$

*Примечание.* Здесь $\equiv$ - знак тождественного совпадения функций (т.е. совпадения при всех возможных значениях аргумента $x$).

*Пример 15.* Пусть $H_0$ – гипотеза 3 из приведенного выше списка, а согласно $H_1$ результаты наблюдений имеют функцию распределения $F(x)$, не являющуюся нормальной. Тогда

$$H_0 : F(x) \equiv \Phi\left(\frac{x-m}{\sigma}\right) \text{ при некоторых } m, \text{у};$$

$H_1$: для любых $m$, у найдется $x_0 = x_0(m, \text{у})$ такое, что $F(x_0) \neq \Phi\left(\frac{x_0 - m}{\sigma}\right)$.

*Пример 16.* Пусть $H_0$ – гипотеза 4 из приведенного выше списка, согласно вероятностной модели две выборки извлечены из совокупностей с функциями распределения $F(x)$ и $G(x)$, являющихся нормальными с параметрами $m_1$, $\text{у}_1$ и $m_2$, $\text{у}_2$ соответственно, а $H_1$ – отрицание $H_0$. Тогда

$$H_0: m_1 = m_2, \text{у}_1 = \text{у}_2, \text{ причем } m_1 \text{и } \text{у}_1 \text{ произвольны};$$

$$H_1: m_1 \neq m_2 \text{ и/или } \text{у}_1 \neq \text{у}_2.$$

*Пример 17.* Пусть в условиях примера 16 дополнительно известно, что $\text{у}_1 = \text{у}_2$. Тогда

$$H_0: m_1 = m_2, \text{у} > 0, \text{ причем } m_1 \text{и у произвольны};$$

$$H_1: m_1 \neq m_2, \text{у} > 0.$$

*Пример 18.* Пусть $H_0$ – гипотеза 5 из приведенного выше списка, согласно вероятностной модели две выборки извлечены из совокупностей с функциями распределения $F(x)$ и $G(x)$ соответственно, а $H_1$ – отрицание $H_0$. Тогда

$$H_0: F(x) \equiv G(x), \text{ где } F(x) \text{ – произвольная функция распределения};$$

$$H_1: F(x) \text{ и } G(x) \text{ - произвольные функции распределения, причем}$$

$$F(x) \neq G(x) \text{ при некоторых } x.$$

*Пример 19.* Пусть в условиях примера 17 дополнительно предполагается, что функции распределения $F(x)$ и $G(x)$ отличаются только сдвигом, т.е. $G(x) = F(x - a)$ при некотором $a$. Тогда

$$H_0: F(x) \equiv G(x), \text{ где } F(x) \text{ – произвольная функция распределения};$$

$$H_1: G(x) = F(x - a),\ a \neq 0, \text{ где } F(x) \text{ – произвольная функция распределения.}$$

*Пример 20.* Пусть в условиях примера 14 дополнительно известно, что согласно вероятностной модели ситуации $F(x)$ - функция нормального распределения с единичной дисперсией, т.е. имеет вид $N(m, 1)$. Тогда

$$H_0: m = 0 \text{ (т.е. } F(x) = \Phi(x) \text{ при всех } x \text{ );(записывается как } F(x) \equiv \Phi(x)\text{);}$$

$$H_1: m \neq 0 \text{ (т.е. неверно, что } F(x) \equiv \Phi(x)\text{).}$$

*Пример 21.* При статистическом регулировании технологических, экономических, управленческих или иных процессов [2] рассматривают выборку, извлеченную из совокупности с нормальным распределением и известной дисперсией, и гипотезы

$$H_0: m = m_0,$$

$$H_1: m = m_1,$$

где значение параметра $m = m_0$ соответствует налаженному ходу процесса, а переход к $m = m_1$ свидетельствует о разладке.

*Пример 22.* При статистическом приемочном контроле [2] число дефектных единиц продукции в выборке подчиняется гипергеометрическому распределению, неизвестным параметром является $p = D/N$ – уровень дефектности, где $N$ – объем партии продукции, $D$ – общее число дефектных единиц продукции в партии. Используемые в нормативно-технической и коммерческой документации (стандартах, договорах на поставку и др.) планы контроля часто нацелены на проверку гипотезы

$$H_0: p \leq AQL$$

против альтернативной гипотезы

$$H_1: p \geq LQ,$$

где $AQL$ – приемочный уровень дефектности, $LQ$ – браковочный уровень дефектности (очевидно, что $AQL < LQ$).

*Пример 23.* В качестве показателей стабильности технологического, экономического, управленческого или иного процесса используют ряд характеристик распределений контролируемых показателей, в частности, коэффициент вариации $v = y/M(X)$. Требуется проверить нулевую гипотезу

$$H_0: v \leq v_0$$

при альтернативной гипотезе

$$H_1: v > v_0,$$

где $v_0$ – некоторое заранее заданное граничное значение.

*Пример 24.* Пусть вероятностная модель двух выборок – та же, что в примере 18, математические ожидания результатов наблюдений в первой и второй выборках обозначим $M(X)$ и $M(У)$ соответственно. В ряде ситуаций проверяют нулевую гипотезу

$$H_0: M(X) = M(У)$$

против альтернативной гипотезы

$$H_1: M(X) \neq M(У).$$

*Пример 25*. Выше отмечалось большое значение в математической статистике функций распределения, симметричных относительно 0, При проверке симметричности

$$H_0: F(-x) = 1 - F(x) \text{ при всех } x, \text{ в остальном } F \text{ произвольна;}$$

$$H_1: F(-x_0) \neq 1 - F(x_0) \text{ при некотором } x_0, \text{ в остальном } F \text{ произвольна.}$$

В вероятностно-статистических методах принятия решений используются и многие другие постановки задач проверки статистических гипотез. Некоторые из них рассматриваются ниже.

Конкретная задача проверки статистической гипотезы полностью описана, если заданы нулевая и альтернативная гипотезы. Выбор метода проверки статистической гипотезы, свойства и характеристики методов определяются как нулевой, так и альтернативной гипотезами. Для проверки одной и той же нулевой гипотезы при различных альтернативных гипотезах следует использовать, вообще говоря, различные методы. Так, в примерах 14 и 20 нулевая гипотеза одна и та же, а альтернативные – различны. Поэтому в условиях примера 14 следует применять методы, основанные на критериях согласия с параметрическим семейством (типа Колмогорова или типа омега-квадрат), а в условиях примера 20 – методы на основе критерия Стьюдента или критерия Крамера-Уэлча [2,11]. Если в условиях примера 14 использовать критерий Стьюдента, то он не будет решать поставленных задач. Если в условиях примера 20 использовать критерий

согласия типа Колмогорова, то он, напротив, будет решать поставленные задачи, хотя, возможно, и хуже, чем специально приспособленный для этого случая критерий Стьюдента.

При обработке реальных данных большое значение имеет правильный выбор гипотез $H_0$ и $H_1$. Принимаемые предположения, например, нормальность распределения, должны быть тщательно обоснованы, в частности, статистическими методами. Отметим, что в подавляющем большинстве конкретных прикладных постановок распределение результатов наблюдений отлично от нормального [2].

Часто возникает ситуация, когда вид нулевой гипотезы вытекает из постановки прикладной задачи, а вид альтернативной гипотезы не ясен. В таких случаях следует рассматривать альтернативную гипотезу наиболее общего вида и использовать методы, решающие поставленную задачу при всех возможных $H_1$. В частности при проверке гипотезы 2 (из приведенного выше списка) как нулевой следует в качестве альтернативной гипотезы использовать $H_1$ из примера 14, а не из примера 20, если нет специальных обоснований нормальности распределения результатов наблюдений при альтернативной гипотезе.

Статистические гипотезы бывают параметрические и непараметрические. Предположение, которое касается неизвестного значения параметра распределения, входящего в некоторое параметрическое семейство распределений, называется параметрической гипотезой (напомним, что параметр может быть и многомерным). Предположение, при котором вид распределения неизвестен (т.е. не предполагается, что оно входит в некоторое параметрическое семейство распределений), называется непараметрической гипотезой. Таким образом, если распределение $F(x)$ результатов наблюдений в выборке согласно принятой вероятностной модели входит в некоторое параметрическое семейство $\{F(x;и), и \in И\}$, т.е. $F(x) = F(x;и_0)$ при некотором $и_0 \in И$, то рассматриваемая гипотеза – параметрическая, в противном случае – непараметрическая.

Если и $H_0$ и $H_1$ – параметрические гипотезы, то задача проверки статистической гипотезы – параметрическая. Если хотя бы одна из гипотез $H_0$ и $H_1$ – непараметрическая, то задача проверки статистической гипотезы – непараметрическая. Другими словами, если вероятностная модель ситуации – параметрическая, т.е. полностью описывается в терминах того или иного параметрического семейства распределений вероятностей, то и задача проверки статистической гипотезы – параметрическая. Если же вероятностная модель ситуации – непараметрическая, т.е. ее нельзя полностью описать в терминах какого-либо параметрического семейства распределений вероятностей, то и задача проверки статистической гипотезы – непараметрическая. В примерах 11-13, 16, 17, 20-22 даны постановки параметрических задач проверки гипотез, а в примерах 14, 15, 18, 19, 23-25 – непараметрических. Непараметрические задачи делятся на два класса: в одном из них речь идет о проверке утверждений, касающихся функций распределения (примеры 14, 15, 18, 19, 25), во втором – о проверке утверждений, касающихся характеристик распределений (примеры 23, 24).

Статистическая гипотеза называется простой, если она однозначно задает распределение результатов наблюдений, вошедших в выборку. В противном случае статистическая гипотеза называется сложной. Гипотеза 2 из приведенного выше списка, нулевые гипотезы в примерах 11, 12, 14, 20, нулевая и альтернативная гипотезы в примере 21 – простые, все остальные упомянутые выше гипотезы – сложные.

Однозначно определенный способ проверки статистических гипотез называется статистическим критерием. Статистический критерий строится с помощью статистики $U(x_1, x_2, \ldots, x_n)$ – функции от результатов наблюдений $x_1, x_2, \ldots, x_n$. В пространстве значений статистики $U$ выделяют критическую область Ш, т.е. область со следующим свойством: если значения применяемой статистики принадлежат данной области, то отклоняют (иногда говорят - отвергают) нулевую гипотезу, в противном случае – не отвергают (т.е. принимают).

Статистику $U$, используемую при построении определенного статистического критерия, называют статистикой этого критерия. Например, в задаче проверки статистической гипотезы, приведенной в примере 14, применяют критерий Колмогорова, основанный на статистике

$$D_n = \sqrt{n} \sup_x | F_n(x) - F_0(x) | .$$

При этом $D_n$ называют статистикой критерия Колмогорова.

Частным случаем статистики $U$ является векторзначная функция результатов наблюдений $U_0(x_1, x_2, \ldots, x_n) = (x_1, x_2, \ldots, x_n)$, значения которой – набор результатов наблюдений. Если $x_i$ – числа, то $U_0$ – набор $n$ чисел, т.е. точка $n$–мерного пространства. Ясно, что статистика критерия $U$ является функцией от $U_0$, т.е. $U = f(U_0)$. Поэтому можно считать, что Ш – область в том же $n$–мерном пространстве, нулевая гипотеза отвергается, если $(x_1, x_2, \ldots, x_n) \in$ Ш, и принимается в противном случае.

В вероятностно-статистических методах принятия решений, статистические критерии, как правило, основаны на статистиках $U$, принимающих числовые значения, и критические области имеют вид

$$\text{Ш} = \{U(x_1, x_2, \ldots, x_n) > C\}, \quad (9)$$

где $C$ – некоторые числа.

Статистические критерии делятся на параметрические и непараметрические. Параметрические критерии используются в параметрических задачах проверки статистических гипотез, а непараметрические – в непараметрических задачах.

При проверке статистической гипотезы возможны ошибки. Есть два рода ошибок. Ошибка первого рода заключается в том, что отвергают нулевую гипотезу, в то время как в действительности эта гипотеза верна. Ошибка второго рода состоит в том, что принимают нулевую гипотезу, в то время как в действительности эта гипотеза неверна.

Вероятность ошибки первого рода называется уровнем значимости и обозначается б. Таким образом, б = $P\{U \in \text{Ш} \mid H_0\}$, т.е. уровень значимости б – это вероятность события $\{U \in \text{Ш}\}$, вычисленная в предположении, что верна нулевая гипотеза $H_0$.

Уровень значимости однозначно определен, если $H_0$ – простая гипотеза. Если же $H_0$ – сложная гипотеза, то уровень значимости, вообще говоря, зависит от функции распределения результатов наблюдений, удовлетворяющей $H_0$. Статистику критерия $U$ обычно строят так, чтобы вероятность события $\{U \in \text{Ш}\}$ не зависела от того, какое именно распределение (из удовлетворяющих нулевой гипотезе $H_0$) имеют результаты наблюдений. Для статистик критерия $U$ общего вида под уровнем значимости понимают максимально возможную ошибку первого рода. Максимум (точнее, супремум) берется по всем возможным распределениям, удовлетворяющим нулевой гипотезе $H_0$, т.е. б = $\sup P\{U \in \text{Ш} \mid H_0\}$.

Если критическая область имеет вид, указанный в формуле (9), то

$$P\{U > C \mid \text{Н}_0\} = \text{б}. \quad (10)$$

Если $C$ задано, то из последнего соотношения определяют б. Часто поступают по иному - задавая б (обычно б = 0,05, иногда б = 0,01 или б = 0,1, другие значения б используются гораздо реже), определяют $C$ из уравнения (10), обозначая его $C_б$, и используют критическую область Ш = $\{U > C_б\}$ с заданным уровнем значимости б.

Вероятность ошибки второго рода есть $P\{U \notin \text{Ш} \mid H_1\}$. Обычно используют не эту вероятность, а ее дополнение до 1, т.е. $P\{U \in \text{Ш} \mid H_1\} = 1 - P\{U \notin \text{Ш} \mid H_1\}$. Эта величина носит название мощности критерия. Итак, мощность критерия – это вероятность того, что нулевая гипотеза будет отвергнута, когда альтернативная гипотеза верна.

Понятия уровня значимости и мощности критерия объединяются в понятии функции мощности критерия – функции, определяющей вероятность того, что нулевая гипотеза будет отвергнута. Функция мощности зависит от критической области Ш и действительного распределения результатов наблюдений. В параметрической задаче проверки гипотез распределение результатов наблюдений задается параметром и. В этом случае функция мощности обозначается $M$(Ш,и) и зависит от критической области Ш и действительного значения исследуемого параметра и. Если

$$H_0: \text{и} = \text{и}_0,$$
$$H_1: \text{и} = \text{и}_1,$$

то

$$M(\text{Ш},\text{и}_0) = \text{б},$$
$$M(\text{Ш},\text{и}_1) = 1 - \text{в},$$

где б – вероятность ошибки первого рода, в - вероятность ошибки второго рода. В статистическом приемочном контроле б – риск изготовителя, в – риск потребителя. При статистическом регулировании технологического процесса б – риск излишней наладки, в – риск незамеченной разладки.

Функция мощности $M$(Ш,и) в случае одномерного параметра и обычно достигает минимума, равного б, при и = $и_0$, монотонно возрастает при удалении от $и_0$ и приближается к 1 при | и - $и_0$ | → ∞.

В ряде вероятностно-статистических методов принятия решений используется оперативная характеристика $L$(Ш,и) - вероятность принятия нулевой гипотезы в зависимости от критической области Ш и действительного значения исследуемого параметра и. Ясно, что

$$L(Ш,и) = 1 - M(Ш,и).$$

Основной характеристикой статистического критерия является функция мощности. Для многих задач проверки статистических гипотез разработан не один статистический критерий, а целый ряд. Чтобы выбрать из них определенный критерий для использования в конкретной практической ситуации, проводят сравнение критериев по различным показателям качества [2, приложение 3], прежде всего с помощью их функций мощности. В качестве примера рассмотрим лишь два показателя качества критерия проверки статистической гипотезы – состоятельность и несмещенность.

Пусть объем выборки $n$ растет, а $U_n$ и $Ш_n$ – статистики критерия и критические области соответственно. Критерий называется состоятельным, если

$$\lim_{n\to\infty} P\{U_n \in \Psi_n \mid H_1\} = 1,$$

т.е. вероятность отвергнуть нулевую гипотезу стремится к 1, если верна альтернативная гипотеза.

Статистический критерий называется несмещенным, если для любого $и_0$, удовлетворяющего $H_0$, и любого $и_1$ , удовлетворяющего $H_1$, справедливо неравенство

$$Р\{U \in Ш \mid и_0\} < Р\{U \in Ш \mid и_1\},$$

т.е. при справедливости $H_0$ вероятность отвергнуть $H_0$ меньше, чем при справедливости $H_1$.

При наличии нескольких статистических критериев в одной и той же задаче проверки статистических гипотез следует использовать состоятельные и несмещенные критерии.

### 1.2.6. Некоторые типовые задачи прикладной статистики и методы их решения

**Статистические данные и прикладная статистика.** Под прикладной статистикой понимают часть математической статистики, посвященную методам обработки реальных статистических данных, а также соответствующее математическое и программное обеспечение. Таким образом, чисто математические задачи не включают в прикладную статистику.

Под статистическими данными понимают числовые или нечисловые значения контролируемых параметров (признаков) исследуемых объектов, которые получены в результате наблюдений (измерений, анализов, испытаний, опытов и т.д.) определенного числа признаков, у каждой единицы, вошедшей в исследование. Способы получения статистических данных и объемы выборок устанавливают, исходя из постановок конкретной прикладной задачи на основе методов математической теории планирования эксперимента.

Результат наблюдения $x_i$ исследуемого признака $X$ (или совокупности исследуемых признаков $X$) у $i$ – ой единицы выборки отражает количественные и/или качественные свойства обследованной единицы с номером $i$ (здесь $i$ = 1, 2, … , $n$, где $n$ – объем выборки). Деление прикладной статистики на направления соответственно виду обрабатываемых результатов наблюдений (т.е. на статистику случайных величин, многомерный статистический анализ, статистику временных рядов и статистику объектов нечисловой природы) обсуждалось выше.

Результаты наблюдений $x_1, x_2, \ldots, x_n$, где $x_i$ – результат наблюдения $i$ – ой единицы выборки, или результаты наблюдений для нескольких выборок, обрабатывают с помощью методов прикладной статистики, соответствующих поставленной задаче. Используют, как правило, аналитические методы, т.е. методы, основанные на численных расчетах (объекты нечисловой природы при этом описывают с помощью чисел). В отдельных случаях допустимо применение графических методов (визуального анализа).

Количество разработанных к настоящему времени методов обработки данных весьма велико. Они описаны в сотнях тысяч книг и статей, а также в стандартах и других нормативно-технических и инструктивно-методических документах.

Многие методы прикладной статистики требуют проведения трудоемких расчетов, поэтому для их реализации необходимо использовать компьютеры. Программы расчетов на ЭВМ должны соответствовать современному научному уровню. Однако для единичных расчетов при отсутствии соответствующего программного обеспечения успешно используют микрокалькуляторы.

**Задачи статистического анализа точности и стабильности технологических процессов и качества продукции.** Статистические методы используют, в частности, для анализа точности и стабильности технологических процессов и качества продукции. Цель - подготовка решений, обеспечивающих эффективное функционирование технологических единиц и повышение качества и конкурентоспособности выпускаемой продукции. Статистические методы следует применять во всех случаях, когда по результатам ограниченного числа наблюдений требуется установить причины улучшения или ухудшения точности и стабильности технологического оборудования. Под точностью технологического процесса понимают свойство технологического процесса, обусловливающее близость действительных и номинальных значений параметров производимой продукции. Под стабильностью технологического процесса понимают свойство технологического процесса, обусловливающее постоянство распределений вероятностей для его параметров в течение некоторого интервала времени без вмешательства извне.

Целями применения статистических методов анализа точности и стабильности технологических процессов и качества продукции на стадиях разработки, производства и эксплуатации (потребления) продукции являются, в частности:

• определение фактических показателей точности и стабильности технологического процесса, оборудования или качества продукции;
• установление соответствия качества продукции требованиям нормативно-технической документации;
• проверка соблюдения технологической дисциплины;
• изучение случайных и систематических факторов, способных привести к появлению дефектов;
• выявление резервов производства и технологии;
• обоснование технических норм и допусков на продукцию;
• оценка результатов испытаний опытных образцов при обосновании требований к продукции и нормативов на нее;
• обоснование выбора технологического оборудования и средств измерений и испытаний;
• сравнение различных образцов продукции;
• обоснование замены сплошного контроля статистическим;
• выявление возможности внедрения статистических методов управления качеством продукции, и т.д.

Для достижения перечисленных выше целей применяют различные методы описания данных, оценивания и проверки гипотез. Приведем примеры постановок задач.

**Задачи одномерной статистики (статистики случайных величин).** Сравнение математических ожиданий проводят в тех случаях, когда необходимо установить соответствие показателей качества изготовленной продукции и эталонного образца. Это – задача проверки гипотезы:

$$H_0: M(\mathrm{X}) = m_0,$$

где $m_0$ – значение соответствующее эталонному образцу; $X$ – случайная величина, моделирующая результаты наблюдений. В зависимости от формулировки вероятностной модели ситуации и альтернативной гипотезы сравнение математических ожиданий проводят либо параметрическими, либо непараметрическими методами.

Сравнение дисперсий проводят тогда, когда требуется установить отличие рассеивания показателя качества от номинального. Для этого проверяют гипотезу:

$$H_0 : D(X) = \sigma_0^2.$$

Ряд иных постановок задач одномерной статистики приведен ниже. Не меньшее значение, чем задачи проверки гипотез, имеют задачи оценивания параметров. Они, как и задачи проверки гипотез, в зависимости от используемой вероятностной модели ситуации делятся на параметрические и непараметрические.

В параметрических задачах оценивания принимают вероятностную модель, согласно которой результаты наблюдений $x_1, x_2, \ldots, x_n$ рассматривают как реализации $n$ независимых случайных величин с функцией распределения $F(x;и)$. Здесь и – неизвестный параметр, лежащий в пространстве параметров И заданном используемой вероятностной моделью. Задача оценивания состоит в определении точечной оценок и доверительных границ (либо доверительной области) для параметра и.

Параметр и – либо число, либо вектор фиксированной конечной размерности. Так, для нормального распределения и = ($m$, $y^2$) – двумерный вектор, для биномиального и = $p$ – число, для гамма-распределения и = ($a, b, c$) – трехмерный вектор, и т.д.

В современной математической статистике разработан ряд общих методов определения оценок и доверительных границ – метод моментов, метод максимального правдоподобия, метод одношаговых оценок, метод устойчивых (робастных) оценок, метод несмещенных оценок и др. Кратко рассмотрим первые три из них. Теоретические основы различных методов оценивания и полученные с их помощью конкретные правила определения оценок и доверительных границ для тех или иных параметрических семейств распределений рассмотрены в специальной литературе, включены в нормативно-техническую и инструктивно-методическую документацию.

Метод моментов основан на использовании выражений для моментов рассматриваемых случайных величин через параметры их функций распределения. Оценки метода моментов получают, подставляя выборочные моменты вместо теоретических в функции, выражающие параметры через моменты.

В методе максимального правдоподобия, разработанном в основном Р.А.Фишером, в качестве оценки параметра и берут значение и*, для которого максимальна так называемая функция правдоподобия

$$f(x_1, и)\, f(x_2, и)\, \ldots f(x_n, и),$$

где $x_1, x_2, \ldots, x_n$ - результаты наблюдений; $f(x, и)$ – их плотность распределения, зависящая от параметра и, который необходимо оценить.

Оценки максимального правдоподобия, как правило, эффективны (или асимптотически эффективны) и имеют меньшую дисперсию, чем оценки метода моментов. В отдельных случаях формулы для них выписываются явно (нормальное распределение, экспоненциальное распределение без сдвига). Однако чаще для их нахождения необходимо численно решать систему трансцендентных уравнений (распределения Вейбулла-Гнеденко, гамма). В подобных случаях целесообразно использовать не оценки максимального правдоподобия, а другие виды оценок, прежде всего одношаговые оценки. В литературе их иногда не вполне точно называют «приближенные оценки максимального правдоподобия». При достаточно больших объемах выборок они имеют столь же хорошие свойства, как и оценки максимального правдоподобия. Поэтому их следует рассматривать не как «приближенные», а как оценки, полученные по *другому* методу, не менее обоснованному и эффективному, чем метод максимального правдоподобия. Одношаговые оценки вычисляют по явным формулам (см. главу 2.2, а также [14]).

В непараметрических задачах оценивания принимают вероятностную модель, в которой результаты наблюдений $x_1, x_2, \ldots, x_n$ рассматривают как реализации $n$ независимых случайных величин с функцией распределения $F(x)$ общего вида. От $F(x)$ требуют лишь выполнения некоторых условий типа непрерывности, существования математического ожидания и дисперсии и т.п. Подобные условия не являются столь жесткими, как условие принадлежности к определенному параметрическому семейству.

В непараметрической постановке оценивают либо характеристики случайной величины (математическое ожидание, дисперсию, коэффициент вариации), либо ее функцию распределения, плотность и т.п. Так, в силу закона больших чисел выборочное среднее арифметическое $\bar{x}$ является состоятельной оценкой математического ожидания $M(X)$ (при любой функции распределения $F(x)$ результатов наблюдений, для которой математическое ожидание существует). С помощью центральной предельной теоремы определяют асимптотические доверительные границы

$$(M(X))_{\text{Н}} = \bar{x} - u\left(\frac{1+\gamma}{2}\right)\frac{s}{\sqrt{n}},\ (M(\text{X}))_{\text{В}} = \bar{x} + u\left(\frac{1+\gamma}{2}\right)\frac{s}{\sqrt{n}}.$$

где г – доверительная вероятность, $u\left(\frac{1+\gamma}{2}\right)$ - квантиль порядка $\frac{1+\gamma}{2}$ стандартного нормального распределения N(0;1) с нулевым математическим ожиданием и единичной дисперсией, $\bar{x}$ - выборочное среднее арифметическое, $s$ – выборочное среднее квадратическое отклонение. Термин «асимптотические доверительные границы» означает, что вероятности

$$P\{(M(X))_H < M(X)\},\ P\{(M(X))_B > M(X)\},$$
$$P\{(M(X))_H < M(X) < (M(X))_B\}$$

стремятся к $\frac{1+\gamma}{2}$, $\frac{1+\gamma}{2}$ и г соответственно при $n \to \infty$, но, вообще говоря, не равны этим значениям при конечных $n$. Практически асимптотические доверительные границы дают достаточную точность при $n$ порядка 10.

Второй пример непараметрического оценивания – оценивание функции распределения. По теореме Гливенко эмпирическая функция распределения $F_n(x)$ является состоятельной оценкой функции распределения $F(x)$. Если $F(x)$ – непрерывная функция, то на основе теоремы Колмогорова доверительные границы для функции распределения $F(x)$ задают в виде

$$(F(x))_{\text{Н}} = \max\left\{0, F_n(x) - \frac{k(\gamma,n)}{\sqrt{n}}\right\},\ (F(x))_{\text{В}} = \min\left\{1, F_n(x) + \frac{k(\gamma,n)}{\sqrt{n}}\right\},$$

где $k(г,n)$ – квантиль порядка г распределения статистики Колмогорова при объеме выборки $n$ (напомним, что распределение этой статистики не зависит от $F(x)$).

Правила определения оценок и доверительных границ в параметрическом случае строятся на основе параметрического семейства распределений $F(x;и)$. При обработке реальных данных возникает вопрос – соответствуют ли эти данные принятой вероятностной модели? Т.е. статистической гипотезе о том, что результаты наблюдений имеют функцию распределения из семейства $\{F(x;и), и \in И\}$ при некотором и = $и_0$? Такие гипотезы называют гипотезами согласия, а критерии их проверки – критериями согласия.

Если истинное значение параметра и = $и_0$ известно, функция распределения $F(x;и_0)$ непрерывна, то для проверки гипотезы согласия часто применяют критерий Колмогорова, основанный на статистике

$$D_n = \sqrt{n}\sup_x |F_n(x) - F(x,\theta_0|,$$

где $F_n(x)$ – эмпирическая функция распределения.

Если истинное значение параметра $и_0$ неизвестно, например, при проверке гипотезы о нормальности распределения результатов наблюдения (т.е. при проверке принадлежности этого распределения к семейству нормальных распределений), то иногда используют статистику

$$D_n(\theta^*) = \sqrt{n}\sup_x |F_n(x) - F(x,\theta^*)|.$$

Она отличается от статистики Колмогорова $D_n$ тем, что вместо истинного значения параметра $и_0$ подставлена его оценка и*.

Распределение статистики $D_n$(и*) сильно отличается от распределения статистики $D_n$. В качестве примера рассмотрим проверку нормальности, когда и = ($m$, $у^2$), а и* = ($\bar{x}$ , $s^2$). Для этого случая квантили распределений статистик $D_n$ и $D_n$(и*) приведены в табл.1 (см., например, [15]). Таким образом, квантили отличаются примерно в 1,5 раза.

Таблица 1.

Квантили статистик $D_n$ и $D_n$(и*) при проверке нормальности

| $p$ | 0,85 | 0,90 | 0,95 | 0,975 | 0,99 |
|---|---|---|---|---|---|
| Квантили порядка $p$ для $D_n$ | 1,138 | 1,224 | 1,358 | 1,480 | 1,626 |
| Квантили порядка $p$ для $D_n$(и*) | 0,775 | 0,819 | 0,895 | 0,955 | 1,035 |

При первичной обработке статистических данных важной задачей является исключение результатов наблюдений, полученных в результате грубых погрешностей и промахов. Например, при просмотре данных о весе (в килограммах) новорожденных детей наряду с числами 3,500, 2,750, 4,200 может встретиться число 35,00. Ясно, что это промах, и получено

ошибочное число при ошибочной записи – запятая сдвинута на один знак, в результате результат наблюдения ошибочно увеличен в 10 раз.

Статистические методы исключения резко выделяющихся результатов наблюдений основаны на предположении, что подобные результаты наблюдений имеют распределения, резко отличающиеся от изучаемых, а потому их следует исключить из выборки.

Простейшая вероятностная модель такова. При нулевой гипотезе результаты наблюдений рассматриваются как реализации независимых одинаково распределенных случайных величин $X_1, X_2, \ldots, X_n$ с функцией распределения $F(x)$. При альтернативной гипотезе $X_1, X_2, \ldots, X_{n-1}$ – такие же, как и при нулевой гипотезе, а $X_n$ соответствует грубой погрешности и имеет функцию распределения $G(x) = F(x - c)$, где $c$ велико. Тогда с вероятностью, близкой к 1 (точнее, стремящейся к 1 при росте объема выборки),

$$X_n = \max \{ X_1, X_2, \ldots, X_n\} = X_{max},$$

т.е. при описании данных в качестве возможной грубой ошибки следует рассматривать $X_{max}$.
Критическая область имеет вид

$$Ш = \{x: x \geq d\}.$$

Критическое значение $d = d(б,n)$ выбирают в зависимости от уровня значимости б и объема выборки $n$ из условия

$$P\{X_{max} \geq d \mid H_0\} = б. \quad (1)$$

Условие (1) эквивалентно при больших $n$ и малых б следующему:

$$F(d) = \sqrt[n]{1-\alpha} \approx 1 - \frac{\alpha}{n}. \quad (2)$$

Если функция распределения результатов наблюдений $F(x)$ известна, то критическое значение $d$ находят из соотношения (2). Если $F(x)$ известна с точностью до параметров, например, известно, что $F(x)$ – нормальная функция распределения, то также разработаны правила проверки рассматриваемой гипотезы [8].

Однако часто вид функции распределения результатов наблюдений известен не абсолютно точно и не с точностью до параметров, а лишь с некоторой погрешностью. Тогда соотношение (2) становится практически бесполезным, поскольку малая погрешность в определении $F(x)$, как можно показать, приводит к большой погрешности при определении критического значения $d$ из условия (2), а при фиксированном $d$ уровень значимости критерия может существенно отличаться от номинального [2].

Поэтому в ситуации, когда о $F(x)$ нет полной информации, однако известны математическое ожидание $M(X)$ и дисперсия $у^2 = D(X)$ результатов наблюдений $X_1, X_2, \ldots, X_n$, можно использовать непараметрические правила отбраковки, основанные на неравенстве Чебышёва. С помощью этого неравенства найдем критическое значение $d = d(б,n)$ такое, что

$$P\left\{\max_{1 \leq i \leq n} | X_i - M(X) | \geq d\right\} \leq \alpha.$$

Так как

$$P\left\{\max_{1 \leq i \leq n} | X_i - M(X) | < d\right\} = \left[P\{| X - M(X) | < d\}\right]^n,$$

то соотношение (3) будет выполнено, если

$$P\{| X - M(X) | \geq d\} \leq 1 - \sqrt[n]{1-\alpha} \approx \frac{\alpha}{n}. \quad (4)$$

По неравенству Чебышёва

$$P\{| X - M(X) | \geq d\} \leq \frac{\sigma^2}{d^2}, \quad (5)$$

поэтому для того, чтобы (4) было выполнено, достаточно приравнять правые части формул (4) и (5), т.е. определить $d$ из условия

$$\frac{\sigma^2}{d^2} = \frac{\alpha}{n}, \quad d = \frac{\sigma\sqrt{n}}{\sqrt{\alpha}}. \quad (6)$$

Правило отбраковки, основанное на критическом значении $d$, вычисленном по формуле (6), использует минимальную информацию о функции распределения F(x) и поэтому исключает лишь результаты наблюдений, весьма далеко отстоящие от основной массы. Другими словами,

значение $d_1$, заданное соотношением (1), обычно много меньше, чем значение $d_2$, заданное соотношением (6).

**Многомерный статистический анализ.** Перейдем к многомерному статистическому анализу. Его применяют при решении следующих задач:

• исследование зависимости между признаками;

• классификация объектов или признаков, заданных векторами;

• снижение размерности пространства признаков.

При этом результат наблюдений – вектор значений фиксированного числа количественных и иногда качественных признаков, измеренных у объекта. Напомним, что количественный признак – признак наблюдаемой единицы, который можно непосредственно выразить числом и единицей измерения. Количественный признак противопоставляется качественному - признаку наблюдаемой единицы, определяемому отнесением к одной из двух или более условных категорий (если имеется ровно две категории, то признак называется альтернативным). Статистический анализ качественных признаков – часть статистики объектов нечисловой природы. Количественные признаки делятся на признаки, измеренные в шкалах интервалов, отношений, разностей, абсолютной.

А качественные – на признаки, измеренные в шкале наименований и порядковой шкале. Методы обработки данных должны быть согласованы со шкалами, в которых измерены рассматриваемые признаки (см. раздел 2.1 о теории измерений).

Целями исследования зависимости между признаками являются доказательство наличия связи между признаками и изучение этой связи. Для доказательства наличия связи между двумя случайными величинами $X$ и $Y$ применяют корреляционный анализ. Если совместное распределение $X$ и $Y$ является нормальным, то статистические выводы основывают на выборочном коэффициенте линейной корреляции, в остальных случаях используют коэффициенты ранговой корреляции Кендалла и Спирмена, а для качественных признаков – критерий хи-квадрат.

Регрессионный анализ применяют для изучения функциональной зависимости количественного признака $Y$ от количественных признаков $x(1)$, $x(2)$, … , $x(k)$. Эту зависимость называют регрессионной или, кратко, регрессией. Простейшая вероятностная модель регрессионного анализа (в случае $k = 1$) использует в качестве исходной информации набор пар результатов наблюдений $(x_i, y_i)$, $i = 1, 2, \dots , n$, и имеет вид

$$y_i = ax_i + b + e_i, \; i = 1, 2, \dots , n,$$

где $e_i$ – ошибки наблюдений. Иногда предполагают, что $e_i$ – независимые случайные величины с одним и тем же нормальным распределением $N(0, y^2)$. Поскольку распределение ошибок наблюдения обычно отлично от нормального, то целесообразно рассматривать регрессионную модель в непараметрической постановке [2], т.е. при произвольном распределении $e_i$.

Основная задача регрессионного анализа состоит в оценке неизвестных параметров $a$ и $b$, задающих линейную зависимость $y$ от $x$. Для решения этой задачи применяют разработанный еще К.Гауссом в 1794 г. метод наименьших квадратов, т.е. находят оценки неизвестных параметров модели$a$ и $b$ из условия минимизации суммы квадратов

$$\sum_{1 \le i \le n} (y_i - ax_i - b)^2$$

по переменным $a$ и $b$.

Теория регрессионного анализа описана и расчетные формулы даны в специальной литературе [2, 16, 17]. В этой теории разработаны методы точечного и интервального оценивания параметров, задающих функциональную зависимость, а также непараметрические методы оценивания этой зависимости, методы проверки различных гипотез, связанных с регрессионными зависимостями. Выбор планов эксперимента, т.е. точек $x_i$, в которых будут проводиться эксперименты по наблюдению $y_i$ – предмет теории планирования эксперимента [18].

Дисперсионный анализ применяют для изучения влияния качественных признаков на количественную переменную. Например, пусть имеются $k$ выборок результатов измерений количественного показателя качества единиц продукции, выпущенных на $k$ станках, т.е. набор чисел $(x_1(j), x_2(j), \dots , x_n(j))$, где $j$ – номер станка, $j = 1, 2, \dots, k$, а $n$ – объем выборки. В распространенной постановке дисперсионного анализа предполагают, что результаты

измерений независимы и в каждой выборке имеют нормальное распределение $N(m(j), y^2)$ с одной и той же дисперсией. Хорошо разработаны и непараметрические постановки [19].

Проверка однородности качества продукции, т.е. отсутствия влияния номера станка на качество продукции, сводится к проверке гипотезы

$$H_0: m(1) = m(2) = \ldots = m(k).$$

В дисперсионном анализе разработаны методы проверки подобных гипотез. Теория дисперсионного анализа и расчетные формулы рассмотрены в специальной литературе [20].

Гипотезу $H_0$ проверяют против альтернативной гипотезы $H_1$, согласно которой хотя бы одно из указанных равенств не выполнено. Проверка этой гипотезы основана на следующем «разложении дисперсий», указанном Р.А.Фишером:

$$(kn)s^2 = n\sum_{j=1}^{k} s^2(j) + (kn)s_1^2, \quad (7)$$

где $s^2$ – выборочная дисперсия в объединенной выборке, т.е.

$$s^2 = \frac{1}{kn}\sum_{i=1}^{n}\sum_{j=1}^{k}(x_i(j) - \bar{x})^2, \quad \bar{x} = \frac{1}{kn}\sum_{i=1}^{n}\sum_{j=1}^{k} x_i(j).$$

Далее, $s^2(j)$ – выборочная дисперсия в $j$-ой группе,

$$s^2(j) = \frac{1}{n}\sum_{i=1}^{n}(x_i(j) - \bar{x}(j))^2, \quad \bar{x}(j) = \frac{1}{n}\sum_{i=1}^{n} x_i(j), \quad j = 1,2,...,k.$$

Таким образом, первое слагаемое в правой части формулы (7) отражает внутригрупповую дисперсию. Наконец, $s_1^2$ - межгрупповая дисперсия,

$$s_1^2 = \frac{1}{k}\sum_{j=1}^{k}(\bar{x}(j) - \bar{x})^2.$$

Область прикладной статистики, связанную с разложениями дисперсии типа формулы (7), называют дисперсионным анализом. В качестве примера задачи дисперсионного анализа рассмотрим проверку приведенной выше гипотезы $H_0$ в предположении, что результаты измерений независимы и в каждой выборке имеют нормальное распределение $N(m(j), y^2)$ с одной и той же дисперсией. При справедливости $H_0$ первое слагаемое в правой части формулы (7), деленное на $y^2$, имеет распределение хи-квадрат с $k(n\text{-}1)$ степенями свободы, а второе слагаемое, деленное на $y^2$, также имеет распределение хи-квадрат, но с $(k\text{-}1)$ степенями свободы, причем первое и второе слагаемые независимы как случайные величины. Поэтому случайная величина

$$F = \frac{k(n-1)}{k-1}\frac{(kn)s_1^2}{n\sum_{j=1}^{k} s^2(j)} = \frac{k^2(n-1)s_1^2}{(k-1)\sum_{j=1}^{k} s^2(j)}$$

имеет распределение Фишера с $(k\text{-}1)$ степенями свободы числителя и $k(n\text{-}1)$ степенями свободы знаменателя. Гипотеза $H_0$ принимается, если $F \le F_{1\text{-}б}$, и отвергается в противном случае, где $F_{1\text{-}б}$ – квантиль порядка 1-б распределения Фишера с указанными числами степеней свободы. Такой выбор критической области определяется тем, что при $H_1$ величина $F$ безгранично увеличивается при росте объема выборок $n$. Значения $F_{1\text{-}б}$ берут из соответствующих таблиц [8].

Разработаны непараметрические методы решения классических задач дисперсионного анализа [19], в частности, проверки гипотезы $H_0$.

Следующий тип задач многомерного статистического анализа – задачи классификации. Они согласно [2, 20] делятся на три принципиально различных вида – дискриминантный анализ, кластер-анализ, задачи группировки.

Задача дискриминантного анализа состоит в нахождении правила отнесения наблюдаемого объекта к одному из ранее описанных классов. При этом объекты описывают в математической модели с помощью векторов, координаты которых – результаты наблюдения ряда признаков у каждого объекта. Классы описывают либо непосредственно в математических терминах, либо с помощью обучающих выборок. Обучающая выборка – это выборка, для каждого элемента которой указано, к какому классу он относится.

Рассмотрим пример применения дискриминантного анализа для принятия решений в технической диагностике. Пусть по результатам измерения ряда параметров продукции

необходимо установить наличие или отсутствие дефектов. В этом случае для элементов обучающей выборки указаны дефекты, обнаруженные в ходе дополнительного исследования, например, проведенного после определенного периода эксплуатации. Дискриминантный анализ позволяет сократить объем контроля, а также предсказать будущее поведение продукции. Дискриминантный анализ сходен с регрессионным – первый позволяет предсказывать значение качественного признака, а второй – количественного. В статистике объектов нечисловой природы разработана математическая схема, частными случаями которой являются регрессионный и дискриминантный анализы [21].

Кластерный анализ применяют, когда по статистическим данным необходимо разделить элементы выборки на группы. Причем два элемента группы из одной и той же группы должны быть «близкими» по совокупности значений измеренных у них признаков, а два элемента из разных групп должны быть «далекими» в том же смысле. В отличие от дискриминантного анализа в кластер-анализе классы не заданы, а формируются в процессе обработки статистических данных. Например, кластер-анализ может быть применен для разбиения совокупности марок стали (или марок холодильников) на группы сходных между собой.

Другой вид кластер-анализа – разбиение признаков на группы близких между собой. Показателем близости признаков может служить выборочный коэффициент корреляции. Цель кластер-анализа признаков может состоять в уменьшении числа контролируемых параметров, что позволяет существенно сократить затраты на контроль. Для этого из группы тесно связанных между собой признаков (у которых коэффициент корреляции близок к 1 – своему максимальному значению) измеряют значение одного, а значения остальных рассчитывают с помощью регрессионного анализа.

Задачи группировки решают тогда, когда классы заранее не заданы и не обязаны быть «далекими» друг от друга. Примером является группировка студентов по учебным группам. В технике решением задачи группировки часто является параметрический ряд – возможные типоразмеры группируются согласно элементам параметрического ряда. В литературе, нормативно-технических и инструктивно-методических документах по прикладной статистике также иногда используется группировка результатов наблюдений (например, при построении гистограмм).

Задачи классификации решают не только в многомерном статистическом анализе, но и тогда, когда результатами наблюдений являются числа, функции или объекты нечисловой природы. Так, многие алгоритмы кластер-анализа используют только расстояния между объектами. Поэтому их можно применять и для классификации объектов нечисловой природы, лишь бы были заданы расстояния между ними. Простейшая задача классификации такова: даны две независимые выборки, требуется определить, представляют они два класса или один. В одномерной статистике эта задача сводится к проверке гипотезы однородности [2].

Третий раздел многомерного статистического анализа – задачи снижения размерности (сжатия информации). Цель их решения состоит в определении набора производных показателей, полученных преобразованием исходных признаков, такого, что число производных показателей значительно меньше числа исходных признаков, но они содержат возможно большую часть информации, имеющейся в исходных статистических данных. Задачи снижения размерности решают с помощью методов многомерного шкалирования, главных компонент, факторного анализа и др. Например, в простейшей модели многомерного шкалирования исходные данные – попарные расстояния $\rho_{ij}, i, j = 1,2,...,k, i \neq j,$ между $k$ объектами, а цель расчетов состоит в представлении объектов точками на плоскости. Это дает возможность в буквальном смысле слова увидеть, как объекты соотносятся между собой. Для достижения этой цели необходимо каждому объекту поставить в соответствие точку на плоскости так, чтобы попарные расстояния $s_{ij}$ между точками, соответствующими объектам с номерами $i$ и $j$, возможно точнее воспроизводили расстояния $c_{ij}$ между этими объектами. Согласно основной идее метода наименьших квадратов находят точки на плоскости так, чтобы величина

$$\sum_{i=1}^{k}\sum_{j=1}^{k}(s_{ij}-\rho_{ij})^2$$

достигала своего наименьшего значения. Есть и многие другие постановки задач снижения размерности и визуализации данных.

**Статистика случайных процессов и временных рядов**. Методы статистики случайных процессов и временных рядов применяют для постановки и решения, в частности, следующих задач:

• предсказание будущего развития случайного процесса или временного ряда;

• управление случайным процессом (временным рядом) с целью достижения поставленных целей, например, заданных значений контролируемых параметров;

• построение вероятностной модели реального процесса, обычно длящегося во времени, и изучение свойств этой модели.

*Пример 1.* При внедрении статистического регулирования технологического процесса необходимо проверить, что в налаженном состоянии математическое ожидание контролируемого параметра не меняется со временем. Если подобное изменение будет обнаружено, то необходимо установить подналадочное устройство.

*Пример 2.* Следящие системы, например, входящие в состав автоматизированной системы управления технологическим процессом, должны выделять полезный сигнал на фоне шумов. Это – задача оценивания (полезного сигнала), в то время как в примере 1 речь шла о задаче проверки гипотезы.

Методы статистики случайных процессов и временных рядов описаны в литературе [2,20].

**Статистика объектов нечисловой природы**. Методы статистики объектов нечисловой природы применяют всегда, когда результаты наблюдений являются объектами нечисловой природы. Например, сообщениями о годности или дефектности единиц продукции. Информацией о сортности единиц продукции. Разбиениями единиц продукции на группы соответственно значения контролируемых параметров. Упорядочениями единиц продукции по качеству или инвестиционных проектов по предпочтительности. Фотографиями поверхности изделия, пораженной коррозией, и т.д. Итак, объекты нечисловой природы – это измерения по качественному признаку, множества, бинарные отношения (разбиения, упорядочения и др.) и многие другие математические объекты [2]. Они используются в различных вероятностно-статистических методах принятия решений. В частности, в задачах управления качеством продукции, а также, например, в медицине и социологии, как для описания результатов приборных измерений, так и для анализа экспертных оценок.

Для описания данных, являющихся объектами нечисловой природы, применяют, в частности, таблицы сопряженности, а в качестве средних величин – решения оптимизационных задач [2]. В качестве выборочных средних для измерений в порядковой шкале используют медиану и моду, а в шкале наименований – только моду. О методах классификации нечисловых данных говорилось выше.

Для решения параметрических задач оценивания используют оптимизационный подход, метод одношаговых оценок, метод максимального правдоподобия, метод устойчивых оценок. Для решения непараметрических задач оценивания наряду с оптимизационными подходами к оцениванию характеристик используют непараметрические оценки распределения случайного элемента, плотности распределения, функции, выражающей зависимость [2].

В качестве примера методов проверки статистических гипотез для объектов нечисловой природы рассмотрим критерий «хи-квадрат» (обозначают $\chi^2$), разработанный К.Пирсоном для проверки гипотезы однородности (другими словами, совпадения) распределений, соответствующих двум независимым выборкам.

Рассматриваются две выборки объемов $n_1$ и $n_2$, состоящие из результатов наблюдений качественного признака, имеющего $k$ градаций. Пусть $m_{1j}$ и $m_{2j}$ – количества элементов первой и второй выборок соответственно, для которых наблюдается $j$–я градация, а $p_{1j}$ и $p_{2j}$ – вероятности того, что эта градация будет принята, для элементов первой и второй выборок, $j = 1, 2, \ldots, k$.

Для проверки гипотезы однородности распределений, соответствующих двум независимым выборкам,

$$H_0: p_{1j} = p_{2j}, j = 1, 2, \ldots, k,$$

применяют критерий $\chi^2$ (хи-квадрат) со статистикой

$$X^2 = n_1 n_2 \sum_{j=1}^{k} \frac{1}{m_{1j} + m_{2j}} \left( \frac{m_{1j}}{n_1} - \frac{m_{2j}}{n_2} \right)^2 .$$

Установлено [9, 11], что статистика $X^2$ при больших объемах выборок $n_1$ и $n_2$ имеет асимптотическое распределение хи-квадрат с $(k-1)$ степенью свободы.

Таблица 1

Распределения плавок стали по процентному содержанию серы

| Содержание серы, в % | Число плавок | |
|---|---|---|
| | Завод А | Завод Б |
| 0,00 ч 0,02 | 82 | 63 |
| 0,02 ч 0,04 | 535 | 429 |
| 0,04 ч 0,06 | 1173 | 995 |
| 0,06 ч 0,08 | 1714 | 1307 |

*Пример 3.* В табл.1 приведены данные о содержании серы в углеродистой стали, выплавляемой двумя металлургическими заводами. Проверим, можно ли считать распределения примеси серы в плавках стали этих двух заводов одинаковыми.

Расчет по данным табл.1 дает $X^2 = 3{,}39$. Квантиль порядка 0,95 распределения хи-квадрат с $k - 1 = 3$ степенями свободы равен $\chi^2_{0,95}(3) = 7{,}8$, а потому гипотезу о совпадении функций распределения содержания серы в плавках двух заводов нельзя отклонить, т.е. ее следует принять (на уровне значимости б = 0,05).

Подробнее методы статистики объектов нечисловой природы рассмотрены в третьей части..

Выше дано лишь краткое описание содержания прикладной статистики на современном этапе. Подробное изложение конкретных методов содержится в дальнейших главах учебника и в специальной литературе.

**Некоторые постановки задач прикладной статистики, широко используемые в практической деятельности и в научных исследованиях.** Чтобы дать представление о богатом содержании теории рассматриваемых методов, приведем краткий перечень основных типов постановок задач в соответствии с описанной выше классификацией областей прикладной статистики. Основные из них рассматриваются в дальнейших главах настоящего учебника.

1. Одномерная статистика.

1.1. Описание материала
1.1.1. Расчет выборочных характеристик распределения.
1.1.2. Построение гистограмм и полигонов часто.
1.1.3. Приближение эмпирических распределений с помощью распределений из системы Пирсона и других систем…

1.2. Оценивание.

1.2.1. Параметрическое оценивание.
1.2.1.1. Правила определения оценок и доверительных границ для параметров устойчивого распределения.
1.2.1.2. Правила определения оценок и доверительных границ для параметров логистического распределения.
1.2.1.3. Правила определения оценок и доверительных границ для параметров экспоненциального распределения и смеси экспоненциальных распределений… (и так далее для различных семейств распределений).

1.2.2. Непараметрическое оценивание.
1.2.2.1. Непараметрическое точечное и доверительное оценивание основных характеристик распределения – математического ожидания, дисперсии, среднего квадратического отклонения, коэффициента вариации, квантилей, прежде всего медианы.
1.2.2.2. Непараметрические оценки плотности и функции распределения.
1.2.2.3. Непараметрическое оценивание параметра сдвига…

1.3. Проверка гипотез.

1.3.1. Параметрические задачи проверки гипотез.
1.3.1.1. Проверка равенства математических ожиданий для двух нормальных совокупностей.
1.3.1.2. Проверка равенства дисперсий для двух нормальных совокупностей.
1.3.1.3. Проверка равенства коэффициентов вариации для двух нормальных совокупностей.
1.3.1.4. Проверка равенства математических ожиданий и дисперсий для двух нормальных совокупностей.
1.3.1.5. Проверка равенства математического ожидания нормального распределения определенному значению.
1.3.1.6. Проверка равенства дисперсии нормального распределения определенному значению…
1.3.1.7. Проверка равенства параметров двух экспоненциальных совокупностей… (и так далее – проверка утверждений о параметрах для различных семейств распределений).

1.3.2. Непараметрические задачи проверки гипотез.
1.3.2.1. Непараметрическая проверка равенства математических ожиданий для двух совокупностей.
1.3.2.2. Непараметрическая проверка равенства дисперсий для двух совокупностей.
1.3.2.3. Непараметрическая проверка равенства коэффициентов вариации для двух совокупностей.
1.3.2.4. Непараметрическая проверка равенства математических ожиданий и дисперсий для двух совокупностей.
1.3.2.5. Непараметрическая проверка равенства математического ожидания определенному значению.
1.3.2.6. Непараметрическая проверка равенства дисперсии определенному значению…
1.3.2.7. Проверка гипотезы согласия с равномерным распределением по критерию Колмогорова.
1.3.2.8. Проверка гипотезы согласия с равномерным распределением по критерию омега-квадрат (Крамера-Мизеса-Смирнова).
1.3.2.9. Проверка гипотезы согласия с равномерным распределением по критерию Смирнова.
1.3.2.10. Проверка гипотезы согласия с нормальным семейством распределений по критерию типа Колмогорова при известной дисперсии.
1.3.2.11. Проверка гипотезы согласия с нормальным семейством распределений по критерию типа Колмогорова при известном математическом ожидании.
1.3.2.12. Проверка гипотезы согласия с нормальным семейством распределений по критерию типа Колмогорова (оба параметра неизвестны).
1.3.2.13. Проверка гипотезы согласия с нормальным семейством распределений по критерию типа омега-квадрат при известной дисперсии.
1.3.2.14. Проверка гипотезы согласия с нормальным семейством распределений по критерию типа омега-квадрат при известном математическом ожидании.
1.3.2.15. Проверка гипотезы согласия с нормальным семейством распределений по критерию типа омега-квадрат (оба параметра неизвестны).
1.3.2.16. Проверка гипотезы согласия с экспоненциальным семейством распределений по критерию типа омега-квадрат… ( и так далее для различных семейств распределений, тех или иных предположениях о параметрах, всевозможных критериев).
1.3.2.17. Проверка гипотезы однородности двух выборок методом Смирнова.
1.3.2.18. Проверка гипотезы однородности двух выборок методом омега-квадрат.
1.3.2.19. Проверка гипотезы однородности двух выборок с помощью критерия Вилкоксона.
1.3.2.20. Проверка гипотезы однородности двух выборок по критерию Ван-дер-Вардена.
1.3.2.21. Проверка гипотезы симметрии функции распределения относительно 0 методом Смирнова.
1.3.2.22. Проверка гипотезы симметрии функции распределения относительно 0 с помощью критерия типа омега-квадрат (Орлова).
1.3.2.23. Проверка гипотезы независимости элементов выборки.
1.3.2.24. Проверка гипотезы одинаковой распределенности элементов выборки…(и т.д.).

2. Многомерный статистический анализ.

2.1. Описание материала.
2.1.1. Расчет выборочных характеристик (вектора средних, ковариационной и корреляционной матриц и др.).
2.1.2. Таблицы сопряженности.
2.1.3. Детерминированные методы приближения функциональной зависимости.
2.1.3.1. Метод наименьших квадратов.
2.1.3.2. Метод наименьших модулей
2.1.3.3. Сплайны и др.
2.1.4. Методы снижения размерности.
2.1.4.1. Алгоритмы факторного анализа.
2.1.4.2. Алгоритмы метода главных компонент
2.1.4.3. Алгоритмы многомерного метрического шкалирования.
2.1.4.4. Алгоритмы многомерного неметрического шкалирования.
2.1.4.5. Методы оптимального проецирования и др.
2.1.5. Методы классификации.
2.1.5.1. Методы кластер-анализа – иерархические процедуры.
2.1.5.2. Методы кластер-анализа – оптимизационный подход.
2.1.5.3. Методы кластер-анализа – итерационные процедуры…
2.1.5.4. Методы группировки…

2.2. Оценивание.

2.2.1. Параметрическое оценивание.
2.2.1.1. Оценивание параметров многомерного нормального распределения.
2.2.1.2. Оценивание параметров в нормальной модели линейной регрессии.
2.2.1.3. Методы расщепления смесей.
2.2.1.4. Оценивание компонент дисперсии в дисперсионном анализе (в нормальной модели).
2.2.1.5. Оценивание размерности и структуры модели в регрессионном анализе (в нормальной модели).
2.2.1.6. Оценивание в дискриминантном анализе (в нормальной модели).
2.2.1.7. Оценивание в методах снижения размерности (в нормальной модели).
2.2.1.8. Нелинейная регрессия.
2.2.1.9. Методы планирования эксперимента.

2.2.2. Непараметрическое оценивание.
2.2.2.1. Непараметрические оценки многомерной плотности.
2.2.2.2. Непараметрическая регрессия (с погрешностями наблюдений произвольного вида).
2.2.2.3. Непараметрическая регрессия (на основе непараметрических оценок многомерной плотности).
2.2.2.4. Монотонная регрессия.
2.2.2.5. Непараметрический дискриминантный анализ.
2.2.2.6. Непараметрический дисперсионный анализ…

2.3. Проверка гипотез.

2.3.1. Параметрические задачи проверки гипотез.
2.3.1.1. Корреляционный анализ (нормальная модель).
2.3.1.2. Проверка гипотез об отличии коэффициентов при предикторах от 0 в линейной регрессии при справедливости нормальной модели.
2.3.1.3. Проверка гипотезы о равенстве математических ожиданий нормальных совокупностей (дисперсионный анализ).
2.3.1.4. Проверка гипотезы о совпадении двух линий регрессии (нормальная модель)…(и т.д.)

2.3.2. Непараметрические задачи проверки гипотез.
2.3.2.1. Непараметрический корреляционный анализ.
2.3.2.2. Проверка гипотез об отличии коэффициентов при предикторах от 0 в линейной регрессии (непараметрическая постановка).
2.3.2.3. Проверка гипотез в непараметрическом дисперсионном анализе.
2.3.2.4. Проверка гипотезы о совпадении двух линий регрессии (непараметрическая постановка)…(и т.д.)

Здесь остановимся, поскольку продолжение предполагало бы знакомство со многими достаточно сложными методами, о которых нет упоминаний в этой книге. Приведенный выше перечень ряда основных типов постановок задач, используемых в прикладной статистике, дает первоначальное представление об объеме арсенала разработанных к настоящему времени интеллектуальных инструментов в рассматриваемой области.

## Литература

1. Вероятность и математическая статистика: Энциклопедия/Гл. ред. Ю.В.Прохоров. – М.: Большая Российская энциклопедия, 1999. – 910с.
2. Орлов А.И. Эконометрика. – М.: Экзамен, 2002. - 576 с.
3. Рекомендации. Прикладная статистика. Методы обработки данных. Основные требования и характеристики / Орлов А.И., Фомин В.Н. и др. - М.: ВНИИСтандартизации, 1987. 62 с.
4. Колмогоров А.Н. Основные понятия теории вероятностей. – М.-Л.: ОНТИ, 1936. 80 с.
5. Колмогоров А.Н. Теория информации и теория алгоритмов. – М.: Наука, 1987. 304 с.
6. Гнеденко Б.В. Курс теории вероятностей: Учебник. 7-е изд., исправл. - М.: Эдиториал УРСС, 2001. 320 с.
7. Орлов А.И. Устойчивость в социально-экономических моделях. – М.: Наука, 1979. 296 с.
8. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1965 (1-е изд.), 1968 (2-е изд.), 1983 (3-е изд.).
9. Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики для технических приложений. – М.: Наука, 1969. 512 с.
10. Колмогоров А.Н. О логарифмически нормальном законе распределения размеров частиц при дроблении / Доклады АН СССР. 1941. Т.31. С.99-101.
11. Крамер Г. Математические методы статистики. – М.: Мир, 1975. 648 с.
12. Прохоров Ю.В., Розанов Ю.А. Теория вероятностей. (Основные понятия. Предельные теоремы. Случайные процессы.) – М.: Наука, 1973. 496 с.
13. Камень Ю.Э., Камень Я.Э., Орлов А.И. Реальные и номинальные уровни значимости в задачах проверки статистических гипотез. - Журнал «Заводская лаборатория». 1986. Т.52. No.12. С.55-57.
14. Орлов А.И. О нецелесообразности использования итеративных процедур нахождения оценок максимального правдоподобия. – Журнал «Заводская лаборатория», 1986, т.52, No.5, с.67-69.
15. Орлов А.И. Распространенная ошибка при использовании критериев Колмогорова и омега-квадрат. – Журнал «Заводская лаборатория».1985. Т.51. No.1. С.60-62.
16. Кендалл М.Дж., Стьюарт А. Статистические выводы и связи. - М.: Наука, 1973. – 900 с.
17. Себер Дж. Линейный регрессионный анализ. - М.: Мир, 1980. - 456 с.
18. Математическая теория планирования эксперимента / Под ред. С.М.Ермакова. - М.: Наука, 1983. – 392 с.
19. Холлендер М., Вульф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983. - 518 с.
20. Кендалл М.Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. - М.: Наука, 1976. – 736 с.
21. Орлов А.И. Некоторые неклассические постановки в регрессионном анализе и теории классификации. - В сб.: Программно-алгоритмическое обеспечение анализа данных в медико-биологических исследованиях. - М.: Наука, 1987. с.27-40.

## Контрольные вопросы и задачи

1. Расскажите о понятиях случайного события и его вероятности.
2. Почему закон больших чисел и центральная предельная теорема занимают центральное место в вероятностно-статистических методах принятия решений?
3. Чем многомерный статистический анализ отличается от статистики объектов нечисловой природы?
4. Имеются три одинаковые с виду ящика. В первом $a$ белых шаров и $b$ черных; во втором $c$ белых и $d$ черных; в третьем только белые шары. Некто подходит наугад к одному из ящиков и вынимает из нее один шар. Найдите вероятность того, что этот шар белый.
5. Пассажир может воспользоваться трамваями двух маршрутов, следующих с интервалами $T_1$ и $T_2$ соответственно. Пассажир может прийти на остановку в некоторый произвольный момент времени. Какой может быть вероятность того, что пассажир, пришедший на остановку, будет ждать не дольше $t$, где $0<t<\min(T_1,T_2)$?
6. Два стрелка, независимо один от другого, делают по два выстрела (каждый по своей мишени). Вероятность попадания в мишень при одном выстреле для первого стрелка $p_1$, для второго $p_2$.Выигравшим соревнование считается тот стрелок, в мишени которого будет больше пробоин. Найти вероятность того, что выиграет первый стрелок.
**7.** Полная колода карт(52 листа) делится наугад на две равные пачки по 26 листов. Найти вероятности следующих событий:

A - в каждой из пачек окажется по два туза;
B - в одной из пачек не будет ни одного туза, а в другой все четыре;
C - в одной из пачек будет один туз, а в другой три.

8. Случайная величина $X$ принимает значения 0 и 1, а случайная величина $Y$ - значения (-1), 0 и 1. Вероятности $P(X=i, Y=j)$ задаются таблицей:

| $P(X=i, Y=j)$ | Y = - 1 | Y = 0 | Y = 1 |
|---|---|---|---|
| X = 0 | 1/16 | 1/4 | 1/16 |
| X = 1 | 1/16 | 1/4 | 5/16 |

Найдите распределение случайной величины $Z = XY$, ее математическое ожидание и дисперсию.
9. В условиях задачи 8 найдите распределение случайной величины $W = X/(Y+3)$, ее математическое ожидание и дисперсию.
10. Даны независимые случайные величины $X$ и $Y$ такие, что $M(X) = 1$ , $D(X) = 3$, $M(Y) = -1$, $D(Y) = 2$. Найдите $M(aX + bY)$ и $D(aX + bY)$, где $a= 3$ , $b= -2$.

## Темы докладов, рефератов, исследовательских работ

1. Описание данных с помощью гистограмм и непараметрических оценок плотности.
2. Сравнительный анализ методов оценивания параметров и характеристик.
3. Преимущества одношаговых оценок по сравнению с оценками метода максимального правдоподобия.
4. Непараметрический регрессионный анализ.
5. Аксиоматическое введение метрик и их использование в статистике объектов нечисловой природы.
6. Законы больших чисел в пространствах произвольной природы, в том числе в дискретных пространствах.
7. Оптимизационные постановки в вероятностно-статистических задачах принятия решений.

# Глава 1.3. Выборочные исследования

Термин "выборочные исследования" применяют, когда невозможно изучить все единицы представляющей интерес совокупности. Приходится знакомиться с частью совокупности - с выборкой, а затем с помощью статистических методов и моделей переносить выводы с выборки на всю совокупность. Выборочные исследования – способ получения статистических данных и важная часть прикладной статистики. В качестве примера рассмотрим выборочные исследования предпочтений потребителей, которые часто проводят специалисты по маркетингу.

## 1.3.1. Применение случайной выборки (на примере оценивания функции спроса)

Функция спроса часто встречается в учебниках по экономической теории, но при этом обычно не рассказывается, как она получена. Между тем оценить ее по эмпирическим данным не так уж трудно. Например, можно выяснять ожидаемый спрос с помощью следующего простого приема - спрашиваем потенциальных потребителей: "Какую максимальную цену Вы заплатили бы за такой-то товар?" Пусть для определенности речь идет о конкретном учебном пособии по менеджменту. В одном из экспериментов выборка состояла из 20 опрошенных. Они назвали следующие максимально допустимые для них цены (в рублях по состоянию на сентябрь 1998 г.):

40, 25, 30, 50, 35, 20, 50, 32, 15, 40, 20, 40, 45, 30, 50, 25, 35, 20, 35, 40.

Первым делом названные величины надо упорядочить в порядке возрастания. Результаты представлены в табл.1. В первом столбце - номера различных численных значений (в порядке возрастания), названных потребителями. Во втором столбце приведены сами значения цены, названные ими. В третьем столбце указано, сколько раз названо то или иное значение.

Таблица 1.

Эмпирическая оценка функции спроса и ее использование

| № п/п ($i$) | Цена $p_i$ | Повторы $N_i$ | Спрос $D(p_i)$ | Прибыль ($p$-10)D($p$) | Прибыль ($p$-15)$D(p)$ | Прибыль ($p$-25)$D(p)$ |
|---|---|---|---|---|---|---|
| **1** | **15** | **1** | **20** | **100** | **0** | **-** |
| **2** | **20** | **3** | **19** | **190** | **95** | **-** |
| **3** | **25** | **2** | **16** | **240** | **160** | **0** |
| **4** | **30** | **2** | **14** | **280** | **210** | **70** |
| **5** | **32** | **1** | **12** | **264** | **204** | **84** |
| **6** | **35** | **3** | **11** | **275** | **220** | **110** |
| **7** | **40** | **4** | **8** | **240** | **200** | **120** |
| **8** | **45** | **1** | **4** | **140** | **120** | **80** |
| **9** | **50** | **3** | **3** | **120** | **105** | **75** |

Таким образом, 20 потребителей назвали 9 конкретных значений цены (максимально допустимых, или приемлемых для них значений), каждое из значений, как видно из третьего столбца, названо от 1 до 4 раз. Теперь легко построить выборочную функцию спроса в зависимости от цены. Она будет представлена в четвертом столбце, который заполним снизу вверх. Спрос как функция от цены $p$ обозначен $D(p)$ (от *demand*(англ.) – спрос). Если мы будем предлагать товар по цене свыше 50 руб., то его не купит никто из опрошенных. При цене 50 руб. появляются 3 покупателя. Записываем 3 в четвертый столбец в девятую строку. А если цену понизить до 45? Тогда товар купят четверо – тот единственный, для кого максимально возможная цена - 45, и те трое, кто был согласен на более высокую цену – 50 руб. Таким образом, легко заполнить столбец 4, действуя по правилу: значение в клетке четвертого столбца равно сумме значений в находящейся слева клетке третьего столбца и в лежащей снизу клетке четвертого столбца. Например, за 30 руб. купят товар 14 человек, а за 20 руб. - 19.

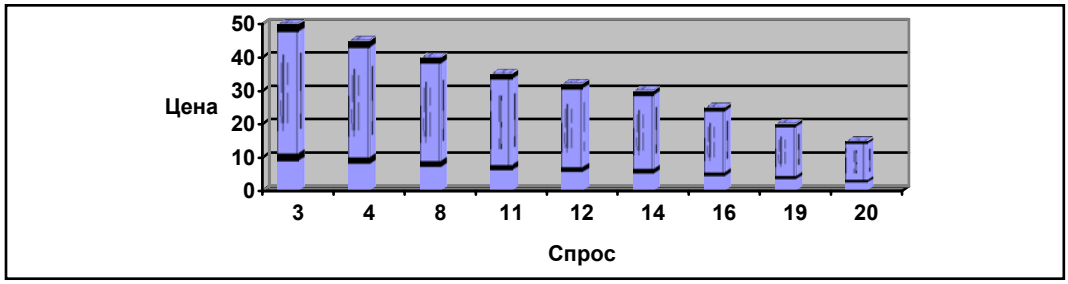
Зависимость спроса от цены - это зависимость четвертого столбца от второго. Табл.1 дает нам девять точек такой зависимости. Зависимость можно представить на рисунке, в координатах «спрос – цена». Если абсцисса - это спрос, а ордината - цена, то девять точек на кривой спроса, перечисленные в порядке возрастания абсциссы, имеют вид:

(3; 50), (4; 45), (8; 40), (11; 35), (12; 32), (14; 30),
(16; 25), (19; 20), (20; 15).

Эти девять точек можно использовать для построения кривой спроса каким-либо графическим или расчетным способом, например, методом наименьших квадратов (см. ниже главу 3.2). Кривая спроса, как и должно быть согласно учебникам экономической теории,

убывает, имея направления от левого верхнего угла чертежа к правому. Однако заметны отклонения от гладкого вида функции, связанные, в частности, с естественным пристрастием потребителей к круглым числам. Заметьте, все опрошенные, кроме одного, назвали числа, кратные 5 руб.

**Рис.1. Кривая спроса**




.

Данные табл.1 могут быть использованы для выбора цены продавцом-монополистом. Или организацией, действующем на рынке монополистической конкуренции. Пусть расходы на изготовление или оптовую покупку единицы товара равны 10 руб. Например, оптовая цена книги - 10 руб. По какой цене ее продавать на том рынке, функцию спроса для которого мы только что нашли? Для ответа на этот вопрос вычислим суммарную прибыль, т.е. произведение прибыли на одном экземпляре ($p$-10) на число проданных (точнее, запрошенных) экземпляров $D(p)$. Результаты приведены в пятом столбце табл.1. Максимальная прибыль, равная 280 руб., достигается при цене 30 руб. за экземпляр. При этом из 20 потенциальных покупателей окажутся в состоянии заплатить за книгу 14, т.е. 70% .

Если же удельные издержки производства, приходящиеся на одну книгу (или оптовая цена), повысятся до 15 руб., то данные столбца 6 табл.1 показывают, что максимальная прибыль, равная 220 руб. (она, разумеется, меньше, чем в предыдущем случае), достигается при более высокой цене - 35 руб. Эта цена доступна 11 потенциальным покупателям, т.е. 55% от всех возможных покупателей. При дальнейшем повышении издержек, скажем, до 25 руб., как вытекает из данных столбца 7 табл.1, максимальная прибыль, равная 120 руб., достигается при цене 40 руб. за единицу товара, что доступно 8 лицам, т.е. 40% покупателей. Отметьте, что при повышении оптовой цены на 10 руб. оказалось выгодным увеличить розничную лишь на 5, поскольку более резкое повышение привело бы к такому сокращению спроса, которое перекрыло бы эффект от повышения удельной прибыли (т.е. прибыли, приходящейся на одну проданную книгу).

Представляет интерес анализ оптимального объема выпуска при различных значениях удельных издержек (табл.2).

В табл.2 звездочками указаны максимальные значения прибыли при том или ином значении издержек, не включенном в табл.1. Для легкости обозрения результаты об оптимальных объемах выпуска и соответствующих ценах из табл. 1 и 2 приведены в табл.3.

Таблица 2.

Прибыль при различных значениях издержек

| № ($i$) | Цена $p_i$ | Спрос $D(p_i)$ | Прибыль (р-5)$D(p)$ | Прибыль ($p$-20)$D(p)$ | Прибыль ($p$-30)$D(p)$ | Прибыль ($p$-35)$D(p)$ | Прибыль ($p$-40)$D(p)$ |
|---|---|---|---|---|---|---|---|
| **1** | **15** | **20** | **200** | **-** | **-** | **-** | **-** |
| **2** | **20** | **19** | **285** | **0** | **-** | **-** | **-** |
| **3** | **25** | **16** | **320** | **80** | **-** | **-** | **-** |
| **4** | **30** | **14** | **350 *** | **140** | **0** | **-** | **-** |
| **5** | **32** | **12** | **324** | **144** | **24** | **-** | **-** |
| **6** | **35** | **11** | **330** | **165 *** | **55** | **0** | **-** |
| **7** | **40** | **8** | **280** | **160** | **80 *** | **40** | **0** |
| **8** | **45** | **4** | **160** | **100** | **60** | **40** | **20** |
| **9** | **50** | **3** | **135** | **90** | **60** | **45 *** | **30 *** |

Таблица 3.

Зависимость оптимального выпуска и цены от издержек

| **Издержки** | **5** | **10** | **15** | **20** | **25** | **30** | **35** | **40** |
|---|---|---|---|---|---|---|---|---|
| **Оптимальный выпуск** | **14** | **14** | **11** | **11** | **8** | **8** | **3** | **3** |
| **Цена** | **30** | **30** | **35** | **35** | **40** | **40** | **50** | **50** |

Как видно из табл.3, с ростом издержек оптимальный выпуск падает, а цена растет. При этом изменение издержек на 5 единиц может вызывать, а может и не вызывать повышения цены. В этом проявляется микроструктура функции спроса – небольшое повышение цены может привести к тому, что значительные группы покупателей откажутся от покупок, и прибыль упадет.

Этот эффект напоминает известное в экономической теории разделение налогового бремени между производителем и потребителем. Неверно говорить, что производитель перекладывает издержки или, конкретно, налоги, на потребителя, повышая цену на их величину, поскольку при этом сокращается спрос (и выпуск), а потому и прибыль производителя.

Дальнейшее ясно - если оптовая цена будет повышаться, то и дающая максимальную прибыль розничная цена также будет повышаться, и все меньшая доля покупателей сможет приобрести товар. Крайняя точка - оптовая цена, равная 45 руб. Тогда только трое (15 %) купят товар за 50 руб., а прибыль продавца составит только 15 руб. Наглядно видно, что повышение издержек производства приводит к ориентации производителя на наиболее богатые слои населения, но и повышение цен (до оптимального для монополиста-производителя уровня) не приводит к повышению прибыли, напротив, она снижается, и при этом большинство потенциальных потребителей не в состоянии купить товар. Таково влияние инфляции издержек на экономическую жизнь.

Отметим, что рыночные структуры не в состоянии обеспечить всех желающих – это просто не выгодно. Так, из 20 опрошенных лишь 14, т.е. 70%, могут рассчитывать на покупку, даже при минимальных издержках и ценах. Если общество желает чем-либо обеспечить всех граждан, оно должно раздавать это благо бесплатно, как это делается, например, с учебниками в школах.

Описанный выше метод оценивания спроса был разработан в Институте высоких статистических технологий и эконометрики в 1993 г.

Для изучения предпочтений потребителей часто используют более изощренные методы. Рассмотрим некоторые из них.

### 1.3.2. Маркетинговые опросы потребителей

Потенциального покупателя интересует не только цена, но и качество товара, красота упаковки (например, для подарочных наборов конфет) и многое другое. Хочешь узнать, чего

желает потребитель - спроси его. Эта простая мысль объясняет популярность маркетинговых опросов.

Бесспорно, что основная цель производственной и торговой деятельности - удовлетворение потребностей людей. Как получить представление об этих потребностях? Очевидно, необходимо опросить потребителей. В американском учебнике по рекламному делу [1] подробно рассматриваются различные методы опроса потребителей и обработки результатов с помощью методов эконометрики. Расскажем о результатах опроса потребителей растворимого кофе. Исследование проведено Институтом высоких статистических технологий и эконометрики по заказу АОЗТ "Д-2" в апреле 1994 г. в Москве.

**Сбор данных.** Один из важнейших разделов прикладной статистики – сбор данных. Обсудим постановку задачи в случае опроса потребителей растворимого кофе. Заказчика интересуют предпочтения как продавцов кофе (розничных и мелкооптовых), так и непосредственно потребителей. В результате совместного обсуждения было признано целесообразным использовать для опроса и тех, и других одну и ту же анкету из 14 основных и 4 социально-демографических вопросов с добавлением двух вопросов специально для продавцов. Анкета была разработана совместно представителями заказчика и исполнителя и утверждена заказчиком. В табл.4 приведен несколько сокращенный вариант этой анкеты.

Таблица 4.

Анкета для потребителей растворимого кофе (в сокращении)

---

Дорогой потребитель растворимого кофе,

Институт высоких статистических технологий и эконометрики просит Вас ответить на несколько простых вопросов о том, какой кофе Вы любите. Ваши ответы позволят составить объективное представление о вкусах российских любителей кофе и будут способствовать повышению качества этого товара на российском рынке.

1.Часто ли Вы пьете растворимый кофе: иногда, каждый день 1 чашку, 2-3 чашки, больше, чем 3 чашки.

(Здесь и далее подчеркните нужное.)

2. Что Вы цените в кофе: вкус, аромат, крепость, цвет, отсутствие вредных для здоровья веществ, что-либо еще (сообщите нам, что именно).

3. Как часто покупаете кофе: по мере надобности или по возможности?

4. Любите ли Вы бразильский растворимый кофе? Да, нет, не знаю.

5. Какой объем упаковки Вы предпочитаете: в пакетиках, маленькая банка, средняя банка, большая банка, обязательно стеклянная банка, все равно.

6. Где покупаете растворимый кофе: в ларьках, в продуктовых магазинах, в специализированных отделах и магазинах, все равно, где купить, где-либо еще (опишите, пожалуйста).

7. Были ли случаи, когда купленный Вами кофе оказывался низкого качества? Да, нет.

8. Согласны ли Вы, что за высокое и гарантированное качество продукта можно и заплатить несколько дороже? Да, нет.

9. Какой кофе Вы предпочтете купить: банка неизвестного качества за 2000 руб. или продукт того же веса, безопасность которого гарантирована Минздравом России, за 2500 руб.? Первый, второй.

10. Считаете ли Вы нужным, чтобы производитель принял меры для того, чтобы вредные для здоровья вещества, в частности, ионы тяжелых металлов, не проникали из материала упаковки непосредственно в растворимый кофе? Да, нет.

Институт высоких статистических технологий и эконометрики предполагает сравнить потребительские предпочтения различных категорий жителей нашей страны. Поэтому просим ответить еще на несколько вопросов.

11. Пол: женский, мужской.

12. Возраст: до 20, 20-30, 30-50, более 50.

13. Род занятий: учащийся, работающий, пенсионер, инженер, врач, преподаватель, служащий, менеджер, предприниматель, научный работник, рабочий, др. (пожалуйста, расшифруйте).

14. Вся Ваша семья любит растворимый кофе или же Вы - единственный любитель этого восхитительного напитка современного человека? Вся семья, я один (одна).

15. Согласились бы Вы и в дальнейшем участвовать в опросах потребителей относительно качества различных пищевых продуктов (чай, джем и др.). Если "да", то сообщите свой адрес, телефон, имя и отчество.

Спасибо за Ваше содействие работе по повышению качества продуктов на российском рынке!

---

**Выбор метода опроса.** Широко применяются процедуры опроса, когда респонденты (так социологи и маркетологи называют тех, от кого получают информацию, т.е. опрашиваемых) самостоятельно заполняют анкеты (розданные им или полученные по почте), а также личные и телефонные интервью. Из этих процедур нами было выбрано личное интервью по следующим причинам.

Возврат почтовых анкет сравнительно невелик (в данном случае можно было ожидать не более 5-10%), оттянут по времени и искажает структуру совокупности потребителей (наиболее динамичные люди вряд ли найдут время для ответа на подобную анкету). Кроме того, есть проблемы с почтовой связью (постоянное изменение тарифов затрудняет возмещение респондентам почтовых расходов и др.).

Самостоятельное заполнение анкеты, как показали специально проведенные эксперименты, не позволяет получить полные ответы на поставленные вопросы (респондент утомляется или отвлекается, отказывается отвечать на часть вопросов, иногда не понимает их или отвечает не по существу). Некоторые категории респондентов, например, продавцы в киосках, отказываются заполнять анкеты, но готовы устно ответить на вопросы.

Телефонный опрос искажает совокупность потребителей, поскольку наиболее активных индивидуумов трудно застать дома и уговорить ответить на вопросы анкеты. Репрезентативность нарушается также и потому, что на один номер телефона может приходиться различное количество продавцов и потребителей растворимого кофе, а некоторые из них не имеют телефонов вообще. Анкета достаточно длинна, и разговор по домашнему и тем более служебному телефону респондента может быть прекращен досрочно по его инициативе. Иногородних продавцов и потребителей растворимого кофе, приехавших в Москву, по телефону опросить практически невозможно.

Метод личного интервью лишен перечисленных недостатков. Соответствующим образом подготовленный интервьюер, получив согласие на интервью, удерживает внимание собеседника на анкете, добивается получения ответов на все её вопросы, контролируя при этом соответствие ответов реальной позиции респондента. Ясно, что успех интервьюирования зависит от личных качеств и подготовки интервьюера. Однако расходы на получение одной анкеты при использовании этого метода больше, чем для других рассмотренных методов.

**Формулировки вопросов**. В маркетинговых и социологических опросах используют три типа вопросов - закрытые, открытые и полузакрытые, они же полуоткрытые. При ответе на закрытые вопросы респондент может выбирать лишь из сформулированных составителями анкеты вариантов ответа. В качестве ответа на открытые вопросы респондента просят изложить свое мнение в свободной форме. Полузакрытые, они же полуоткрытые вопросы занимают промежуточное положение - кроме перечисленных в анкете вариантов, респондент может добавить свои соображения.

В социологических публикациях, посвященных выборочным исследованиям, продолжается дискуссия по поводу "мягких" и "жестких" форм сбора данных, т.е. фактически о том, какого типа вопросы более целесообразно использовать - открытые или закрытые (см., например, статью директора Института социологии РАН В.А. Ядова [2]).

Преимущество открытых вопросов состоит в том, что респондент может свободно высказать свое мнение так, как сочтет нужным. Их недостаток - в сложности сопоставления мнений различных респондентов. Для такого сопоставления и получения сводных характеристик организаторы опроса вынуждены сами шифровать ответы на открытые вопросы, применяя разработанную ими схему шифровки.

Преимущество закрытых вопросов в том и состоит, что такую шифровку проводит сам респондент. Однако при этом организаторы опроса уподобляются древнегреческому мифическому персонажу Прокрусту. Как известно, Прокруст приглашал путников заночевать у него. Укладывал их на кровать. Если путник был маленького роста, он вытягивал его ноги так, чтобы они доставали до конца кровати. Если же путник оказывался высоким и ноги его торчали - он обрубал их так, чтобы достигнуть стандарта: "рост" путника должен равняться длине кровати. Так и организаторы опроса, применяя закрытые вопросы, заставляют респондента "вытягивать" или "обрубать" свое мнение, чтобы выразить его с помощью приведенных в формулировке вопроса возможных ответов.

Ясно, что для обработки данных по группам и сравнения групп между собой нужны формализованные данные, и фактически речь может идти лишь о том, кто - респондент или маркетолог (социолог, психолог и др.) - будет шифровать ответы. В проекте "Потребители растворимого кофе" практически для всех вопросов варианты ответов можно перечислить заранее, т.е. можно широко использовать закрытые вопросы. В отличие от опросов с вопросами типа: "Одобряете ли Вы идущие в России реформы?", в которых естественно просить респондента расшифровать, что он понимает под "реформами" (открытый вопрос). Поэтому в используемой в описываемом проекте анкете использовались в основном закрытые и полузакрытые вопросы. Как показали результаты обработки, этот подход оказался правильным - лишь в небольшом числе анкет оказались вписаны свои варианты ответов. Вместе с тем демонстрировалось уважение к мнению респондента, не выдвигалось требование обязательного выбора из заданного множества ответов - респондент мог добавить свое, но редко пользовался этой возможностью (не более чем в 5% случаев).

В последнем вопросе анкеты респонденту предлагалось стать постоянным участником опросов о качестве товаров народного потребления. Ряд респондентов откликнулся на это предложение, в результате стало возможным развертывание постоянной сети "экспертов по качеству", подобной аналогичным в США и других странах.

**Обоснование объема выборки и проведение опроса.** Математико-статистические вероятностные модели выборочных маркетинговых и социологических исследований часто опираются на предположение о том, что выборку можно рассматривать как "случайную выборки из конечной совокупности" (см. главу 1.2). Типа той, когда из списков избирателей с помощью датчика случайных чисел отбирается необходимое число номеров для формирования жюри присяжных заседателей. В рассматриваемом проекте нельзя обеспечить формирование подобной выборки - не существует реестра потребителей растворимого кофе. Однако в этом и нет необходимости. Поскольку гипергеометрическое распределение хорошо приближается биномиальным, если объем выборки по крайней мере в 10 раз меньше объема всей совокупности (в рассматриваемом случае это так), то правомерно использование биномиальной модели, согласно которой мнение респондента (ответы на вопросы анкеты) рассматривается как случайный вектор, а все такие вектора независимы между собой. Другими словами, можно использовать модель простой случайной выборки. В среде специалистов, изучающих поведение человека (маркетологов, социологов, психологов, политологов и др.) давно идет дискуссия о роли случайности в поведении человека. А именно, о том, есть ли случайность в поведении отдельно взятого человека или же случайность проявляется лишь в отборе выборки из генеральной совокупности. Таким образом, сформулированные выше результаты прикладной статистики показывают, что позиция в давней дискуссии практически не влияет на алгоритмы обработки данных.

В биномиальной модели выборки оценивание характеристик происходит тем точнее, чем объем выборки больше. Часто спрашивают: "Какой объем выборки нужен?" В прикладной статистике есть методы определения необходимого объема выборки. Они основаны на разных подходах. Либо на задании необходимой точности оценивания параметров. Либо на явной формулировке альтернативных гипотез, между которыми необходимо сделать выбор. Либо на учете погрешностей измерений (методы статистики интервальных данных, см. главу 3.5). Ни один из этих подходов нельзя применить в рассматриваемом случае.

**Биномиальная модель выборки.** Она применяется для описания ответов на закрытые вопросы, имеющие две подсказки, например, "да" и "нет". Конечно, пары подсказок могут быть

иными. Например, "согласен" и "не согласен". Или при опросе потребителей кондитерских товаров первая подсказка может иметь такой вид: "Больше люблю "Марс", чем "Сникерс". А вторая тогда такова: "Больше люблю "Сникерс", чем "Марс".

Пусть объем выборки равен $n$. Тогда ответы опрашиваемых можно представить как $X_1$, $X_2$,...,$X_n$, где $X_i = 1$, если $i$-й респондент выбрал первую подсказку, и $X_i = 0$, если $i$-й респондент выбрал вторую подсказку, $i$=1,2,…,$n$. В вероятностной модели предполагается, что случайные величины $X_1$, $X_2$,...,$X_n$ независимы и одинаково распределены. Поскольку эти случайные величины принимают два значения, то ситуация описывается одним параметром $p$ - долей выбирающих первую подсказку во всей генеральной совокупности. Тогда

$$P(X_i = 1) = p,\ P(X_i = 0) = 1-p,\ i=1,2,...,n.$$

Пусть $m = X_1 + X_2 + ... + X_n$. Оценкой вероятности p является частота $p^* = m/n$. При этом математическое ожидание $M(p^*)$ и дисперсия $D(p^*)$ имеют вид

$$M(p^*) = p,\ D(p^*) = p(1-p)/n.$$

По Закону Больших Чисел (ЗБЧ) теории вероятностей (в данном случае - по теореме Бернулли) частота $p^*$ сходится (т.е. безгранично приближается) к вероятности $p$ при росте объема выборки. Это и означает, что оценивание проводится тем точнее, чем больше объем выборки. Точность оценивания можно указать. Займемся этим.

По теореме Муавра-Лапласа теории вероятностей

$$\lim_{n\to\infty} P\{\frac{m-np}{\sqrt{np(1-p)}} \le x\} = \Phi(x),$$

где $\Phi(x)$ - функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy,$$

где $\pi = 3{,}1415925\ldots$-отношение длины окружности к ее диаметру, e= 2,718281828… - основание натуральных логарифмов. График плотности стандартного нормального распределения

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$

был очень точно изображен на германской денежной банкноте в 10 немецких марок. Эта банкнота была посвящена великому немецкому математику Карлу Гауссу (1777-1855), среди основных работ которого есть относящиеся к нормальному распределению. К сожалению для прикладной статистики, в настоящее время национальные валюты, в том числе немецкая, заменены на единую – евро.

В настоящее время нет необходимости вычислять функцию стандартного нормального распределения и ее плотность по приведенным выше формулам, поскольку давно составлены подробные таблицы (см., например, [3]), а распространенные программные продукты содержат алгоритмы нахождения этих функций.

С помощью теоремы Муавра-Лапласа могут быть построены доверительные интервалы для неизвестной эконометрику вероятности. Сначала заметим, что из этой теоремы непосредственно следует, что

$$\lim_{n\to\infty} P\{-x \le \frac{m-np}{\sqrt{np(1-p)}} \le x\} = \Phi(x) - \Phi(-x).$$

Поскольку функция стандартного нормального распределения симметрична относительно 0, т.е. $\Phi(x) + \Phi(-x) = 1,$ то $\Phi(x) - \Phi(-x) = 2\Phi(x) - 1.$

Зададим доверительную вероятность $\gamma$. Пусть $U(\gamma)$ удовлетворяет условию

$$\Phi(U(\gamma)) - \Phi(-U(\gamma)) = \gamma,$$

т.е.

$$U(\gamma) = \Phi^{-1}(\frac{1+\gamma}{2}).$$

Из последнего предельного соотношения следует, что

$$\lim_{n\to\infty} P\{p^* - U(\gamma)\frac{\sqrt{p(1-p)}}{\sqrt{n}} \le p \le p^* + U(\gamma)\frac{\sqrt{p(1-p)}}{\sqrt{n}}\} = \gamma.$$

К сожалению, это соотношение нельзя непосредственно использовать для доверительного оценивания, поскольку верхняя и нижняя границы зависят от неизвестной вероятности. Однако с помощью метода наследования сходимости (см. главу 1.4 или [4, п.2.4]) можно доказать, что

Следовательно, нижняя доверительная граница имеет вид

$$p_{нижн} = p^* - U(\gamma)\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}},$$

в то время как верхняя доверительная граница такова:

$$\lim_{n\to\infty} P\{p^* - U(\gamma)\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}} \le p \le p^* + U(\gamma)\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}}\} = \gamma.$$

$$p_{верх} = p^* + U(\gamma)\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}}.$$

Наиболее распространенным (в прикладных исследованиях) значением доверительной вероятности является $\gamma = 0{,}95$. Иногда употребляют термин "95% доверительный интервал". Тогда $U(\gamma) = 1{,}96$.

*Пример 1*. Пусть *n*=500, *m*=200. Тогда *p** =0,40. Найдем доверительный интервал для $\gamma = 0{,}95$:

$$p_{нижн} = 0{,}40 - 1{,}96\frac{\sqrt{0{,}4\times 0{,}6}}{\sqrt{500}} = 0{,}40 - 0{,}043 = 0{,}357, \quad p_{верх} = 0{,}40 + 0{,}043 = 0{,}443.$$

Таким образом, хотя в достаточно большой выборке 40% респондентов говорят "да", можно утверждать лишь, что во всей генеральной совокупности таких от 35,7% до 44,3% - крайние значения отличаются на 8,6%.

*Замечание.* С достаточной для практики точностью можно заменить 1,96 на 2.

Удобные для использования в практической работе специалиста по выборочным исследованиям, маркетолога и социолога таблицы точности оценивания разработаны во ВЦИОМ (Всероссийском центре по изучению общественного мнения). Приведем здесь несколько модифицированный вариант одной из них.

Таблица 5.

Допустимая величина ошибки выборки (в процентах)

| Объем группы<br>Доля p* | 1000 | 750 | 600 | 400 | 200 | 100 |
|---|---|---|---|---|---|---|
| Около 10% или 90% | 2 | 3 | 3 | 4 | 5 | 7 |
| Около 20% или 80% | 3 | 4 | 4 | 5 | 7 | 9 |
| Около 30% или 70% | 4 | 4 | 4 | 6 | 9 | 10 |
| Около 40% или 60% | 4 | 4 | 5 | 6 | 8 | 11 |
| Около 50% | 4 | 4 | 5 | 6 | 8 | 11 |

В условиях рассмотренного выше примера надо взять вторую снизу строку. Объема выборки 500 нет в таблице, но есть объемы 400 и 600, которым соответствуют ошибки в 6% и 5% соответственно. Следовательно, в условиях примера целесообразно оценить ошибку как ((5+6)/2)% = 5,5%. Эта величина несколько больше, чем рассчитанная выше (4,3%). С чем связано это различие? Дело в том, что таблица ВЦИОМ связана не с доверительной вероятностью $\gamma = 0{,}95$, а с доверительной вероятностью $\gamma = 0{,}99$, которой соответствует множитель $U(\gamma) = 2{,}58$. Расчет ошибки по приведенным выше формулам дает 5,65%, что практически совпадает со значением, найденным по табл.5.

Минимальный из обычно используемых объемов выборки n в маркетинговых или социологических исследованиях - 100, максимальный - до 5000 (обычно в исследованиях, охватывающих ряд регионов страны, т.е. фактически разбивающихся на ряд отдельных исследований - как в ряде исследований ВЦИОМ). По данным Института социологии Российской академии наук [5], среднее число анкет в социологическом исследовании не превышает 700. Поскольку стоимость исследования растет по крайней мере как линейная функция объема выборки, а точность повышается как квадратный корень из этого объема, то верхняя граница объема выборки определяется обычно из экономических соображений. Объемы пилотных исследований (т.е. проводящихся впервые, предварительно или как первые в сериях подобных) обычно ниже, чем объемы исследований по обкатанной программе.

Нижняя граница определяется тем, что в минимальной по численности анализируемой подгруппе должно быть несколько десятков человек (не менее 30), поскольку по ответам попавших в эту подгруппу необходимо сделать обоснованные заключения о предпочтениях соответствующей подгруппы в совокупности всех потребителей растворимого кофе. Учитывая деление опрашиваемых на продавцов и покупателей, на мужчин и женщин, на четыре градации по возрасту и восемь - по роду занятий, наличие 5 - 6 подсказок во многих вопросах, приходим к выводу о том, что в рассматриваемом проекте объем выборки должен быть не менее 400 - 500. Вместе с тем существенное превышение этого объема нецелесообразно, поскольку исследование является пилотным.

Поэтому в проекте «Потребители растворимого кофе» объем выборки был выбран равным 500. Анализ полученных результатов (см. ниже) позволяет утверждать, что в соответствии с целями исследования выборку следует считать репрезентативной.

**Организация опроса.** Интервьерами работали молодые люди – студенты первого курса экономико-математического факультета Московского государственного института электроники и математики (технического университета) и лицея No.1140, проходившие обучение по экономике, всего 40 человек, имеющих специальную подготовку по изучению рынка и проведению маркетинговых опросов потребителей и продавцов (в объеме 8 часов). Опрос продавцов проводился на рынках г. Москвы, действующих в Лужниках, у Киевского вокзала и в других местах. Опрос покупателей проводился на рынках, в магазинах, на улицах около киосков и ларьков, а также в домашней и служебной обстановке.

Большое внимание уделялось качеству заполнения анкет. Интервьюеры были разбиты на шесть бригад, бригадиры персонально отвечали за качество заполнения анкет. Второй уровень контроля осуществляла специально созданная "группа организации опроса", третий происходил при вводе информации в базу данных. Каждая анкета заверена подписями интервьюера и бригадира, на ней указано место и время интервьюирования. Поэтому необходимо признать высокую достоверность собранных анкет.

**Обработка данных.** В соответствии с целью исследования основной метод первичной обработки данных - построение частотных таблиц для ответов на отдельные вопросы. Кроме того, проводилось сравнение различных групп потребителей и продавцов, выделенных по социально-демографическим данным, с помощью критериев проверки однородности выборок (см. ниже). При более углубленном анализе применялись различные методы статистики объектов нечисловой природы (более 90 % маркетинговых и социологических данных имеют нечисловую природу [6]). Использовались средства графического представления данных.

**Итоги опроса.** Итак, по заданию одной из торговых фирм были изучены предпочтения покупателей и мелкооптовых продавцов растворимого кофе. Совместно с представителями заказчика был составлен опросный лист (анкета типа социологической) из 16 основных вопросов и 4 дополнительных, посвященных социально-демографической информации. Опрос проводился в форме интервью с 500 покупателями и продавцами кофе. Места опроса - рынки, лотки, киоски, продуктовые и специализированные магазины. Другими словами, были охвачены все виды мест продаж кофе. Интервью проводили более 40 специально подготовленных (примерно по 8-часовой программе) студентов, разбитых на 7 бригад. После тщательной проверки бригадирами и группой обработки информация была введена в специально созданную базу данных. Затем проводилась разнообразная статистическая обработка, строились таблицы и

диаграммы, проверялись статистические гипотезы и т.д. Заключительный этап - осмысление и интерпретация данных, подготовка итогового отчета и предложений для заказчиков.

Технология организации и проведения маркетинговых опросов лишь незначительно отличается от технологии социологических опросов, многократно описанной в литературе. Так, мы предпочли использовать полуоткрытые вопросы, в которых для опрашиваемого дан перечень подсказок, а при желании он может высказать свое мнение в свободной форме. Не уложившихся в подсказки оказалось около 5 % , их мнения были внесены в базу данных и анализировались дополнительно. Для повышения надежности опроса о наиболее важных с точки зрения маркетинга моментах спрашивалось в нескольких вопросах. Были вопросы - ловушки, с помощью которых контролировалась "осмысленность" заполнения анкеты. Например, в вопросе: "Что Вы цените в кофе: вкус, аромат, крепость, наличие пенки..." ловушкой является включение "крепости" - ясно, что крепость зависит не от кофе самого по себе, а от его количества в чашке. В ловушку никто из 500 не попался - никто не отметил "крепость". Этот факт свидетельствует о надежности выводов проведенного опроса. Мы считали нецелесообразным задавать вопрос об уровне доходов (поскольку в большинстве случаев отвечают "средний", что невозможно связать с определенной величиной). Вместо такого вопроса мы спрашивали: "Как часто Вы покупаете кофе: по мере надобности или по возможности?". Поскольку кофе не является дефицитным товаром, первый ответ свидетельствовал о наличии достаточных денежных средств, второй - об их ограниченности (потребитель не всегда имел возможность позволить себе купить банку растворимого кофе).

Стоимость подобных исследований - 5-10 долларов США на одного обследованного. При этом трудоемкость (и стоимость) начальной стадии - подготовки анкеты и интервьюеров, пробный опрос и др. - 30 % от стоимости исследования, стоимость непосредственно опроса - тоже 30 %, ввод информации в компьютер и проведение расчетов, построение таблиц и графиков - 20 %, интерпретация результатов, подготовка итогового отчета и предложений для заказчиков - 20 % . Таким образом, стоимость собственно опроса в два с лишним раза меньше стоимости остальных стадий исследования. И в выполнении работы участвуют различные специалисты. На первой стадии – в основном нужны высококвалифицированные аналитики. На второй – многочисленные интервьюеры, в роли которых могут выступать студенты и школьники, прошедшие конкретный курс обучения в 8-10 часов. На третьей – работа с компьютером (надо уметь строить и обсчитывать электронные таблицы или базы данных, использовать статистические пакеты, составлять и печатать таблицы и диаграммы и т.п.). На четвертой – опять в основном нужны высококвалифицированные аналитики.

Приведем некоторые из полученных результатов.

а) В отличие от западных потребителей, отечественные не отдавали предпочтения стеклянным банкам по сравнению с жестяными. Поскольку жестяные банки дешевле стеклянных, то можно было порекомендовать (в 1994 г., когда проходил опрос) с целью снижения расходов закупку кофе в жестяных банках.

б) Отечественные потребители готовы платить на 10-20% больше за экологически безопасный кофе более высокого качества, имеющий сертификат Минздрава и символ экологической безопасности на упаковке.

в) Средний объем потребления растворимого кофе - 850 г в месяц (на семью потребителя).

г) Потребители растворимого кофе могут быть разделены на классы (кластеры в терминологии главы 3.2). Есть "продвинутые" потребители, обращающие большое внимание на качество и экологическую безопасность, марку и страну производства, терпимо относящиеся к изменению цены. Эти "тонкие ценители" - в основном женщины от 30 до 50 лет, служащие, менеджеры, научные работники, преподаватели, врачи (т.е. лица с высшим образованием), пьющие кофе как дома, так и на работе, причем "кофейный ритуал" зачастую входит в процедуру деловых переговоров или совещаний. Противоположный по потребительскому поведению класс состоит из мужчин двух крайних возрастных групп - школьников и пенсионеров. Для них важна только цена, что очевидным образом объясняется недостатком денег.

Результаты были использованы заказчиком в рекламной кампании. В частности, обращалось внимание на сертификат Минздрава и на экологическую безопасность упаковки.

Приведем пример еще одной анкеты из нашего опыта, предназначенной для изучения спроса на образовательные услуги (табл. 6).

Таблица 6.
Исследование рынка образовательных услуг

**ИССЛЕДОВАНИЕ РЫНКА ОБРАЗОВАТЕЛЬНЫХ УСЛУГ**

Анкета для студентов первого курса экономико-математического факультета МГИЭМ(ту).

**А. Объективные данные**

1. Группа
2. Пол
3. Год рождения
4. Женат(замужем) - да/нет

**Б. Общее изучение рынка**

5. Почему Вы выбрали специальность экономиста?
6. Почему Вы выбрали именно МГИЭМ(ту) среди всех вузов Москвы, готовящих экономистов?
7. Как Вы представляете себе будущую деятельность по окончании МГИЭМ(ту)?
8. Есть ли у Вас надежда на то, что приобретаемые сейчас знания окажутся полезными в практической работе? Если нет, то зачем Вы учитесь?

**В. Отношение к платному образованию**

9. Если бы обучение в МГИЭМ(ту) было платным (порядка 1 миллиона руб. в год в ценах февраля 1994 г.), стали бы Вы поступать в МГИЭМ(ту)?
10. Если обучение в МГИЭМ(ту) станет платным, то останетесь ли Вы учиться в МГИЭМ(ту)? (Например, организация оплаты за учебу такова: некоторая фирма заключает контракт со студентом и оплачивает его учебу; студент самостоятельно ищет такую фирму.)
11. Представляет ли для Вас интерес возможность параллельно с дипломом МГИЭМ(ту) получить диплом бакалавра Межкультурного открытого университета (штаб-квартира в Нидерландах) по специальности "бизнес администрейшн" (обучение заочное, стоимость 1780 долларов США за курс)?

**Г. О курсе "Основы экономики"**

12. Нужно ли рассказывать содержание реферата-дайджеста учебника К. Макконнелла и С. Брю "Экономикс: Принципы, проблемы и политика" или считать его общеизвестным и говорить о том, чего в нём нет?
13. Полезен ли электронный учебник? Если нет, то почему?
14. Нужны ли Вам индивидуальные занятия в аудитории (а не в компьютерном классе с электронным учебником) и в каком виде?
15. Какие темы Вы считаете полезным рассмотреть дополнительно?
16. Сформулируйте иные Ваши замечания и предложения по курсу "Основы экономики": по лекциям, практическим и индивидуальным занятиям.

**Д. Дополнительная информация**

17. Какие предметы обучения - самые трудные, какие - самые легкие на первом семестре?
18. Подрабатываете ли Вы? Если согласны, укажите примерную (среднюю) сумму в месяц.
19. Существенна ли для Вас стипендия?
20. Есть ли у Вас дома компьютер?
21. Участвуете ли Вы в каких-либо политических движениях, партиях? Если согласны, назовите.

---

Выпускник программы «Топ-менеджер» Академии народного хозяйства при Правительстве Российской Федерации А.А.Пивень в 2003 г. оценил функцию спроса на продукцию своего предприятия. Расчет и установление оптимальной цены на изделие с точки

зрения максимизации прибыли был произведен по описанному выше методу. В табл. 7 приведена функция ожидаемого спроса в зависимости от цены. Как подсчитал А.А. Пивень, уровень издержек на производство 1 изделия составляет 42824,7 руб. (1350 у.е.). Для удобства все расчеты будем производить в условных единицах.

Таблица 7.

Функция ожидаемого спроса в зависимости от цены.

| № п/п | Цена, у.е. | Объем продаж в год, шт | Издержки на объ-ем производства | Выручка, у.е. | Прибыль, у.е. |
|---|---|---|---|---|---|
| 1 | 1 400 | 1 600 | 2 160 000 | 2 240 000 | 80 000 |
| 2 | 1 500 | 1 500 | 2 025 000 | 2 250 000 | 225 000 |
| 3 | 1 600 | 1 200 | 1 620 000 | 1 920 000 | 300 000 |
| 4 | 1 700 | 1 000 | 1 350 000 | 1 700 000 | 350 000 |
| 5 | 1 800 | 720 | 972 000 | 1 246 000 | 324 000 |
| 6 | 1 900 | 500 | 675 000 | 950 000 | 275 000 |
| 7 | 2 000 | 320 | 432 000 | 640 000 | 208 000 |
| 8 | 2 100 | 170 | 229 500 | 357 000 | 127 500 |
| 9 | 2 200 | 110 | 148 500 | 242 000 | 93 500 |

Как видно из приведенных расчетов, оптимальная цена на подъемник должна находиться в диапазоне 1600 – 1700 у.е.

На основе многомерной регрессионной зависимости (см. главу 3.2) была построена математическая модель рынка. Она довольно точно отражает реальное положение дел. При исходной цене 1650 у.е. продажи ориентировочно должны составить 1010 шт. На рис. 2 приведена кривая спроса.

Эти расчеты были сделаны при допущении, что издержки не меняются в течение длительного промежутка времени. Однако, в реальных условиях постоянный рост стоимости энергоресурсов и непрекращающаяся инфляция издержек (рост затрат на сырье, материалы, комплектующие изделия, рабочую силу) приводит к увеличению издержек. Поэтому А.А. Пивень проанализировал оптимальный объем выпуска при их различных значениях. Данные приведены в табл. 8.

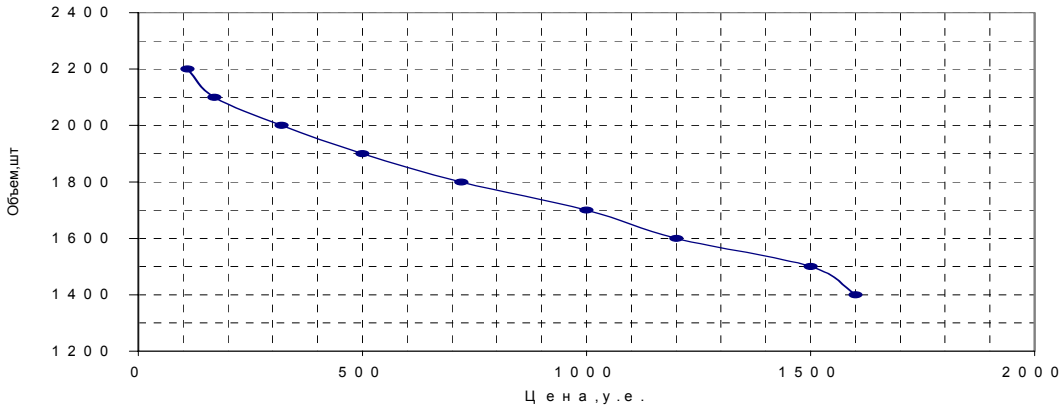



Рис.2. Кривая спроса на изделие.

Таблица 8.

Прибыль в зависимости от цены и издержек.

| № п/п | Цена, у.е. | Объем продаж, шт. | Прибыль (тыс у.е.) при издержках на единицу продукции, у.е. | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1350 | 1400 | 1450 | 1500 | 1550 | 1600 | 1650 | 1700 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1400 | 1600 | 80 | 0 | - | - | - | - | - | - |
| 2 | 1 500 | 1500 | 225 | 150 | 75 | 0 | - | - | - | - |
| 3 | 1 600 | 1200 | 300 | 240 | 180 | 120 | 60 | 0 | - | - |
| 4 | 1 700 | 1000 | 350 | 300 | 250 | 200 | 150 | 100 | 50 | 0 |
| 5 | 1 800 | 720 | 324 | 288 | 252 | 216 | 180 | 144 | 108 | 72 |
| 6 | 1 900 | 500 | 275 | 250 | 225 | 200 | 175 | 150 | 125 | 100 |
| 7 | 2 000 | 320 | 208 | 192 | 176 | 160 | 144 | 128 | 112 | 96 |
| 8 | 2 100 | 170 | 127,5 | 119 | 110,5 | 102 | 93,5 | 85 | 76,5 | 68 |
| 9 | 2 200 | 110 | 93,5 | 88 | 82,5 | 77 | 71,5 | 66 | 60,5 | 55 |

Для удобства рассмотренные результаты оптимальных объемов производства при соответствующих ценах приведены в табл. 9.

Таблица 9.

Оптимальные выпуск и цена в зависимости от издержек

| Издержки | 1350 | 1400 | 1450 | 1500 | 1550 | 1600 | 1650 | 1700 |
|---|---|---|---|---|---|---|---|---|
| Оптимальный выпуск | 1000 | 1000 | 720 | 720 | 720 | 500 | 500 | 500 |
| Цена | 1700 | 1700 | 1800 | 1800 | 1800 | 1900 | 1900 | 1900 |

Как видно из табл. 9, увеличение издержек ведет к снижению оптимального выпуска при росте цены. Хотя изменение издержек на 50 у.е. может не сразу привести к изменению цены. Необоснованная цена может “переключить” большую группу потребителей на другое, аналогичное изделие, имеющее сходный по уровню набор технических характеристик, но более низкую рыночную цену.

По данным функции спроса (табл. 8) проведем расчет эластичности спроса по цене. Под ценовой эластичностью спроса понимается степень реагирования рыночного спроса на изменение цен. В классическом понимании эластичность спроса по цене показывает, насколько изменится объем спроса при изменении цены на 1%. Спрос квалифицируется как эластичный, если понижение цены вызывает такой рост оборота, при котором увеличение объема продаж с лихвой компенсирует более низкие цены. Если же понижение цены, приводя к некоторому увеличению объема продаж, тем не менее не ведет к увеличению оборота или даже уменьшает его, то такой спрос называется неэластичным. Коэффициент ценовой эластичности спроса определяется по формуле:

$$K_{ЦЭС} = \frac{(Q_1 - Q_2) \big/ (Q_1 + Q_2)}{(P_1 - P_2) \big/ (P_1 + P_2)},$$

где $Q_1$, $Q_2$ – значения объема продаж; $P_1$, $P_2$ – значения цены изделия.

В рассматриваемом случае $K_{ЦЭС}$ будет различен на протяжении всей функции спроса (рис. 2). Однако, произведем расчет на той части кривой (в том диапазоне), где присутствует расчетная цена подъемника, а именно: $Q_1$=1200 шт. ; $Q_2$=720 шт.; $P_1$=1600 у.е.; $P_2$=1800у.е. В этом случае

$$K_{ЦЭС} = \frac{(1200 - 720)/(1200 + 720)}{(1600 - 1800)/(1600 + 1800)} = -4{,}25.$$

Коэффициент $K_{ЦЭС}$ имеет отрицательный знак и абсолютную величину, значительно превышающую 1. Это говорит о сильной обратной зависимости объемов продаж от цены. Спрос на подъемник эластичен. Валовая выручка увеличивается при снижении цены и уменьшается при ее повышении. Компании необходимо быть готовой к тому, что покупатели очень чутко реагируют на всякое повышение цены на изделие значительным снижением объемов закупок. Как отмечает А.А. Пивень, снижение же эластичности спроса на изделие возможно только при общем росте благосостояния населения страны и в частности, значительного роста доходной части бюджетов промышленных предприятий.

### 1.3.3. Проверка однородности двух биномиальных выборок

Проверка однородности – одна из базовых проблем прикладной статистики. Она часто обсуждается в литературе, а методы проверки однородности применяются при решении многих практических задач. Например, как сравнить две группы - мужчин и женщин, молодых и пожилых, и т.п.? В маркетинге это важно для сегментации рынка. Если две группы не отличаются по ответам, значит, их можно объединить в один сегмент и проводить по отношению к ним одну и туже маркетинговую политику, в частности, осуществлять одни и те же рекламные воздействия. Если же две группы различаются, то и относиться к ним надо по-разному. Это - представители двух разных сегментов рынка, требующих разного подхода при борьбе за их завоевание.

Обсуждаемая далее постановка задачи в терминах прикладной статистики такова. Рассматривается вопрос с двумя возможными ответами, например, "да" и "нет". В первой группе из $n_1$ опрошенных $m_1$ человек сказали "да", а во второй группе из $n_2$ опрошенных $m_2$ сказали "да". В вероятностной модели предполагается, что $m_1$ и $m_2$ - биномиальные случайные величины $B(n_1 , p_1 )$ и $B(n_2 , p_2 )$ соответственно. (Запись $B(n , p)$ означает, что случайная величина $m$, имеющая биномиальное распределение $B(n , p)$ с параметрами $n$ - объем выборки и $p$ - вероятность определенного ответа (скажем, ответа "да"), может быть представлена в виде $m = X_1 + X_2 +...+X_n$ , где случайные величины $X_1 , X_2 ,...,X_n$ независимы, одинаково распределены, принимают два значения1 и 0, причем $P(X_i = 1) = p, P(X_i = 0)= 1-p, i=1,2,...,n.$)

Однородность двух групп означает, что соответствующие им вероятности равны, неоднородность - что эти вероятности отличаются. В терминах прикладной математической статистики: необходимо проверить гипотезу однородности

$$H_0 : p_1 = p_2$$

при альтернативной гипотезе

$$H_1 : p_1 \neq p_2 .$$

(Иногда представляют интерес односторонние альтернативные гипотезы $H_1^{'} : p_1 > p_2$ и $H_1^{''} : p_1 < p_2 .$)

Оценкой вероятности $p_1$ является частота $p_1^*=m_1/n_1$, а оценкой вероятности $p_2$ является частота $p_2^*=m_2/n_2$ . Даже при совпадении вероятностей $p_1$ и $p_2$ частоты, как правило, различаются. Как говорят, "по чисто случайным причинам". Рассмотрим случайную величину $p_1^* - p_2^*$. Тогда

$$M(p_1^* - p_2^*) = p_1 - p_2 , D(p_1^* - p_2^*) = p_1 (1 - p_1 )/ n_1 + p_2 (1-p_2 )/ n_2 .$$

Из теоремы Муавра-Лапласа и теоремы о наследовании сходимости (глава 1.4 и [4, п.2.4]) следует, что

$$\lim_{n_1\to\infty,n_2\to\infty} P\{\frac{p_1^* - p_2^* - M(p_1^* - p_2^*)}{\sqrt{D(p_1^* - p_2^*)}} \le x\} = \Phi(x),$$

где $\Phi(x)$ - функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Для практического применения этого соотношения следует заменить неизвестную статистику дисперсию разности частот на оценку этой дисперсии:

$$D^*(p_1^* - p_2^*) = p^*_1 (1 - p^*_1 )/ n_1 + p^*_2 (1-p^*_2 )/ n_2 .$$

(Могут использоваться и другие оценки рассматриваемой дисперсии, например, по объединенной выборке). С помощью указанной выше математической техники можно показать,

$$\lim_{n_1\to\infty,n_2\to\infty} P\{\frac{p_1^* - p_2^* - M(p_1^* - p_2^*)}{\sqrt{D^*(p_1^* - p_2^*)}} \le x\} = \Phi(x).$$

что

При справедливости гипотезы однородности $M(p_1^* - p_2^*) = 0$. Поэтому правило принятия решения при проверке однородности двух выборок выглядит так:

1. Вычислить статистику

$$Q = \frac{p_1^* - p_2^*}{\sqrt{\frac{p_1^*(1-p_1^*)}{n_1} + \frac{p_2^*(1-p_2^*)}{n_2}}}.$$

2. Сравнить значение модуля статистика $|Q|$ с граничным значением $K$. Если $|Q| \le K$, то принять гипотезу однородности $H_0$ . Если же $|Q| > K$, то заявить об отсутствии однородности и принять альтернативную гипотезу $H_1$ .

Граничное значение $K$ определяется выбором уровня значимости статистического критерия проверки однородности. Из приведенных выше предельных соотношений следует, что при справедливости гипотезы однородности $H_0$ для уровня значимости $\alpha = P(|Q| > K)$ имеем (при $n_1 \to \infty, n_2 \to \infty$)

$$\alpha \to \Phi(K) - \Phi(-K) = 2\Phi(K) - 1.$$

Следовательно, граничное значение в зависимости от уровня значимости целесообразно выбирать из условия

$$K = K(\alpha) = \Phi^{-1}\left(\frac{1+\alpha}{2}\right).$$

Здесь $\Phi^{-1}(\cdot)$ - функция, обратная к функции стандартного нормального распределения. В социально-экономических исследованиях наиболее распространен 5% уровень значимости, т.е. $\alpha = 0{,}05$. Для него $K = 1{,}96$.

*Пример 2.* Пусть в первой группе из 500 опрошенных ответили "да" 200, а во второй группе из 700 опрошенных сказали "да" 350. Есть ли разница между генеральными совокупностями, представленными этими двумя группами, по доле отвечающих "да"?

Уберем из формулировки примера термин "генеральная совокупность". Получим следующую постановку.

Пусть из 500 опрошенных мужчин ответили "да, я люблю пепси-колу" 200, а из 700 опрошенных женщин 350 сказали "да, я люблю пепси-колу". Есть ли разница между мужчинами и женщинами по доле отвечающих "да" на вопрос о любви к пепси-коле?

В рассматриваемом примере нужные для расчетов величины таковы: $n_1 = 500, p_1^* = 200/500 = 0{,}4; n_2 = 700, p_2^* = 350/700 = 0{,}5$. Вычислим статистику

$$Q = \frac{0{,}4 - 0{,}5}{\sqrt{\frac{0{,}4 \cdot 0{,}6}{500} + \frac{0{,}5 \cdot 05}{700}}} = \frac{-0{,}1}{\sqrt{\frac{0{,}24}{500} + \frac{0{,}25}{700}}} = \frac{-0{,}1}{\sqrt{0{,}00048 + 0{,}0003571}}$$

$$= \frac{-0{,}1}{\sqrt{0{,}0008371}} = \frac{-0{,}1}{0{,}029} = -3{,}45.$$

Поскольку $|Q| = 3{,}45 > 1{,}96$, то необходимо отклонить нулевую гипотезу и принять альтернативную. Таким образом, мужчины и женщины отличаются по рассматриваемому признаку - любви к пепси-коле.

Необходимо отметить, что результат проверки гипотезы однородности зависит не только от частот, но и от объемов выборок. Предположим, что частоты (доли) зафиксированы, а объемы выборок растут. Тогда числитель статистики $Q$ не меняется, а знаменатель уменьшается, значит, вся дробь возрастает. Поскольку знаменатель стремится к 0, то дробь возрастает до бесконечности и рано или поздно превзойдет любую границу. Есть только одно исключение - когда в числителе стоит 0. Следовательно, при строгом подходе к формулировкам вывод статистика должен выглядеть так: "различие обнаружено" или "различие не обнаружено". Во втором случае различие, возможно, было бы обнаружено при увеличении объемов выборок.

Как и для доверительного оценивания вероятности, во ВЦИОМ разработаны две полезные таблицы, позволяющие оценить вызванные чисто случайными причинами допустимые расхождения между частотами в группах. Эти таблицы рассчитаны при выполнении нулевой гипотезы однородности и соответствуют ситуациям, когда частоты близки к 50% (табл.10) или к 20% (табл.11). Если наблюдаемые частоты - от 30% до 70%, то рекомендуется пользоваться первой из этих таблиц, если от 10% до 30% или от 70% до 90% - то второй. Если наблюдаемые

частоты меньше 10% или больше 90%, то теорема Муавра-Лапласа и основанные на ней асимптотические формулы дают не очень хорошие приближения, целесообразно применять иные, более продвинутые математические средства, в частности, приближения с помощью распределения Пуассона.

Таблица 10.

Допустимые расхождения (в %) между частотами в двух группах, когда наблюдаются частоты от 30% до 70%

| Объемы Групп | 750 | 600 | 400 | 200 | 100 |
|---|---|---|---|---|---|
| 750 | 6 | 7 | 7 | 10 | 12 |
| 600 | 7 | 8 | 8 | 11 | 13 |
| 400 | 7 | 8 | 10 | 11 | 14 |
| 200 | 10 | 11 | 11 | 13 | 16 |
| 100 | 12 | 13 | 14 | 16 | 18 |

Таблица 11.

Допустимые расхождения (в %) между частотами в двух группах, когда наблюдаются частоты от 10% до30% или от 70% до 90%

| Объемы Групп | 750 | 600 | 400 | 200 | 100 |
|---|---|---|---|---|---|
| 750 | 5 | 5 | 6 | 8 | 10 |
| 600 | 5 | 6 | 7 | 8 | 10 |
| 400 | 6 | 7 | 8 | 9 | 11 |
| 200 | 8 | 8 | 9 | 10 | 12 |
| 100 | 10 | 10 | 11 | 12 | 14 |

В условиях разобранного выше примера табл.10 дает допустимое расхождение 7%. Действительно, объем первой группы 500 отсутствует в таблице, но строки, соответствующие объемам 400 и 600, совпадают для первых двух столбцов слева. Эти столбцы соответствуют объемам второй группы 750 и 600, между которыми расположен объем 700, данный в примере. Он ближе к 750, поэтому берем величину расхождения, стоящую на пересечении первого столбца и второй (и третьей) строк, т.е. 7%. Поскольку реальное расхождение (10%) больше, чем 7%, то делаем вывод о наличии значимого различия между группами. Естественно, этот вывод совпадает с полученным ранее расчетным путем.

Как и для табл.5, значения в таблицах 10 и 11 несколько больше, чем рассчитанные по приведенным выше формулам. Как и раньше, дело в том, что таблицы ВЦИОМ связаны не с уровнем значимости $\alpha = 0{,}05$, а с уровнем значимости $\alpha = 0{,}01$, которому соответствует граничное значение 2,58.

Допустимое расхождение $\Delta = \Delta(\alpha)$ между частотами нетрудно получить расчетным путем. Для этого достаточно воспользоваться формулой для статистики $Q$ и определить, при каком максимальном расхождении частот все еще делается вывод о том, что верна гипотеза однородности. Следовательно, допустимое расхождение $\Delta = \Delta(\alpha)$ находится из уравнения

$$K(\alpha) = \frac{\Delta(\alpha)}{\sqrt{\frac{p_1^*(1-p_1^*)}{n_1} + \frac{p_2^*(1-p_2^*)}{n_2}}}.$$

Таким образом,

$$\Delta(\alpha) = K(\alpha)\sqrt{\frac{p_1^*(1-p_1^*)}{n_1} + \frac{p_2^*(1-p_2^*)}{n_2}}.$$

Для данных примера 2 $\Delta = \Delta(\alpha) = 1{,}96 \times 0{,}029 = 0{,}057$, или 5,7%, для уровня значимости 0,05.

Для других уровней значимости надо использовать другие коэффициенты $K(\alpha)$. Так, $K(0{,}01) = 2{,}58$ для уровня значимости 1% и $K(0{,}10) = 1{,}64$ для уровня значимости 10%. Для данных примера $\Delta = \Delta(\alpha) = 2{,}58 \times 0{,}029 = = 0{,}7482 \approx 0{,}075$, или 7,5%, для уровня значимости 0,01. Если округлить до ближайшего целого числа процентов, то получим 7%, как при использовании таблицы 7 выше.

Анализ таблиц 10 и 11 показывает, что для констатации различия частоты должны отличаться не менее чем на 6%, а при некоторых объемах выборок - более чем на 10%, при объемах выборок 100 и 100 - на 19%. Если частоты отличаются на 5% или менее, можно сразу сказать, что статистический анализ приведет к выводу о том, что различие не обнаружено (для выборок объемов не более 750).

В связи с этим возникает вопрос: каково типовое отличие частот в двух выборках из одной и той же совокупности? Разность частот в этом случае имеет нулевое математическое ожидание и дисперсию

$$p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right) = \frac{p(1-p)(n_1+n_2)}{n_1 n_2}.$$

Величина $p(1\text{-}p)$ достигает максимума при $p=1/2$, и этот максимум равен 1/4. Если $p=1/2$, а объемы двух выборок совпадают и равны 500, то дисперсия разности частот равна

$$\frac{0{,}25 \times 1000}{500 \times 500} = \frac{250}{250 \times 1000} = \frac{1}{1000}.$$

Следовательно, среднее квадратического отклонение $\sigma$ равно 0,032, или 3,2%. Поскольку для стандартной нормальной случайной величины в 50% случаев ее значение не превосходит по модулю 0,67 (а в 50% случаев - больше 0,67), то типовой разброс равен $0{,}67\sigma$, а в рассматриваемом случае - 2,1%. Приведенные соображения дают возможность построить метод контроля за правильностью проведения повторных опросов. Если частоты излишне устойчивы, значения при повторных опросах слишком близки - это подозрительно! Возможно, нарушены правила проведения опросов, выборки не являются случайными, и т.д.

## Литература

1. Сэндидж Ч., Фрайбургер В., Ротцолл К. Реклама: теория и практика: Пер. с англ. - М.: Прогресс, 1989. - 630 с.
2. Ядов В.А. Стратегии и методы качественного анализа данных. - Журнал "Социология: методология, методы, математические модели", 1991, No.1, с.14-31.
3. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1983. - - 416 с.
4. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука, 1979. - 296 с.
5. Опыт применения ЭВМ в социологических исследованиях. - М.: Институт социологических исследований АН СССР, Советская социологическая ассоциация, 1977. - 158 с.
6. Орлов А.И. Общий взгляд на статистику объектов нечисловой природы. - В сб.: Анализ нечисловой информации в социологических исследованиях (научные редакторы: В.Г. Андреенков, А.И.Орлов, Ю.Н.Толстова). - М.: Наука, 1985. - С.58-92.

## Контрольные вопросы и задачи

1. Почему выборочные исследования необходимы для решения многих практических задач?
2. Рассчитайте коэффициент ценовой эластичности спроса по данным табл.1 при цене $p = 35$ и при цене $p = 40$.
3. Какова роль теоремы Муавра-Лапласа в теории выборочных исследований?

В задачах 4 - 7 выберите наиболее подходящий вариант ответа.

4. При какой цене максимальна прибыль в условиях пункта 1.3.1, если издержки (оптовая цена книги) равны 12?

А) 30; Б) 32; В) 35.

5. При исследовании предпочтений потребителей открытые вопросы: А) труднее для опрашиваемых, но легче для обработки;

Б) легче для опрашиваемых, но труднее для обработки.

6. Пусть из 657 опрошенных 289 сказали «да». Доверительный интервал для доли отвечающих «да» в генеральной совокупности, соответствующий доверительной вероятности 0,95, таков:

А) [0,245; 0,398]; Б) [0,435; 0,445]; В) [0,405; 0,556];

Г) [0,402; 0,478]; Д) [0,247; 0,633] .

7. Из 513 юношей 193 любят «Сникерс», а из 748 девушек – 327. Значение статистического критерия $Q$ для проверки гипотезы о равенстве вероятностей равно:

А) 3,38; Б) -2,176; В) 0,25; Г) 12,56; Д) -0,173.

Гипотеза об одинаковой привлекательности «Сникерса» для юношей и девушек (на уровне значимости 0,05):

А) принимается; Б) отклоняется.

8. Как понятие допустимого расхождения между частотами можно использовать при планировании выборочных исследований?

## Темы докладов, рефератов, исследовательских работ

1. Проведите выборочное исследование с целью построения оценки функции ожидаемого спроса на выбранный Вами товар (услугу).
2. Найдите адекватное приближение функции спроса в табл.1 с помощью метода наименьших квадратов.
3. Постройте экономико-математическую модель оптимизации цены при заданных функциях спроса (в зависимости о цены) и издержек (в зависимости от выпуска).
4. Сопоставьте теорию квотной выборки с теорией простой случайной выборки.
5. Рассмотрите статистическую теорию доверительного оценивания и проверки гипотез о равенстве вероятностей ответов в случае нескольких возможных ответов (с использованием мультиномиального распределения вместо биномиального).
6. В каких случаях может быть использована теория малых выборок (теорема Пуассона) для доверительного оценивания вероятности определенного ответа?

## 1.4. Теоретическая база прикладной статистики

В настоящем разделе собраны основные математико-статистические утверждения, постоянно используемые при математическом обосновании методов прикладной статистики. Эти утверждения отнюдь не всегда легко найти в литературе по теории вероятностей и математической статистике. Например, такие рассматриваемые далее теоремы и методы, как многомерная центральная предельная теорема, теоремы о наследовании сходимости и метод линеаризации, даже не включены в энциклопедию «Вероятность и математическая статистика» [1] – наиболее полный свод знаний по этой тематике. Последний факт наглядно демонстрирует разрыв между математической дисциплиной «теория вероятностей и математическая статистика» и потребностями прикладной статистики.

### 1.4.1. Законы больших чисел

Законы больших чисел позволяют описать поведение сумм случайных величин. Примером является следующий результат, обобщающий полученный ранее в подразделе 1.2.2. Там было доказано следующее утверждение.

*Теорема Чебышёва.* Пусть случайные величины $X_1, X_2, \ldots, X_k$ попарно независимы и существует число $C$ такое, что $D(X_i) \le C$ при всех $i = 1, 2, \ldots, k$. Тогда для любого положительного е выполнено неравенство

$$P\left\{\left|\frac{X_1 + X_2 + \ldots + X_k}{k} - \frac{M(X_1) + M(X_2) + \ldots + M(X_k)}{k}\right| \ge \varepsilon\right\} \le \frac{C}{k\varepsilon^2}. \quad (1)$$

Частным случаем теоремы Чебышева является теорема Бернулли – первый в истории вариант закона больших чисел.

*Теорема Бернулли.* Пусть $m$ – число наступлений события $A$ в $k$ независимых (попарно) испытаниях, и $p$ есть вероятность наступления события $A$ в каждом из испытаний. Тогда при любом $\varepsilon > 0$ справедливо неравенство

$$P\left\{\left|\frac{m}{k} - p\right| \ge \varepsilon\right\} \le \frac{p(1-p)}{k\varepsilon^2}. \quad (2)$$

Ясно, что при росте $k$ выражения в правых частях формул (1) и (2) стремятся к 0. Таким образом, среднее арифметическое попарно независимых случайных величин сближается со средним арифметическим их математических ожиданий.

Напомним, что в разделе 1.2 шла речь лишь о пространствах элементарных событий из конечного числа элементов. Однако приведенные теоремы верны и в общем случае, для произвольных пространств элементарных событий. Однако в условие закона больших чисел необходимо добавить требование существования дисперсий. Легко видеть, что если существуют дисперсии, то существуют и математические ожидания. Закон больших чисел в форме Чебышёва приобретает следующий вид.

*Теорема Чебышева* [2, с.147]. Если $X_1, X_2, \ldots, X_k$ ,… - последовательность попарно независимых случайных величин, имеющих конечные дисперсии, ограниченные одной и той же постоянной,

$$D(X_1) \le C,\ D(X_2) \le C, \ldots\ D(л_i) \le C, \ldots$$

то, каково бы ни было постоянное е > 0,

$$\lim_{k\to\infty} P\left\{\left|\frac{1}{n}\sum_{j=1}^{k} X_j - \frac{1}{n}\sum_{j=1}^{k} MX_j\right| < \varepsilon\right\} = 1. \quad (3)$$

С точки зрения прикладной статистики ограниченность дисперсий вполне естественна. Она вытекает, например, из ограниченности диапазона изменения практически всех величин, используемых при реальных расчетах.

В 1923 г. А.Я. Хинчин показал, что если случайные величины не только независимы, но и одинаково распределены, то существование у них математического ожидания является необходимым и достаточным условием для применимости закона больших чисел [2, с.150].

*Теорема* [2, с.150-151]. Для того чтобы для последовательности $X_1, X_2, ..., X_k$ ,…(как угодно зависимых) случайных величин при любом положительном е выполнялось соотношение (3), необходимо и достаточно, чтобы при $n \to \infty$

$$M \frac{\left( \sum_{j=1}^{k} (X_j - MX_j) \right)^2}{n^2 + \left( \sum_{j=1}^{k} (X_j - MX_j) \right)^2} \to 0.$$

Законы больших чисел для случайных величин служат основой для аналогичных утверждений для случайных элементов в пространствах более сложной природы. В частности, в пространствах произвольной природы (см. подраздел 2.1.5 далее). Однако здесь мы ограничимся классическими формулировками, служащими основой для современной прикладной статистики.

Смысл классических законов больших чисел состоит в том, что выборочное среднее арифметическое независимых одинаково распределенных случайных величин приближается (сходится ) к математическому ожиданию этих величин. Другими словами, *выборочные средние сходятся к теоретическому среднему.*

Это утверждение справедливо и для других видов средних. Например, выборочная медиана сходится к теоретической медиане. Это утверждение – тоже закон больших чисел, но не классический.

Существенным продвижением в теории вероятностей во второй половине XX в. явилось введение средних величин в пространствах произвольной природы и получение для них законов больших чисел, т.е. утверждений, состоящих в том, что эмпирические (т.е. выборочные )средние сходятся к теоретическим средним. Эти результаты будут рассмотрены в подразделе 2.1.5 ниже.

### 1.4.2. Центральные предельные теоремы

В разделе 1.2. уже был приведен простейший вариант Центральной предельной теоремы (ЦПТ) теории вероятностей.

*Центральная предельная теорема* (для одинаково распределенных слагаемых). Пусть $X_1, X_2, ..., X_n, \ldots$– независимые одинаково распределенные случайные величины с математическими ожиданиями $M(X_i) = m$ и дисперсиями $D(X_i) = \sigma^2$, $i = 1, 2, \ldots, n, \ldots$ Тогда для любого действительного числа $x$ существует предел

$$\lim_{n\to\infty} P\left( \frac{X_1 + X_2 + ... + X_n - nm}{\sigma\sqrt{n}} < x \right) = \Phi(x),$$

где $\Phi(x)$ – функция стандартного нормального распределения.

Эту теорему иногда называют теоремой Линдеберга-Леви [3, с.122].

В ряде прикладных задач не выполнено условие одинаковой распределенности. В таких случаях центральная предельная теорема обычно остается справедливой, однако на последовательность случайных величин приходится накладывать те или иные условия. Суть этих условий состоит в том, что ни одно слагаемое не должно быть доминирующим, вклад каждого слагаемого в среднее арифметическое должен быть пренебрежимо мал по сравнению с итоговой суммой. Наиболее часто используется теорема Ляпунова.

*Центральная предельная теорема* (для разнораспределенных слагаемых) – *теорема Ляпунова*. Пусть $X_1, X_2, ..., X_n, \ldots$– независимые случайные величины с математическими ожиданиями $M(X_i) = m_i$ и дисперсиями $D(X_i) = \sigma_i^2 \neq 0$, $i = 1, 2, \ldots, n, \ldots$ Пусть при некотором д>0 у всех рассматриваемых случайных величин существуют центральные моменты порядка 2+д и безгранично убывает «дробь Ляпунова»:

$$\lim_{k\to\infty} \frac{1}{B_n^{2+\delta}} \sum_{k=1}^{n} M \mid X_k - m_k \mid^{2+\delta} = 0,$$

где

$$B_k^2 = \sum_{i=1}^{k} \sigma_i^2 = D\left( \sum_{i=1}^{k} X_i \right).$$

Тогда для любого действительного числа *x* существует предел

$$\lim_{n\to\infty} P\left( \frac{X_1 + X_2 + ... + X_n - m_1 - m_2 - ... - m_n}{B_n} < x \right) = \Phi(x), \quad (1)$$

где *Ф(x)* – функция стандартного нормального распределения.

В случае одинаково распределенных случайных слагаемых

$$m_1 = m_2 = ... = m_n = m, \quad B_n = D(X_1 + X_2 + ... + X_n) = \sigma\sqrt{n},$$

и теорема Ляпунова переходит в теорему Линдеберга-Леви.

История получения центральных предельных теорем для случайных величин растянулась на два века – от первых работ Муавра в 30-х годах 18-го века для необходимых и достаточных условий, полученных Линдебергом и Феллером в 30-х годах 20-го века.

*Теорема Линдеберга-Феллера.* Пусть $X_1, X_2, ..., X_n, \ldots$ – независимые случайные величины с математическими ожиданиями $M(X_i) = m_i$ и дисперсиями $D(X_i) = \sigma_i^2 \neq 0$, $i = 1, 2, \ldots, n, \ldots$ Предельное соотношение (1), т.е. центральная предельная теорема, выполнено тогда и только тогда, когда при любом ф>0

$$\lim_{n\to\infty} \frac{1}{B_n^2} \sum_{k=1}^{n} \int_{|x-m_k|>\tau B_n} (x - m_k)^2 \, dF_k \, (x0 = 0,$$

где $F_k(x)$ обозначает функцию распределения случайной величины $X_k$.

Доказательства перечисленных в настоящем подразделе центральных предельных теорем для случайных величин можно найти в классическом курсе теории вероятностей [2].

Для прикладной статистики большое значение имеет многомерная центральная предельная теорема. В ней речь идет не о сумме случайных величин, а о сумме случайных векторов.

*Необходимое и достаточное условие многомерной сходимости* [3, с.124]. Пусть $F_n$ обозначает совместную функцию распределения *k*-мерного случайного вектора $(X_n^{(1)}, ..., X_n^{(k)})$, $n = 1,2,\ldots$, и $F_{лn}$ – функция распределения линейной комбинации $\lambda_1 X_n^{(1)} + ... + \lambda_л X_n^{(k)}$. Необходимое и достаточное условие для сходимости $F_n$ к некоторой *k*-мерной функции распределения *F* состоит в том, что $F_{лn}$ имеет предел для любого вектора л.

Приведенная теорема ценна тем, что сходимость векторов сводит к сходимости линейных комбинаций их координат, т.е. к сходимости обычных случайных величин, рассмотренных ранее. Однако она не дает возможности непосредственно указать предельное распределение. Это можно сделать с помощью следующей теоремы.

*Теорема о многомерной сходимости.* Пусть $F_n$ и $F_{лn}$ – те же, что в предыдущей теореме. Пусть *F* - совместная функция распределения *k*-мерного случайного вектора $(X_1, ..., X_k)$. Если функция распределения $F_{лn}$ сходится при росте объема выборки к функции распределения $F_л$ для любого вектора л, где $F_л$ – функция распределения линейной комбинации $\lambda_1 X_1 + ... + \lambda_k X_k$, то $F_n$ сходится к *F*.

Здесь сходимость $F_n$ к *F* означает, что для любого *k*-мерного вектора $(x_1, ..., x_k)$ такого, что функция распределения *F* непрерывна в $(x_1, ..., x_k)$, числовая последовательность $F_n(x_1, ..., x_k)$ сходится при росте *n* к числу $F(x_1, ..., x_k)$. Другими словами, сходимость функций распределения понимается ровно также, как при обсуждении предельных теорем для случайных величин выше. Приведем многомерный аналог этих теорем.

*Многомерная центральная предельная теорема* [3]. Рассмотрим независимые одинаково распределенные *k*-мерные случайные вектора

$$U_n^{'} = (U_{1n}, ..., U_{kn}), \quad n = 1,2,...,$$

где штрих обозначает операцию транспонирования вектора. Предположим, что случайные вектора $U_n$ имеют моменты первого и второго порядка, т.е.

$$M(U_n) = м,\ D(U_n) = У,$$

где *м* – вектор математических ожиданий координат случайного вектора, У – его ковариационная матрица. Введем последовательность средних арифметических случайных векторов:

$$\overline{U}_n = (\overline{U}_{1n},...,\overline{U}_{kn}),\ \ n = 1,2,...,\ \ \overline{U}_{in} = \frac{1}{n}\sum_{j=1}^{n} U_{ij}.$$

Тогда случайный вектор $\sqrt{n}(\overline{U}_n - \mu)$ имеет асимптотическое *k*-мерное нормальное распределение $N_k(0,\Sigma)$, т.е. он асимптотически распределен так же, как *k*-мерная нормальная величина с нулевым математическим ожиданием, ковариационной У и плотностью

$$N_k(u \mid 0,\Sigma) = (2\pi)^{-k/2} \mid \Sigma \mid^{-1/2} \exp\{-\frac{1}{2}u'\Sigma^{-1}u\}.$$

Здесь |У| - определитель матрицы У. Другими словами, распределение случайного вектора $\sqrt{n}(\overline{U}_n - \mu)$ сходится к *k*-мерному нормальному распределению с нулевым математическим ожиданием и ковариационной матрицей У.

Напомним, что многомерным нормальным распределением с математическим ожиданием м и ковариационной матрицей У называется распределение, имеющее плотность

$$N_k(u \mid \mu,\Sigma) = (2\pi)^{-k/2} \mid \Sigma \mid^{-1/2} \exp\{-\frac{1}{2}[(u-\mu)'\Sigma^{-1}(u-\mu)]\}.$$

Многомерная центральная предельная теорема показывает, что распределения сумм независимых одинаково распределенных случайных векторов при большом числе слагаемых хорошо приближаются с помощью нормальных распределений, имеющих такие же первые два момента (вектор математических ожиданий координат случайного вектора и его корреляционную матрицу), как и исходные вектора. От одинаковой распределенности можно отказаться, но это потребует некоторого усложнения символики. В целом из теоремы о многомерной сходимости вытекает, что многомерный случай ничем принципиально не отличается от одномерного.

*Пример.* Пусть $X_1$, … $X_n$ ,…– независимые одинаково распределенные случайные величины. Рассмотрим *k*-мерные независимые одинаково распределенные случайные вектора

$$U_n^{'} = (X_n, X_n^2, X_n^3,..., X_n^k),\ \ n = 1,2,...$$

Их математическое ожидание – вектор теоретических начальных моментов, а ковариационная матрица составлена из соответствующих центральных моментов. Тогда $\overline{U}_n$ - вектор выборочных центральных моментов. Многомерная центральная предельная теорема утверждает, что $\overline{U}_n$ имеет асимптотически нормальное распределение. Как вытекает из теорем о наследовании сходимости и о линеаризации (см. ниже), из распределения $\overline{U}_n$ можно вывести распределения различных функций от выборочных начальных моментов. А поскольку центральные моменты выражаются через начальные моменты, то аналогичное утверждение верно и для них.

### 1.4.3. Теоремы о наследовании сходимости

**Суть проблемы наследования сходимости.** Пусть распределения случайных величин $X_n$ при $n \to \infty$ стремятся к распределению случайной величины $X$. При каких функциях $f$ можно утверждать, что распределения случайных величин $f(X_n)$ сходятся к распределению $f(X)$, т.е. наследуется сходимость?

Хорошо известно, что для непрерывных функций $f$ сходимость наследуется [3]. Однако в прикладной статистике используются различные обобщения этого утверждения. Необходимость обобщений связана с тремя обстоятельствами.

1) Статистические данные могут моделироваться не только случайными величинами, но и случайными векторами, случайными множествами, случайными элементами произвольной

природы (т.е. функциями на вероятностном пространстве со значениями в произвольном множестве).

2) Переход к пределу должен рассматриваться не только для случая безграничного возрастания объема выборки, но и в более общих случаях. Например, если в постановке статистической задачи участвуют несколько выборок объемов $n(1)$, $n(2)$, … , $n(k)$, то вполне обычным является предположение о безграничном росте всех этих объемов (что можно описать и как $\min \{n(1), n(2), \dots, n(k)\} \to \infty$).

3) Функция $f$ не обязательно является непрерывной. Она может иметь разрывы. Кроме того, она может зависеть от параметров, по которым происходит переход к пределу. Например, может зависеть от объемов выборок. Например, в главе 3.1 понадобится рассмотреть функцию $f = f(n(1), n(2), \dots, n(k))$.

**Расстояние Прохорова и сходимость по направленному множеству.** Введем необходимые для дальнейшего изложения понятия.

*Расстояние (метрика) Прохорова.* Пусть $C$ – некоторое пространство, $A$ – его подмножество, $d$ – метрика в $C$. Введем понятие е-окрестности множества $A$ в метрике $d$:

$$S(A,\mathrm{e}) = \{x \in C: d(A,x) < \mathrm{e}\}.$$

Таким образом, е-окрестность множества $A$ – это совокупность всех точек пространства $C$, отстоящих от $A$ не более чем на положительное число е. При этом расстояние от точки $x$ до множества $A$ – это точная нижняя грань расстояний от $x$ до точек множества $A$, т.е.

$$d(A,x) = \inf\{d(x,y): y \in A\}.$$

Пусть $P_1$ и $P_2$ – две вероятностные меры на $C$ (т.е. распределения двух случайных элементов со значениями в $C$). Пусть $D_{12}$ – множество чисел $\mathrm{e} > 0$ таких, что

$$P_1(A) \le P_2(S(A,\mathrm{e})+\mathrm{e}$$

для любого замкнутого подмножества $A$ пространства $C$. Пусть $D_{21}$ – множество чисел $\mathrm{e} > 0$ таких, что

$$P_2(A) \le P_1(S(A,\mathrm{e})+\mathrm{e}$$

для любого замкнутого подмножества $A$ пространства $C$. Расстояние Прохорова $L(P_1,P_2)$ между вероятностными мерами (его можно рассматривать и как расстояние между случайными элементами с распределениями $P_1$ и $P_2$ соответственно) вводится формулой

$$L(P_1,P_2) = \max(\inf D_{12}, \inf D_{21}).$$

С помощью метрики Прохорова формализуется понятие сходимости распределений случайных элементов в произвольном пространстве.

Расстояние $L(P_1,P_2)$ введено академиком РАН Юрием Васильевичем Прохоровым в середине XX в. и широко используется в современной теории вероятностей.

*Сходимость по направленному множеству* [4, с.95-96]. Бинарное отношение $\ge$ (упорядочение), заданное на множестве $B$, называется направлением на нем, если $B$ не пусто и

(а) если $m$, $n$ и $p$ – такие элементы множества $B$, что $m \ge n$ и $n \ge p$, то $m \ge p$;

(б) $m \ge m$ для любого $m$ из $B$;

(в) если $m$ и $n$ принадлежат $B$, то найдется элемент $p$ из $B$ такой, что $p \ge m$ и $p \ge n$.

Направленное множество – это пара $(B, \ge)$, где $\ge$ - направление на множестве $B$. Направленностью (или «последовательностью по направленному множеству») называется пара $(f, \ge)$, где $f$ – функция, $\ge$ - направление на ее области определения. Пусть $f$: $B \to Y$, где $Y$ – топологическое пространство. Направленность $(f, \ge)$ сходится в топологическом пространстве $Y$ к точке $y_0$, если для любой окрестности $U$ точки $y_0$ найдется $p$ из $B$ такое, что $f(q) \in U$ при любом $q \ge p$. В таком случае говорят также о сходимости по направленному множеству.

Пусть $B = \{(n(1), n(2), \dots, n(k))\}$ – совокупность векторов, каждый из которых составлен из объемов $k$ выборок. Пусть

$$(n(1), n(2), \dots, n(k)) \ge (n_1(1), n_1(2), \dots, n_1(k))$$

тогда и только тогда, когда $n(i) \ge n_1(i)$ при всех $i = 1, 2, \dots, k$. Тогда $(B, \ge)$ – направленное множество, сходимость по которому эквивалентна сходимости при $\min \{n(1), n(2), \dots, n(k)\} \to \infty$.

Чтобы охватить различные частные случаи, целесообразно предельные теоремы формулировать в терминах сходимости по направленному множеству. Будем писать $B = \{б\}$. Пусть запись $б \to \infty$ обозначает переход к пределу по направленному множеству.

**Формулировка проблемы наследования сходимости.** Пусть случайные элементы $X_б$ со значениями в пространстве $C$ сходятся при $б \to \infty$ к случайному элементу $X$, где через $б \to \infty$ обозначен переход к пределу по направленному множеству. Сходимость случайных элементов означает, что $L(X_б, X) \to 0$ при $б \to \infty$, где $L$ – метрика Прохорова в пространстве $C$.

Пусть $f_б$: $C \to Y$ – некоторые функции. Какие условия надо на них наложить, чтобы из $L(X_б, X) \to 0$ вытекало, что $L_1(f_б(X_б), f_б(X)) \to 0$ при $б \to \infty$, где $L_1$ – метрика Прохорова в пространстве $Y$? Другими словами, какие условия на функции $f_б$: $C \to Y$ гарантируют наследование сходимости?

В работах [5, 6] найдены необходимые и достаточные условия на функции $f_б$: $C \to Y$, гарантирующие наследование сходимости. Описанию этих условий посвящена оставшаяся часть подраздела.

Приведем для полноты изложения строгие формулировки математических предположений (в дальнейшем никому, кроме профессиональных математиков, не понадобятся)

*Математические предположения.* Пусть $C$ и $У$ – полные сепарабельные метрические пространства, Пусть выполнены обычные предположения измеримости: $X_б$ и $X$ – случайные элементы $C$, $f_б(X_б)$ и $f_б(X)$ – случайные элементы в $У$, рассматриваемые ниже подмножества пространств $C$ и $У$ лежат в соответствующих у–алгебрах измеримых подмножеств, и т.д.

Понадобятся некоторые *определения*. Разбиение $T_n = \{C_{1n}, C_{2n}, \dots, C_{nn}\}$ пространства $C$ – это такой набор подмножеств $C_j$, $j = 1, 2, \dots, n$, этого пространства, что пересечение любых двух из них пусто, а объединение совпадает с $C$. Диаметром $diam(A)$ подмножества $A$ множества $C$ называется точная верхняя грань расстояний между элементами $A$, т.е.

$$diam(A) = \sup \{d(x,y), x \in A, y \in A\},$$

где $d(x,y)$ – метрика в пространстве $C$. Обозначим $\partial A$ границу множества $A$, т.е. совокупность точек $x$ таких, что любая их окрестность $U(x)$ имеет непустое пересечение как с $A$, так и с $C \backslash A$. Колебанием $д(f, B)$ функции $f$ на множестве $B$ называется $д(f, B) = \sup \{|f(x) - f(y)|, x \in B, y \in B\}$.

**Достаточное условие для наследования сходимости.** Пусть $L(X_б, X) \to 0$ при $б \to \infty$. Пусть существует последовательность $T_n$ разбиений пространства $C$ такая, что $P(X \in \partial A) = 0$ для любого $A$ из $T_n$ и, основное условие, для любого $е > 0$

$$m_\varepsilon(\alpha, n) = \sum P(X \in A) \to 0 \qquad (1)$$

при $n \to \infty$ и $б \to \infty$, где сумма берется по всем тем $A$ из $T_n$, для которых колебание функции $f_б$ на $A$ больше е, т.е. $д(f_б, A) > е$. Тогда $L_1(f_б(X_б), f_б(X)) \to 0$ при $б \to \infty$.

**Необходимое условие для наследования сходимости.** Пусть $У$ – конечномерное линейное пространство, $У = R^k$. Пусть случайные элементы $f_б(X)$ асимптотически ограничены по вероятности при $б \to \infty$, т.е. для любого $е > 0$ существуют число $S(е)$ и элемент направленного множества $б(е)$ такие, что $P(\|f_б(X)\| > S(е)) < е$ при $б \geq б(е)$, где $\|f_б(X)\|$ - норма (длина) вектора $f_б(X)$. Пусть существует последовательность $T_n$ разбиений пространства $C$ такая, что

$$\lim_{n \to \infty} \max \{diam(C_{jn}),\ C_{jn} \in T_n\} = 0,$$

т.е. последовательность $T_n$ является безгранично измельчающейся. Самое существенное – пусть условие (1) не выполнено для последовательности $T_n$. Тогда существует последовательность случайных элементов $X_б$ такая, что $L(X_б, X) \to 0$ при $б \to \infty$, но $L_1(f_б(X_б), f_б(X))$ не сходится к 0 при $б \to \infty$.

Несколько огрубляя, можно сказать, что *условие (1) является необходимым и достаточным для наследования сходимости*.

*Пример 1.* Пусть $C$ и $У$ – конечномерные линейные пространства, функции $f_б$ не зависят от б, т.е. $f_б \equiv f$, причем функция $f$ ограничена. Тогда условие (1) эквивалентно требованию интегрируемости по Риману-Стилтьесу функции $f$ по мере $G(A) = P(X \in A)$. В частности, условие (1) выполнено для непрерывной функции $f$.

В конечномерных пространствах $C$ вместо сходимости $L(X_б, X) \to 0$ при $б \to \infty$ можно говорить о слабой сходимости функций распределения случайных векторов $X_б$ к функции

распределения случайного вектора $X$. Речь идет о «сходимости по распределению», т.е. о сходимости во всех точках непрерывности функции распределения случайного вектора $X$. В этом случае разбиения могут состоять из многомерных параллелепипедов [5, гл.2].

*Пример 2.* Полученные выше результаты дают обоснование для рассуждений типа следующего (ср., например, утверждения в главе 3.1 ниже). Пусть по двум независимым выборкам объемов $m$ и $n$ соответственно построены статистики $X_m$ и $Y_n$. Пусть известно, что распределения этих статистик сходятся при безграничном росте объемов выборок к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1. Пусть $a(m, n)$ и $b(m, n)$ – некоторые коэффициенты. Тогда согласно результатам примера 1 распределение случайной величины $Z(m, n) = a(m, n)X_m + b(m, n)Y_n$ сближается с распределением нормально распределенной случайной величины с математическим ожиданием 0 и дисперсией $a^2(m, n) + b^2(m, n)$. Если же $a^2(m, n) + b^2(m, n) = 1$, например,

$$a(m,n) = \sqrt{\frac{m}{m+n}}. \quad b(m,n) = \sqrt{\frac{n}{m+n}},$$

то распределение $Z(m, n)$ сходится при безграничном росте объемов выборок к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1.

### 1.4.4. Метод линеаризации

При разработке методов прикладной статистики часто возникает следующая задача [3, с.338]. Имеется последовательность $k$-мерных случайных векторов $X_n = (X_{1n}, X_{2n}, \dots, X_{kn})$, $n = 1, 2, \dots$, такая, что $X_n \to a = (a_1, a_2, \dots, a_k)$ при $n \to \infty$, и последовательность функций $f_n$: $R^k \to R^1$. Требуется найти распределение случайной величины $f_n(X_n)$.

Основная идея – рассмотреть главный линейный член функции $f_n$ в окрестности точки $a$. Из математического анализа известно, что

$$f_n(X_n) - f_n(a) = \sum_{j=1}^{k} \frac{\partial f_n(a)}{\partial x_j}(X_{jn} - a_j) + O_n(\| X_n - a \|^2),$$

где остаточный член является бесконечно малой величиной более высокого порядка малости, чем линейный член. Таким образом, произвольная функция может быть заменена на линейную функцию от координат случайного вектора. Эта замена проводится с точностью до бесконечно малых более высокого порядка. Конечно, должны быть выполнены некоторые математические условия регулярности. Например, функции $f_n$ должны быть дважды непрерывно дифференцируемы в окрестности точки $a$.

Если вектор $X_n$ является асимптотически нормальным с математическим ожиданием $a$ и ковариационной матрицей $\sum/n$, где $\sum = \|y_{ij}\|$, причем $y_{ij} = nM(X_i - a_i)(X_j - a_j)$, то линейная функция от его координат также асимптотически нормальна. Следовательно, при очевидных условиях регулярности $f_n(X_n)$ – асимптотически нормальная случайная величина с математическим ожиданием $f_n(a)$ и дисперсией

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{\partial f_n(a)}{\partial x_i}\frac{\partial f_n(a)}{\partial x_j}\sigma_{ij}.$$

Для практического использования асимптотической нормальности $f_n(X_n)$ остается заменить неизвестные моменты $a$ и $\sum$ на их оценки. Например, если $X_n$ – это среднее арифметическое независимых одинаково распределенных случайных векторов, то $a$ можно заменить на $X_n$, а $\sum$ - на выборочную ковариационную матрицу.

*Пример.* Пусть $Y_1, Y_2, \dots, Y_n$ – независимые одинаково распределенные случайные величины с математическим ожиданием $a$ и дисперсией $y^2$. В качестве $X_n$ ($k = 1$) рассмотрим выборочное среднее арифметическое

$$\overline{Y} = \frac{Y_1 + Y_2 + \dots + Y_n}{n}.$$

Как известно, в силу закона больших чисел $\overline{Y} \to a = M(Y)$. Следовательно, для получения распределений функций от выборочного среднего арифметического можно использовать метод линеаризации. В качестве примера рассмотрим $f_n(y) = f(y) = y^2$. Тогда

$$\left(\overline{Y}\right)^2 - a^2 = \frac{df(a)}{dy}\left(\overline{Y} - a\right) + O\left(\left(\overline{Y} - a\right)^2\right) = 2a\left(\overline{Y} - a\right) + O\left(\left(\overline{Y} - a\right)^2\right).$$

Из этого соотношения следует, что с точностью до бесконечно малых более высокого порядка

$$\left(\overline{Y}\right)^2 = a^2 + 2a\left(\overline{Y} - a\right).$$

Поскольку в соответствии с Центральной Предельной Теоремой выборочное среднее арифметическое является асимптотически нормальной случайной величиной с математическим ожиданием $a$ и дисперсией $\sigma^2/n$, то квадрат этой статистики является асимптотически нормальной случайной величиной с математическим ожиданием $a^2$ и дисперсией $4a^2\sigma^2/n$. Для практического использования может оказаться полезной замена параметров (асимптотического нормального распределения) на их оценки, а именно, математического ожидания – на $\left(\overline{Y}\right)^2$, а дисперсии – на $4\left(\overline{Y}\right)^2 s^2 / n$, где $s^2$ – выборочная дисперсия.

Большое внимание (целая глава!) уделено методу линеаризации в классическом учебнике Е.С. Вентцель [7].

### 1.4.5. Принцип инвариантности

Пусть $Y_1, Y_2, \ldots, Y_n$ – независимые одинаково распределенные случайные величины с непрерывной функцией распределения $F(x)$. Многие используемые в прикладной статистике функции от результатов наблюдений выражаются через эмпирическую функцию распределения $F_n(x)$. К ним относятся статистики Колмогорова, Смирнова, омега-квадрат. Отметим, что и другие статистики выражаются через эмпирическую функцию распределения, например:

$$\overline{Y} = \int_{-\infty}^{+\infty} x dF_n(x).$$

Полезным является преобразование Н.В.Смирнова $t = F(x)$. Тогда независимые случайные величины $Z_j = F(Y_j)$, $j = 1, 2, \ldots, n$, имеют равномерное распределение на отрезке [0; 1]. Рассмотрим построенную по ним эмпирическую функцию распределения $F_n(t)$, $0 \le t \le 1$. *Эмпирическим процессом* называется случайный процесс

$$\xi_n(t) = \sqrt{n}\left(F_n(t) - t\right).$$

Рассмотрим критерии проверки согласия функции распределения выборки с фиксированной функцией распределения $F(x)$. Статистика критерия Колмогорова записывается в виде

$$K_n = \sup_{0 \le t \le 1} |\xi_n(t)|,$$

статистика критерия Смирнова – это

$$S_n = \sup_{0 \le t \le 1} \xi_n(t),$$

а статистика критерия омега-квадрат (Крамера-Мизеса-Смирнова) имеет вид

$$\omega_n^2 = \int_0^1 \xi_n^2(t) dt.$$

Случайный процесс $\omega_n(t)$ имеет нулевое математическое ожидание и ковариационную функцию $M\omega_n(s)\omega_n(t) = \min(s,t) - st$. Рассмотрим гауссовский случайный процесс $\omega(t)$ с такими же математическим ожиданием и ковариационной функцией. Он называется броуновским мостом. (Напомним, что гауссовским процесс именуется потому, что вектор $(\omega(t_1), \omega(t_2), \ldots, \omega(t_k))$ имеет многомерное нормальное распределение при любых наборах моментов времени $t_1, t_2, \ldots, t_k$.)

Пусть $f$ – функционал, определенный на множестве возможных траекторий случайных процессов. *Принцип инвариантности* [1] состоит в том, что последовательность распределений случайных величин $f(\omega_n)$ сходится при $n \to \infty$ к распределению случайной величины $f(\omega)$.

Сходимость по распределению обозначим символом =>. Тогда принцип инвариантности кратко записывается так: $f(o_n)$ => $f(o)$. В частности, согласно принципу инвариантности статистика Колмогорова и статистика омега квадрат сходятся по распределению к распределениям соответствующих функционалов от случайного процесса о:

$$K_n = \sup_{0\le t\le 1} |\xi_n(t)| \Rightarrow \sup_{0\le t\le 1} |\xi(t)|,\ \omega_n^2 = \int_0^1 \xi_n^2(t)dt \Rightarrow \int_0^1 \xi^2(t)dt .$$

Таким образом, от проблем прикладной статистики сделан переход к теории случайных процессов. Методами этой теории найдены распределения случайных величин

$$\sup_{0\le t\le 1} |\xi(t)|,\ \int_0^1 \xi^2(t)dt .$$

Принцип инвариантности – инструмент получения предельных распределений функций от результатов наблюдений, используемых в прикладной статистике.

Обоснование принципу инвариантности может быть дано на основе теории сходимости вероятностных мер в функциональных пространствах [8]. Более простой подход, позволяющий к тому же получать необходимые и достаточные условия в предельной теории статистик интегрального типа (принцип инвариантности к ним нельзя применить), рассмотрен в главе 2.3.

Почему «принцип инвариантности» так назван? Обратим внимание, что предельные распределения рассматриваемых статистик не зависят от их функции распределения $F(x)$. Другими словами, предельное распределение инвариантно относительно выбора $F(x)$.

В более широком смысле термин «принцип инвариантности» применяют тогда, когда предельное распределение не зависит от тех или иных характеристик исходных распределений [1]. В этом смысле наиболее известный «принцип инвариантности» - это Центральная Предельная Теорема, поскольку предельное стандартное нормальное распределение – одно и то же для всех возможных распределений независимых одинаково распределенных слагаемых (лишь бы слагаемые имели конечные математическое ожидание и дисперсию).

### 1.4.6. Нечеткие множества как проекции случайных множеств

**Нечеткость и случайность.** С самого начала появления современной теории нечеткости в 1960-е годы (см. главу 1.1) началось обсуждение ее взаимоотношений с теорией вероятностей. Дело в том, что функция принадлежности нечеткого множества напоминает распределение вероятностей. Отличие только в том, что сумма вероятностей по всем возможным значениям случайной величины (или интеграл, если множество возможных значений несчетно) всегда равна 1, а сумма $S$ значений функции принадлежности (в непрерывном случае - интеграл от функции принадлежности) может быть любым неотрицательным числом. Возникает искушение пронормировать функцию принадлежности, т.е. разделить все ее значения на $S$ (при $S \neq 0$), чтобы свести ее к распределению вероятностей (или к плотности вероятности). Однако специалисты по нечеткости справедливо возражают против такого "примитивного" сведения, поскольку оно проводится отдельно для каждой размытости (нечеткого множества), и определения обычных операций над нечеткими множествами с ним согласовать нельзя. Последнее утверждение означает следующее. Пусть указанным образом преобразованы функции принадлежности нечетких множеств $A$ и $B$. Как при этом преобразуются функции принадлежности $A \cap B, A \cup B, A + B, AB$? Установить это *невозможно в принципе.* Последнее утверждение становится совершенно ясным после рассмотрения нескольких примеров пар нечетких множеств с одними и теми же суммами значений функций принадлежности, но различными результатами теоретико-множественных операций над ними. Причем и суммы значений соответствующих функций принадлежности для этих результатов теоретико-множественных операций, например, для пересечений множеств, также различны.

В работах по нечетким множествам время от времени утверждается, что теория нечеткости является самостоятельным разделом прикладной математики и не имеет отношения к теории вероятностей (см., например, обзор литературы в монографиях [5,9]). Некоторые авторы, сравнивавшие теорию нечеткости и теорию вероятностей, подчеркивали различие между этими

областями теоретических и прикладных исследований. Обычно сравнивают аксиоматику и сравнивают области приложений. Надо сразу отметить, что аргументы при втором типе сравнений не имеют доказательной силы, поскольку по поводу границ применимости даже такой давно выделившейся научной области, как вероятностно-статистические методы, имеются различные мнения. Напомним, что итог рассуждений одного из наиболее известных французских математиков Анри Лебега по поводу границ применимости арифметики таков: "Арифметика применима тогда, когда она применима" (см. его монографию [10, с.21-22]).

При сравнении различных аксиоматик теории нечеткости и теории вероятностей нетрудно увидеть, что списки аксиом различаются. Из этого, однако, отнюдь не следует, что между указанными теориями нельзя установить связь, типа известного сведения евклидовой геометрии на плоскости к арифметике (точнее к теории числовой системы $R^2$ - см., например, монографию [11]). Напомним, что эти две аксиоматики - евклидовой геометрии и арифметики - на первый взгляд весьма сильно различаются.

Можно понять желание энтузиастов теории нечеткости подчеркнуть принципиальную новизну своего научного аппарата. Однако не менее важно установить связи этого подхода с ранее известными.

**Проекция случайного множества**. Как оказалось, теория нечетких множеств тесно связана с теорией случайных множеств. Еще в 1975 г. в работе [12] было показано, что нечеткие множества естественно рассматривать как "проекции" случайных множеств. Рассмотрим этот метод сведения теории нечетких множеств к теории случайных множеств.

*Определение 1.* Пусть $A = A(\omega)$ - случайное подмножество конечного множества *У.* Нечеткое множество *В,* определенное на *У,* называется проекцией *А* и обозначается *Proj A,* если

$$\mu_B(y) = P(y \in A) \qquad (1)$$

при всех $y \in Y$.

Очевидно, каждому случайному множеству *А* можно поставить в соответствие с помощью формулы (1) нечеткое множество *B = Proj A*. Оказывается, верно и обратное.

*Теорема 1***.** Для любого нечеткого подмножества *В* конечного множества *У* существует случайное подмножество *А* множества *У* такое, что *B = Proj A*.

*Доказательство.* Достаточно задать распределение случайного множества *А*. Пусть $Y_1$ - носитель *В* (см. определение 1 в подразделе 1.1.4 выше). Без ограничения общности можно считать, что $Y_1 = \{y_1, y_2, ..., y_m\}$ при некотором *m* и элементы $Y_1$ занумерованы в таком порядке, что

$$0 < \mu_B(y_1) \le \mu_B(y_2) \le ... \le \mu_B(y_m).$$

Введем множества

$$Y(1) = Y_1, Y(2) = \{y_2, ..., y_m\}, ..., Y(t) = \{y_t, ..., y_m\}, ..., Y(m) = \{y_m\}.$$

Положим

$$P(A = Y(1)) = \mu_B(y_1), \quad P(A = Y(2)) = \mu_B(y_2) - \mu_B(y_1), ...,$$

$$P(A = Y(t)) = \mu_B(y_t) - \mu_B(y_{t-1}), ..., P(A = Y(m)) = \mu_B(y_m) - \mu_B(y_{m-1}),$$

$$P(A = \varnothing) = 1 - \mu_B(y_m).$$

Для всех остальных подмножеств *X* множества *У* положим *P(A=X)*=0. Поскольку элемент $y_t$ входит во множества *Y*(1), *Y*(2),…, *Y*(t) и не входит во множества *Y*(*t*+1),..., *Y*(*m*), то из приведенных выше формул следует, что $P(y_t \in A) = \mu_B(y_t)$. Если $y \notin Y_1$, то, очевидно, $P(y \in A) = 0$. Теорема 1 доказана.

Распределение случайного множества с независимыми элементами, как следует из рассмотрений главы 8 монографии [13], полностью определяется его проекцией. Для конечного случайного множества общего вида это не так. Для уточнения сказанного понадобится следующая теорема.

*Теорема 2.* Для случайного подмножества *А* множества *У* из конечного числа элементов наборы чисел $P(A = X), X \subseteq Y,$ и $P(X \subseteq A), X \subseteq Y,$ выражаются один через другой.

*Доказательство.* Второй набор выражается через первый следующим образом:

$$P(X \subseteq A) = \sum_{X' : X \subseteq X'} P(A = X').$$

Элементы первого набора выразить через второй можно с помощью формулы включений и исключений из формальной логики, в соответствии с которой

$P(A = X) = P(X \subseteq A) - \sum P(X \cup \{y\} \subseteq A) + \sum P(X \cup \{y_1, y_2\} \subseteq A) - \ldots \pm P(Y \subseteq A)$. В этой формуле в первой сумме $y$ пробегает все элементы множества $Y \backslash X$, во второй сумме переменные суммирования $y_1$ и $y_2$ не совпадают и также пробегают это множество, и т.д. Ссылка на формулу включений и исключений завершает доказательство теоремы 2.

В соответствии с теоремой 2 случайное множество $A$ можно характеризовать не только распределением, но и набором чисел $P(X \subseteq A), X \subseteq Y$. В этом наборе $P(\varnothing \subseteq A) = 1$, а других связей типа равенств нет. В этот набор входят числа $P(\{y\} \subseteq A) = P(y \in A)$, следовательно, фиксация проекции случайного множества эквивалентна фиксации $k = Card(Y)$ параметров из ($2^k$-1) параметров, задающих распределение случайного множества $A$ в общем случае.

Будет полезна следующая теорема.

*Теорема 3.* Если $Proj\ A = B$, то $\Pr oj\, \overline{A} = \overline{B}$.

Для доказательства достаточно воспользоваться тождеством из теории случайных множеств $P(\overline{A} = X) = P(A = \overline{X})$, формулой для вероятности накрытия $P(y \in A)$, определением отрицания нечеткого множества и тем, что сумма всех $P(A=X)$ равна 1. При этом под формулой для вероятности накрытия имеется в виду следующее утверждение: чтобы найти вероятность накрытия фиксированного элемента $q$ случайным подмножеством $S$ конечного множества $Q$, достаточно вычислить

$$P(q \in S) = \mathrm{P}(\{\omega : q \in S(\omega)\}) = \sum_{A : q \in A, A \subseteq 2^Q} P(S = A),$$

где суммирование идет по всем подмножествам $A$ множества $Q$, содержащим $q$.

**Пересечения и произведения нечетких и случайных множеств.** Выясним, как операции над случайными множествами соотносятся с операциями над их проекциями. В силу законов де Моргана (теорема 1 в подразделе 1.1.4) и теоремы 3 достаточно рассмотреть операцию пересечения случайных множеств.

*Теорема 4.* Если случайные подмножества $A_1$ и $A_2$ конечного множества $У$ независимы, то нечеткое множество $\Pr oj(A_1 \cap A_2)$ является произведением нечетких множеств $Proj\ A_1$ и $Proj\ A_2$.

*Доказательство.* Надо показать, что для любого $y \in Y$

$$P(y \in A_1 \cap A_2) = P(y \in A_1) P(y \in A_2). \qquad (2)$$

По формуле для вероятности накрытия точки случайным множеством (см. выше)

$$P(y \in A_1 \cap A_2) = \sum_{X : y \in X} P((A_1 \cap A_2) = X). \qquad (3)$$

Легко проверить, что распределение пересечения случайных множеств $A_1 \cap A_2$ можно выразить через их совместное распределение следующим образом:

$$P(A_1 \cap A_2 = X) = \sum_{X_1, X_2 : X_1 \cap X_2 = X} P(A_1 = X_1, A_2 = X_2). \qquad (4)$$

Из соотношений (3) и (4) следует, что вероятность накрытия для пересечения случайных множеств можно представить в виде двойной суммы

$$P(y \in A_1 \cap A_2) = \sum_{X : y \in X} \sum_{X_1, X_2 : X_1 \cap X_2 = X} P(A_1 = X_1, A_2 = X_2). \qquad (5)$$

Заметим теперь, что правую часть формулы (5) можно переписать следующим образом:

$$\sum_{X_1, X_2 : e \in X_1, e \in X_2} P(A_1 = X_1, A_2 = X_2). \qquad (6)$$

Действительно, формула (5) отличается от формулы (6) лишь тем, что в ней сгруппированы члены, в которых пересечение переменных суммирования $X_1 \cap X_2$ принимает постоянное

значение. Воспользовавшись определением независимости случайных множеств и правилом перемножения сумм, получаем, что из (5) и (6) вытекает равенство

$$P(y \in A_1 \cap A_2) = \left( \sum_{X_1 : y \in X_1} P(A_1 = X_1) \right) \left( \sum_{X_2 : y \in X_2} P(A_2 = X_2) \right).$$

Для завершения доказательства теоремы 4 достаточно еще раз сослаться на формулу для вероятности накрытия точки случайным множеством.

*Определение 2***.** Носителем случайного множества $C$ называется совокупность всех тех элементов $y \in Y,$ для которых $P(y \in C) > 0.$

*Теорема 5.* Равенство

$$\Pr oj(A_1 \cap A_2) = (\Pr ojA_1) \cap (\Pr ojA_2)$$

верно тогда и только тогда, когда пересечение носителей случайных множеств $\overline{A_1} \cap A_2$ и $A_1 \cap \overline{A_2}$ пусто.

*Доказательство.* Необходимо выяснить условия, при которых

$$P(y \in A_1 \cap A_2) = \min(P(y \in A_1), P(y \in A_2)). \qquad (7)$$

Положим

$$p_1 = P(y \in A_1 \cap A_2), p_2 = P(y \in \overline{A_1} \cap A_2), p_3 = P(y \in A_1 \cap \overline{A_2}).$$

Тогда равенство (7) сводится к условию

$$p_1 = \min(p_1 + p_2, p_1 + p_3). \quad (8)$$

Ясно, что соотношение (8) выполнено тогда и только тогда, когда $p_2 p_3 = 0$ при всех $y \in Y,$ т.е. не существует ни одного элемента $y_0 \in Y$ такого, что одновременно $P(y_0 \in \overline{A}_1 \cap A_2) > 0$ и $P(y_0 \in A_1 \cap \overline{A_2}) > 0$, а это эквивалентно пустоте пересечения носителей случайных множеств $\overline{A_1} \cap A_2$ и $A_1 \cap \overline{A_2}$. Теорема 5 доказана.

**Сведение последовательности операций над нечеткими множествами к последовательности операций над случайными множествами.** Выше получены некоторые связи между нечеткими и случайными множествами. Стоит отметить, что изучение этих связей в работе [12] началось с введения случайных множеств с целью развития и обобщения аппарата нечетких множеств Л. Заде. Дело в том, что математический аппарат нечетких множеств не позволяет в должной мере учитывать различные варианты зависимости между понятиями (объектами), моделируемыми с его помощью, не является достаточно гибким. Так, для описания "общей части" двух нечетких множеств есть лишь две операции - произведение и пересечение. Если применяется первая из них, то фактически предполагается, что множества ведут себя как проекции независимых случайных множеств (см. выше теорему 4). Операция пересечения также накладывает вполне определенные ограничения на вид зависимости между множествами (см. выше теорему 5), причем в этом случае найдены даже необходимые и достаточные условия. Желательно иметь более широкие возможности для моделирования зависимости между множествами (понятиями, объектами). Использование математического аппарата случайных множеств предоставляет такие возможности.

Цель сведения теории нечетких множеств к теории случайных множеств состоит в том, чтобы за любой конструкцией из нечетких множеств увидеть конструкцию из случайных множеств, определяющую свойства первой, аналогично тому, как за плотностью распределения вероятностей мы видим случайную величину. Рассмотрим результаты по сведению алгебры нечетких множеств к алгебре случайных множеств.

*Определение 3.* Вероятностное пространство $\{\Omega, G, P\}$ назовем делимым, если для любого измеримого множества $X \in G$ и любого положительного числа $\alpha$, меньшего $P(X)$, можно указать измеримое множество $Y \subset X$ такое, что $P(Y) = \alpha.$

*Пример.* Пусть $\Omega$ - единичный куб конечномерного линейного пространства, $G$ есть сигма-алгебра борелевских множеств, а $P$ - мера Лебега. Тогда $\{\Omega, G, P\}$ - делимое вероятностное пространство.

Таким образом, делимое вероятностное пространство - это не экзотика. Обычный куб является примером такого пространства.

Доказательство сформулированного в примере утверждения проводится стандартными математическими приемами. Они основаны на том, что измеримое множество можно сколь угодно точно приблизить открытыми множествами, последние представляются в виде суммы не более чем счетного числа открытых шаров, а для шаров делимость проверяется непосредственно (от шара $X$ тело объема $\alpha < P(X)$ отделяется соответствующей плоскостью).

*Теорема 6.* Пусть даны случайное множество $A$ на делимом вероятностном пространстве $\{\Omega, G, P\}$ со значениями во множестве всех подмножеств множества $U$ из конечного числа элементов, и нечеткое множество $D$ на $U$. Тогда существуют случайные множества $C_1$, $C_2$, $C_3$, $C_4$ на том же вероятностном пространстве такие, что

$$\Pr oj(A \cap C_1) = B \cap D, \quad \Pr oj(A \cap C_2) = BD, \quad \Pr oj(A \cup C_3) = B \cup D,$$

$$\Pr oj(A \cup C_4) = B + D, \quad \Pr ojC_i = D, \ i = 1,2,3,4,$$

где $B = Proj\ A$.

*Доказательство.* В силу справедливости законов де Моргана для нечетких (см. теорему 1 в подразделе 1.1.4 выше) и для случайных множеств, а также теоремы 3 выше (об отрицаниях) достаточно доказать существование случайных множеств $C_1$ и $C_2$.

Рассмотрим распределение вероятностей во множестве всех подмножеств множества $U$, соответствующее случайному множеству $C$ такому, что $Proj\ C = D$ (оно существует в силу теоремы 1). Построим случайное множество $C_2$ с указанным распределением, независимое от $A$. Тогда $\Pr oj(A \cap C_2) = BD$ по теореме 4.

Перейдем к построению случайного множества $C_1$. По теореме 7 необходимо и достаточно определить случайное множество $C_1(\omega)$ так, чтобы $Proj\ C_1 = D$ и пересечение носителей случайных множеств $A \cap \overline{C_1}$ и $\overline{A} \cap C_1$ было пусто, т.е.

$$p_3 = P(y \in A \cap \overline{C_1}) = 0$$

для $y \in Y_1 = \{y : \mu_B(y) \le \mu_D(y)\}$ и

$$p_2 = P(y \in \overline{A} \cap C_1) = 0$$

для $y \in Y_2 = \{y : \mu_B(y) \ge \mu_D(y)\}$.

Построим $C_1(\omega)$, исходя из заданного случайного множества $A(\omega)$. Пусть $y_1 \in Y_2$. Исключим элемент $y_1$ из $A(\omega)$ для стольких элементарных событий $\omega$, чтобы для полученного случайного множества $A_1(\omega)$ было справедливо равенство

$$P(y_1 \in A_1) = \mu_D(y_1)$$

(именно здесь используется делимость вероятностного пространства, на котором задано случайное множество $A(\omega)$). Для $y \ne y_1$, очевидно,

$$P(y \in A_1) = P(y \in A).$$

Аналогичным образом последовательно исключаем $y$ из $A(\omega)$ для всех $y \in Y_2$ и добавляем $y$ в $A(\omega)$ для всех $y \in Y_1$, меняя на каждом шагу $P(y \in A_i)$ только для $y = y_i$ так, чтобы

$$P(y_i \in A_i) = \mu_D(y_i)$$

(ясно, что при рассмотрении $y_i \in Y_1 \cap Y_2$ случайное множество $A_i(\omega)$ не меняется). Перебрав все элементы $U$, получим случайное множество $A_k(\omega) = C_1(\omega)$, для которого выполнено требуемое. Теорема 6 доказана.

Основной результат о сведении теории нечетких множеств к теории случайных множеств дается следующей теоремой.

*Теорема 7.* Пусть $B_1, B_2, B_3, ..., B_t$ - некоторые нечеткие подмножества множества $U$ из конечного числа элементов. Рассмотрим результаты последовательного выполнения теоретико-множественных операций

$$B^m = ((...((B_1 \circ B_2) \circ B_3) \circ ...) \circ B_{m-1}) \circ B_m, \quad m = 1,2,...,t,$$

где $\circ$ - символ одной из следующих теоретико-множественных операций над нечеткими множествами: пересечение, произведение, объединение, сумма (на разных местах могут стоять разные символы). Тогда существуют случайные подмножества $A_1, A_2, A_3, ..., A_t$ того же множества $U$ такие, что

$$\Pr oj A_i = B_i, \ i = 1,2,...,t,$$

и, кроме того, результаты теоретико-множественных операций связаны аналогичными соотношениями

$$\Pr oj\{((...((A_1 \otimes A_2) \otimes A_3) \otimes ...) \otimes A_{m-1}) \otimes A_m\} = B^m, \ m = 1,2,...,t,$$

где знак $\otimes$ означает, что на рассматриваемом месте стоит символ пересечения $\cap$ случайных множеств, если в определении $B^m$ стоит символ пересечения или символ произведения нечетких множеств, и соответственно символ объединения $\cup$ случайных множеств, если в $B^m$ стоит символ объединения или символ суммы нечетких множеств.

*Комментарий.* Поясним содержание теоремы. Например, если

$$B^5 = (((B_1 + B_2) \cap B_3)B_4) \cup B_5,$$

то

$$(((A_1 \otimes A_2) \otimes A_3) \otimes A_4) \otimes A_5 = (((A_1 \cup A_2) \cap A_3) \cap A_4) \cup A_5.$$

Как совместить справедливость дистрибутивного закона для случайных множеств (вытекающего из его справедливости для обычных множеств) с теоремой 2 подраздела 1.1.4 выше, в которой показано, что для нечетких множеств, вообще говоря, $(B_1 + B_2)B_3 \neq B_1B_3 + B_2B_3$ ? Дело в том, что хотя в соответствии с теоремой 7 для любых трех нечетких множеств $B_1$, $B_2$ и $B_3$ можно указать три случайных множества $A_1$, $A_2$ и $A_3$ такие, что

$$\Pr oj(A_i) = B_i, \ i = 1,2,3, \ \Pr oj(A_1 \cup A_2) = B_1 + B_2, \ \Pr oj((A_1 \cup A_2) \cap A_3) = B^3,$$

где

$$B^3 = (B_1 + B_2)B_3,$$

но при этом, вообще говоря,

$$\Pr oj(A_1 \cap A_3) \neq B_1B_3$$

и, кроме случаев, указанных в теореме 2 подраздела 1.1.4,

$$\Pr oj((A_1 \cup A_2) \cap A_3) \neq B_1B_3 + B_2B_3.$$

*Доказательство* теоремы 7 проводится по индукции. При $t$=1 распределение случайного множества строится с помощью теоремы 1. Затем конструируется само случайное множество $A_1$, определенное на делимом вероятностном пространстве (нетрудно проверить, что на делимом вероятностном пространстве можно построить случайное подмножество конечного множества с любым заданным распределением именно в силу делимости пространства). Далее случайные множества $A_2$, $A_3$, ..., $A_t$ строим по индукции с помощью теоремы 6. Теорема 7 доказана.

*Замечание.* Проведенное доказательство теоремы 9 проходит и в случае, когда при определении $B^m$ используются отрицания, точнее, кроме $B^m$ ранее введенного вида используются также последовательности результатов теоретико-множественных операций, очередной шаг в которых имеет вид

$$B_1^m = \overline{B^{m-1}} \circ B_m, \quad B_2^m = B^{m-1} \circ \overline{B_m}, \quad B_3^m = \overline{B^{m-1}} \circ \overline{B_m}.$$

А именно, сначала при помощи законов де Моргана (теорема 1 подраздела 1.1.4 выше) проводится преобразование, в результате которого в последовательности $B^m$ остаются только отрицания отдельных подмножеств из совокупности $B_1, B_2, B_3, ..., B_t$, а затем с помощью теоремы 3 вообще удается избавиться от отрицаний и вернуться к условиям теоремы 7.

Итак, в настоящем подразделе описаны связи между такими объектами нечисловой природы, как нечеткие и случайные множества, установленные в нашей стране в первой половине 1970-х годов. Через несколько лет, а именно, в начале 1980-х годов, близкие подходы

стали развиваться и за рубежом. Одна из работ [14] носит примечательное название "Нечеткие множества как классы эквивалентности случайных множеств".

В прикладной статистике и эконометрике [13] разработан ряд методов статистического анализа нечетких данных. В том числе методы классификации, регрессии, проверки гипотез о совпадении функций принадлежности по опытным данным и т.д. При этом оказались полезными общие подходы статистики объектов нечисловой природы (см. главу 3.4 ниже). Методологические и прикладные вопросы теории нечеткости обсуждались и в научно-популярной литературе (см., например, статью [15]).

### 1.4.7. Устойчивость выводов и принцип уравнивания погрешностей

**Устойчивость математических моделей.** Проблемам познания, в том числе в технических исследованиях, естественно-научных и социально-экономических областях, посвящено огромное количество работ. Однако это не значит, что обо всем в этой области уже все сказано. А о некоторых положениях целесообразно говорить еще и еще раз, пока они ни станут общеизвестными.

В идеале каждую модель порождения и анализа данных следовало бы рассматривать как аксиоматическую теорию. В этом идеальном случае создание и использование модели происходит в соответствии с известной триадой "практика - теория - практика". А именно, сначала вводятся некоторые математические объекты, соответствующие интересующим исследователя реальным объектам, и на основе представлений о свойствах реальных объектов формулируются необходимые для успешного моделирования свойства математических объектов, которые и принимаются в качестве аксиом. Затем аксиоматическая теория развивается как часть математики, вне связи с представлениями о реальных объектах. На заключительном этапе полученные в математической теории результаты интерпретируются содержательно. Получаются утверждения о реальных объектах, являющиеся следствиями тех и только тех их свойств, которые ранее были аксиоматизированы.

После построения математической модели реального явления или процесса встает вопрос об ее адекватности. Иногда ответ на этот вопрос может дать эксперимент. Рассогласование модельных и экспериментальных данных следует интерпретировать как признак неадекватности некоторых из принятых аксиом. Однако для проверки адекватности социально-экономических моделей зачастую невозможно поставить решающий эксперимент в отличие, скажем, от физических моделей. С другой стороны, для одного и того же явления или процесса, как правило, можно составить много возможных моделей, если угодно, много разновидностей одной базовой модели. Поэтому необходимы какие-то дополнительные условия, которые позволяли бы их множества возможных моделей и эконометрических методов анализа данных выбрать наиболее подходящие. В качестве одного из подобных условий выдвигается требование *устойчивости* модели и метода анализа данных относительно допустимых отклонений исходных данных и предпосылок модели или условий применимости метода.

Отметим, что в большинстве случаев исследователей и практических работников интересуют не столько сами модели и методы, сколько решения, которые с их помощью принимаются. Ведь модели и методы для того и разрабатываются, чтобы подготавливать решения. Вместе с тем очевидно, что решения, как правило, принимаются в условиях неполноты информации. Так, любые числовые параметры известны лишь с некоторой точностью. Введение в рассмотрение возможных неопределенностей исходных данных требует каких-то заключений относительно устойчивости принимаемых решений по отношению к этим допустимым неопределенностям.

Введем основные понятия согласно монографии [5]. Будем считать, что имеются *исходные данные,* на основе которых принимаются *решения.* Способ переработки (отображения) исходных данных в решение назовем *моделью.* Таким образом, с общей точки зрения модель - это функция, переводящая исходные данные в решение, т.е. способ перехода значения не имеет. Очевидно, любая рекомендуемая для практического использования модель должна быть

исследована *на устойчивость* относительно допустимых отклонений исходных данных. Укажем некоторые возможные применения результатов подобного исследования:

- заказчик научно-исследовательской работы получает представление о точности предлагаемого решения;

- удается выбрать из многих моделей наиболее адекватную;

- по известной точности определения отдельных параметров модели удается указать необходимую точность нахождения остальных параметров;

- переход к случаю "общего положения" позволяет получать более сильные с математической точки зрения результаты.

*Примеры.* По каждому из четырех перечисленных возможных применений в [5, 13] приведены различные примеры. В прикладной статистике точность предлагаемого решения связана с разбросом исходных данных и с объемом выборки. Выбору наиболее адекватной модели посвящены многие рассмотрения в главах 3.1 и 3.2, связанные с обсуждением моделей однородности и регрессии. Использование рационального объема выборки в статистике интервальных данных (глава 3.5) исходит из принципа уравнивания погрешностей. Этот принцип основан на том, что по известной точности определения отдельных параметров модели удается указать необходимую точность нахождения остальных параметров. Другим примером применения принципа уравнивания погрешностей является нахождение необходимой точности оценивания параметров в моделях логистики, рассмотренных в главе 5 монографии [5]. Наконец, переходом к случаю "общего положения" в прикладной статистике является, в частности, переход к непараметрическим методам, необходимый из-за невозможности обосновать принадлежность результатов наблюдений к тем или иным параметрическим семействам.

Специалисты по математическому моделированию и теории управления считают устойчивость одной из важных характеристик технических, социально-экономических, медицинских и иных моделей. Достаточно глубокие исследования ведутся по ряду направлений.

Первоначальное изучение влияния малого изменения одного параметра обычно называют *анализом чувствительности*. Оно описывается значением частной производной. Если модель задается дифференцируемой функцией, то итог анализа чувствительности - вектор значений частных производных в анализируемой точке.

Теория устойчивости решений дифференциальных уравнений развивается по крайней мере с XIX в. [16]. Выработаны соответствующие понятия - устойчивость по Ляпунову, корректность, доказаны глубокие теоремы. Для решения некорректных задач академиком АН СССР А.Н. Тихоновым в начале 1960-х годов был предложен метод регуляризации. Модели явлений и процессов, выражаемые с помощью дифференциальных уравнений, могут быть исследованы на устойчивость путем применения хорошо разработанного математического аппарата.

Вопросы устойчивости изучались практически во всех направлениях прикладных математических методов - и в математическом программировании, и в теории массового обслуживания (теории очередей), и в эколого-экономических моделях, и в различных областях эконометрики.

**Общая схема устойчивости.** Прежде чем переходить к конкретным постановкам, обсудим "общую схему устойчивости", дающую понятийную базу для обсуждения проблем устойчивости в различных предметных областях.

*Определение 1.* Общей схемой устойчивости называется объект $\{\mathrm{A}, \mathrm{B}, d, f, \mathrm{E}\}$.

Здесь $\mathrm{A}$ - множество, интерпретируемое как пространство исходных данных; $\mathrm{B}$ - множество, называемое пространством решений. Однозначное отображение $f : \mathrm{A} \to \mathrm{B}$ называется моделью. Об этих трех составляющих общей схемы устойчивости уже шла речь выше.

Оставшиеся два понятия нужны для уточнения понятий близости в пространстве исходных данных и пространстве решений. Подобные уточнения могут быть сделаны разными способами. Самое "слабое" уточнение - на языке топологических пространств. Тогда возможны качественные выводы (сходится - не сходится), но не количественные расчеты. Самое "сильное" уточнение - на языке метрических пространств. Промежуточный вариант - используются

показатели различия (отличаются от метрик тем, что не обязательно выполняются неравенства треугольника) или вводимые ниже понятия.

Пусть $d$ -показатель устойчивости, т.е. неотрицательная функция, определенная на подмножествах *У* множества В и такая, что из $Y_1 \subseteq Y_2$ вытекает $d(Y_1) \le d(Y_2)$. Часто показатель устойчивости *d(Y)* определяется с помощью метрики, псевдометрики или показателя различия (меры близости) $\rho$ как диаметр множества *У,* т.е.

$$d(Y) = \sup\{\rho(y_1, y_2), y_1 \in Y, y_2 \in Y\}.$$

Таким образом, говоря попросту, в пространстве решений с помощью показателя устойчивости вокруг образа исходных данных может быть сформирована система окрестностей. Но сначала надо такую систему сформировать в пространстве исходных данных.

Пусть $\mathrm{E} = \{E(x,\alpha), x \in \mathrm{A}, \alpha \in \Theta\}$ - совокупность допустимых отклонений. Т.е. система подмножеств множества А такая, что каждому элементу множества исходных данных $x \in \mathrm{A}$ и каждому значению параметра $\alpha$ из некоторого множества параметров $\Theta$ соответствует подмножество $E(x,\alpha)$ множества исходных данных. Оно называется множеством допустимых отклонений в точке *x* при значении параметра, равном $\alpha$. Наглядно можно представить себе, что вокруг точки *x* взята окрестность радиуса $\alpha$.

*Определение 2.* Показателем устойчивости в точке *x* при значении параметра, равном $\alpha$*,* называется число

$$\beta(x, E(x,\alpha)) = d(f(E(x,\alpha)).$$

Другими словами, это - диаметр образа множества допустимых колебаний при рассматриваемом в качестве модели отображении. Очевидно, что этот показатель устойчивости зависит как от исходных данных, так и от диаметра множества возможных отклонений в исходном пространстве. Для непрерывных функций показатель устойчивости обычно называется модулем непрерывности.

Естественно посмотреть, насколько сузится образ окрестности возможных отклонений при максимально возможном сужении этой окрестности.

*Определение 3.* Абсолютным показателем устойчивости в точке x называется число

$$\beta(x, \mathrm{E}) = \inf\{\beta(x, E(x,\alpha), \alpha \in \Theta\}.$$

Если функция $f$ непрерывна, а окрестности - именно те, о которых идет речь в математическом анализе, то максимальное сужение означает сужение к точке и абсолютный показатель устойчивости равен 0. Но в теории измерений и статистике интервальных данных мы сталкиваемся с совсем иными ситуациями. В теории измерений окрестностью исходных данных являются все те вектора, что получаются из исходного путем преобразования координат с помощью допустимого преобразования шкалы, а допустимое преобразование шкалы берется из соответствующей группы допустимых преобразований. В статистике интервальных данных под окрестностью исходных данных естественно понимать - при описании выборки - куб с ребрами $2\Delta$ и центром в исходном векторе. И в том, и в другом случае максимальное сужение не означает сужение к точке.

Естественным является желание ввести характеристики устойчивости на всем пространстве. Не вдаваясь в математические тонкости (см. о них монографию [5]), рассмотрим меру $\mu$ на пространстве А такую, что мера всего пространства равна 1 (т.е. $\mu(\mathrm{A}) = 1$).

*Определение 4.* Абсолютным показателем устойчивости на пространстве исходных данных А по мере $\mu$ называется число

$$\gamma(\mu) = \int_{\mathrm{A}} \beta(x, \mathrm{E}) d\mu.$$

Здесь имеется в виду интеграл Лебега. Интегрирование проводится по (абстрактному) пространству исходных данных А по мере $\mu$. Естественно, должны быть выполнены некоторые внутриматематические условия. Читателю, незнакомому с интегрированием по Лебегу, достаточно мысленно заменить в предыдущей формуле интеграл на сумму (а пространство А считать конечным, хотя и состоящим из большого числа элементов).

*Определение 5.* Максимальным абсолютным показателем устойчивости называется

$$\gamma = \sup\{\beta(x, \mathrm{E}), x \in \mathrm{A}\}.$$

Легко видеть, что $\gamma = \sup \gamma(\mu)$, где супремум берется по всем описанным выше мерам.

Итак, построена иерархия показателей устойчивости математических моделей реальных явлений и процессов. Она с успехом использовалась в различных исследованиях, подробно развивалась, в частности, в монографии [5]. Приведем еще одно полезное определение.

*Определение 6.* Модель $f$ называется абсолютно $\varepsilon$-устойчивой, если $\gamma \le \varepsilon$, где $\gamma$ - максимальный абсолютный показатель устойчивости.

*Пример.* Если показатель устойчивости формируется с помощью метрики $\rho$, совокупность допустимых отклонений $\mathrm{E}$ - это совокупность всех окрестностей всех точек пространства исходных данных $\mathrm{A}$, то 0-устойчивость модели $f$ эквивалентна непрерывности модели $f$ на множестве $\mathrm{A}$.

*Основная проблема в общей схеме устойчивости* - проверка $\varepsilon$-устойчивости данной модели $f$ относительно данной системы допустимых отклонений $\mathrm{E}$.

Часто оказываются полезными следующие два обобщения основной проблемы.

*Проблема А (характеризации устойчивых моделей).* Даны пространство исходных данных $\mathrm{A}$, пространство решений $\mathrm{B}$, показатель устойчивости $d$, совокупность допустимых отклонений $\mathrm{E}$ и неотрицательное число $\varepsilon$. Описать достаточно широкий класс $\varepsilon$-устойчивых моделей $f$. Или: найти все $\varepsilon$-устойчивые модели среди моделей, обладающих данными свойствами, т.е. входящих в данное множество моделей.

*Проблема Б (характеризации систем допустимых отклонений).* Даны пространство исходных данных $\mathrm{A}$, пространство решений $\mathrm{B}$, показатель устойчивости $d$, модель $f$ и неотрицательное число $\varepsilon$. Описать достаточно широкий класс систем допустимых отклонений $\mathrm{E}$, относительно которых модель $f$ *является* $\varepsilon$-устойчивой. Или*:* найти все такие системы допустимых отклонений $\mathrm{E}$ среди совокупностей допустимых отклонений, обладающих данными свойствами, т.е. входящих в данное множество совокупностей допустимых отклонений.

Ясно, что проблемы А и Б можно рассматривать не только для показателя устойчивости $\gamma$, но и для других только что введенных показателей устойчивости, а именно, $\gamma(\mu), \beta(x, \mathrm{E}), \beta(x, E(x, \alpha))$.

Язык общей схемы устойчивости позволяет описывать конкретные задачи специализированных теорий устойчивости в различных областях исследований, выделять в основные элементы в них, ставить проблемы типа А и Б. В частности, на этом языке легко формулируются задачи теории устойчивости решений дифференциальных уравнений, теории робастности статистических процедур (см. главу 2.2.), проблемы адекватности теории измерений, достигаемой точности расчетов в статистике интервальных данных и в логистике (см. монографию [5]), и т.д.

Для примера рассмотрим определение устойчивости по Ляпунову решения $\varphi(t, x)$ нормальной автономной системы дифференциальных уравнений $\dot{y} = g(y)$ с начальными условиями $\varphi(0, x) = x$. Здесь пространство исходных данных $\mathrm{A}$ - конечномерное евклидово пространство, множество допустимых отклонений $E(x, \alpha)-$ окрестность радиуса $\alpha$ точки $x \in \mathrm{A}$, пространство решений $\mathrm{B}$ - множество функций на луче $[0;+\infty)$ с метрикой

$$\rho(y_1, y_2) = \sup_{t \ge 0} | y_1(t) - y_2(t) |.$$

Модель $f$ - отображение, переводящее начальные условия $x$ в решение системы дифференциальных уравнений с этими начальными условиями $\varphi(t, x)$.

В терминах общей схемы устойчивости положение равновесия $a$ называется *устойчивым по Ляпунову*, если $\beta(a, \mathrm{E}) = 0$. Для формулировки определения асимптотической устойчивости по Ляпунову надо ввести в пространстве решений $\mathrm{B}$ псевдометрику

$$\rho_1(y_1, y_2) = \overline{\lim_{t \to \infty}} | y_1(t) - y_2(t) |.$$

Положение равновесия $a$ называется асимптотически устойчивым, если $\beta_1(a, E(a,\varepsilon)) = 0$ для некоторого $\varepsilon > 0$, где показатель устойчивости $\beta_1$ рассчитан с использованием псевдометрики $\rho_1$.

Таким образом, общая схема устойчивости естественным образом включает в себя классические понятия теории устойчивости по Ляпунову. Вместе с тем стоит отметить, что эта схема дает общий подход к различным проблемам устойчивости. Она дает систему понятий, которые в каждом конкретном случае должны приспосабливаться к решаемой задаче.

До настоящего момента для определенности речь шла о допустимых отклонениях в пространстве исходных данных. Часто оказывается необходимым говорить и об отклонениях от предпосылок модели. С чисто формальной точки зрения для этого достаточно расширить понятие "исходные данные" до пары (*x, f*), т.е. включив "прежнюю" модель в качестве второго элемента пары. Все остальные определения остаются без изменения. Теперь отклонения в пространстве решений вызываются не только отклонениями в исходных данных *x*, но и отклонениями от предпосылок модели, т.е. отклонениями *f*. Это соображение нам понадобится в подразделе 2.2.4, посвященном робастности статистических процедур.

**Устойчивость по отношению к объему выборки.** Различные асимптотические постановки в прикладной статистике также естественно рассматривать как задачи устойчивости. Если при безграничном возрастании объема выборки некоторая величина стремится к пределу, то в терминах общей схемы устойчивости это означает, что она 0-устойчива в соответствующей псевдометрике (см. выше обсуждение асимптотической устойчивости по Ляпунову). С содержательной точки зрения употребление термина "устойчивость" в такой ситуации представляется вполне оправданным, поскольку рассматриваемая величина мало меняется при изменении объема выборки.

Рассмотрим проблему и методы оценки близости предельных распределений статистик и распределений, соответствующих конечным объемам выборок. При каких объемах выборок уже можно пользоваться предельными распределениями? Каков точный смысл термина "можно" в предыдущей фразе? Основное внимание уделяется переходу от точных формул допредельных распределений к пределу и применению метода статистических испытаний (Монте-Карло).

Начнем с обсуждения взаимоотношений асимптотической математической статистики и практики анализа статистических данных. Как обычно подходят к обработке реальных данных в конкретной задаче? Первым делом строят статистическую модель. Если хотят перенести выводы с совокупности результатов наблюдений на более широкую совокупность, например, предсказать что-либо, то рассматривают, как правило, вероятностно-статистическую модель. Например, традиционную модель выборки, в которой результаты наблюдений - реализации независимых (в совокупности) одинаково распределенных случайных величин. Очевидно, *любая модель лишь приближенно соответствует реальности.* В частности, естественно ожидать, что распределения результатов наблюдений несколько отличаются друг от друга, а сами результаты связаны между собой, хотя и слабо.

Итак, первый этап - переход от реальной ситуации к математической модели. Далее - неожиданность: на настоящем этапе своего развития математическая теория статистики зачастую не позволяет провести необходимые исследования для имеющихся объемов выборок. Более того, отдельные математики пытаются оправдать свой отрыв от практики соображениями о структуре этой теории, на первый взгляд убедительными. Неосторожная давняя фраза Б.В. Гнеденко и А.Н.Колмогорова: "Познавательная ценность теории вероятностей раскрывается только предельными теоремами" (см. классическую монографию [17], одну из наиболее ценных математических книг XX в.) взята на вооружение и более близкими к нам по времени авторами. Так, И.А. Ибрагимов и Р.З. Хасьминский пишут: "Решение неасимптотических задач оценивания, хотя и весьма важное само по себе, как правило, не может являться объектом достаточно общей математической теории. Более того, соответствующее решение часто зависит от конкретного типа распределения, объема выборки и т.д. Так, теория малых выборок из нормального закона будет отличаться от теории малых выборок из закона Пуассона" (см. напичканную формулами монографию [18, с.7]).

Согласно цитированным и подобным им авторам, основное содержание математической теории статистики - предельные теоремы, полученные в предположении, что объемы рассматриваемых выборок стремятся к бесконечности. Эти теоремы опираются на предельные соотношения теории вероятностей, типа Закона Больших Чисел и Центральной Предельной Теоремы. Ясно, что сами по себе подобные утверждения относятся к математике, т.е. к сфере чистой абстракции, и не могут быть непосредственно применены для анализа реальных данных. Их практическое использование, о котором "чистые" математики предпочитают не думать, опирается на важное предположение: *«При данном объеме выборки достаточно точными являются асимптотические формулы».*

Конечно, в качестве первого приближения представляется естественным воспользоваться асимптотическими формулами, не тратя сил на анализ их точности. Но это - лишь начало долгой цепи исследований. Как же обычно преодолевают разрыв между результатами асимптотической математической статистики и потребностями практики статистического анализа данных? Какие "подводные камни" подстерегают на этом пути?

**Точные формулы и асимптотика.** Начнем с наиболее продвинутой в математическом плане ситуации, когда для статистики известны как предельное распределение, так и распределения при конечных объемах выборки.

Примером является двухвыборочная односторонняя статистика Н.В.Смирнова. Рассмотрим две независимые выборки объемов $m$ и $n$ из непрерывных функций распределения $F(x)$ и $G(x)$ соответственно. Для проверки гипотезы однородности двух выборок (ср. главу 3.1)

$$H_0: F(x) = G(x) \text{ для всех действительных чисел } x$$

в 1939 г. Н.В. Смирнов в статье [19] предложил использовать статистику

$$D^+(m,n) = sup\ (F_m(x) - G_n(x)),$$

где $F_m(x)$ - эмпирическая функция распределения, построенная по первой выборке, $G_n(x)$ - эмпирическая функция распределения, построенная по второй выборке, супремум берется по всем действительным числам $x$. Для обсуждения проблемы соотношения точных и предельных результатов ограничимся случаем равных объемов выборок, т.е. $m = n$. Положим

$$H(n,t) = P(D^+(n,n) \ge \frac{t}{\sqrt{n}}).$$

В цитированной статье [19] Н.В. Смирнов установил, что при безграничном возрастании объема выборки $n$ вероятность $H(n, t)$ стремится к $exp(-t^2)$.

В работе [20] 1951 г. Б.В. Гнеденко и В.С. Королюк показали, что при целом $c = t\sqrt{n}$ (именно при таких $t$ вероятность $H(n, t)$ как функция $t$ имеет скачки, поскольку статистика Смирнова $D^+(n,n)$ кратна $1/n$) рассматриваемая вероятность $H(n, t)$ выражается через биномиальные коэффициенты, а именно,

$$H(n,t) = \binom{2n}{n-c} / \binom{2n}{n}. \qquad (1)$$

К сожалению, непосредственные расчеты по формуле (1) возможны лишь при сравнительно небольших объемах выборок, поскольку величина $n!$ ($n$-факториал) уже при $n$=100 имеет более 200 цифр и не может быть без преобразований использована в вычислениях. Следовательно, наличие точной формулы для интересующей нас вероятности не снимает необходимости использования предельного распределения и изучения точности приближения с его помощью.

Широко известная формула Стирлинга для гамма-функции и, в частности, для факториалов позволяет преобразовать последнее выражение в асимптотическое разложение. Т.е. построить бесконечный степенной ряд (по степеням $n$) такой, что каждая следующая частичная сумма дает все более точное приближение для интересующей нас вероятности $H(x, t)$. Это и было сделано в работе А.А. Боровкова 1962 г. Большое количество подобных разложений для различных статистических задач приведено в работах В.М. Калинина и О.В. Шалаевского конца 1960-х - начала 1970-х годов. (Интересно отметить, что асимптотические разложения в ряде случаев расходятся, т.е. остаточные члены имеют нетривиальную природу.)

Затем в работах конца семидесятых годов была сделана попытка теоретически оценить остаточный член второго порядка. Итоги подведены в монографии [5, §2.2, с.37-45]. Справедливо равенство

$$H(n, t) = exp( - t^2).(1 + f(t)/n + g(n,t)/ n^2 ),$$

где

$$f(t) = t^2 (1/2 - t^2/6).$$

Целью последних из названных работ было получение равномерных по $n, t$ оценок остаточного члена второго порядка $g(n,t)$ сверху и снизу в области, задаваемой условиями

$$0 < \frac{t}{\sqrt{n}} < A, \quad 0 < t < t_{\max}, \quad , n \ge n_0. \qquad (2)$$

где $A$, $t_{\max}$, $n_0$ - некоторые параметры. С помощью длинных цепочек оценок остаточных членов в формулах, получаемых при преобразовании формулы (1) к предельному виду, сформулированная выше цель была достигнута. Для различных наборов параметров $A$, $t_{\max}$, $n_0$ получены равномерные по $n, t$ оценки (сверху и снизу) остаточного члена второго порядка $g(n,t)$ в области (2). Так, например, при $A = 0{,}5$, $t\ max = 1{,}73$, $n_0 = 8$ нижняя граница равна (- 0,71), а верхняя есть 2,65.

Основными недостатками такого подхода являются, во первых, зависимость оценок от параметров $A$, $t_{\max}$, $n_0$, задающих границы областей, во-вторых, завышение оценок, иногда в сотни раз, обусловленное желанием получить равномерные оценки по области (оценкой реальной погрешности в конкретной точке является значение следующего члена асимптотического разложения).

Поэтому при составлении рассчитанной на практическое использование методики [21] проверки однородности двух выборок с помощью статистики Смирнова было решено перейти на несколько другую методологию (назовем ее "методологией заданной точности"), которую кратко можно описать следующим образом.

1) выбирается достаточно малое положительное число $p$, например $p = 0{,}05$ или $p = 0{,}20$;

2) приводятся точные значения $H(n, t)$ для всех значений $n$ таких, что

$$|H(n, t) - exp( - t^2)| > p\ exp( - t^2);$$

3) если же последнее неравенство не выполнено, то используется вместо $H(n, t)$ предельное значение $exp(-t^2)$.

Таким образом, принятая в методике [21] методология предполагает интенсивное использование вычислительной техники. Результатами расчетов являются *граничные значения* объемов выборок $n(p,t)$ такие, что при меньших значениях объемов выборок рекомендуется пользоваться точными значениями функции распределения статистики Смирнова, а при больших - предельными. Описывается этот результат таблицей, а не формулой. Отметим, что при построении реальных таблиц не обойтись без выбора того или иного конкретного значения $p$, задающего объемы таблиц.

**Оценки скорости сходимости.** Теоретические оценки скорости сходимости в различных задачах прикладной математической статистики иногда формулируются в весьма абстрактном виде. Так, в 1960-1970-х годах была популярна задача оценки скорости сходимости распределения классической статистики омега-квадрат (Крамера-Мизеса-Смирнова). Для максимума модуля разности допредельной и предельной функций распределения этой статистики различные авторы доказывали, что для любого $e>0$ существует константа $C(e)$ такая, что он не превосходит $C(e)n^{-w+e}$. Прогресс состоял в увеличении константы $w$. Сформулированный выше результат был доказан последовательно для $w = 1/10, 1/6, 1/5, 1/4, 1/3, 1/2$ и 1 (подробнее история этих исследований рассказана в §2.3 монографии [5]).

Конечно, все эти исследования не могли дать конкретных практических рекомендаций. Однако необходимой исходной точкой является само существование предельного распределения. Представим себе, что некто, не зная, что у распределения Коши нет математического ожидания, моделирует выборочные средние арифметические результатов

наблюдений из этого распределения. Ясно, что его попытки оценить скорость сходимости выборочных средних к пределу обречены на провал.

Последовательное улучшение теоретических оценок скорости сходимости дает надежду на быструю реальную сходимость. Действительно, численные расчеты показали, что предельным распределением для статистики омега-квадрат (Крамера-Мизеса-Смирнова) можно пользоваться уже при объеме выборки, равном 4.

**Использование датчиков псевдослучайных чисел.** Если же предельное распределение известно, то возникает возможность изучить скорость сходимости численно методом статистических испытаний (Монте-Карло). Однако при этом обычно возникают две проблемы.

Во-первых, откуда известно, что скорость сходимости монотонна? Если при данном объеме выборки различие мало, то будет ли оно мало и при дальнейших объемах? Иногда отклонения допредельного распределения от предельного объясняются довольно сложными причинами. Так, для распределения хи-квадрат они связаны с рядом до сих пор не решенных теоретико-числовых проблем о числе целых точек в эллипсоиде растущего диаметра.

Во-вторых, с помощью датчиков псевдослучайных чисел получаем допредельные распределения с погрешностью, которая может преуменьшать различие. Поясним мысль аналогией. Растущий сигнал измеряется с погрешностями. Когда можно гарантировать, что его величина наверняка превзошла заданную границу?

Напомним, что проблема качества датчиков псевдослучайных чисел продолжает оставаться открытой (см. главу 11 в [13]). Для моделирования в пространствах фиксированной размерности датчики псевдослучайных чисел решают поставленные задачи. Но для рассматриваемых здесь задач размерность не фиксирована - мы не знаем, при каком конкретно объеме выборки можно переходить к предельному распределению согласно "методологии заданной точности".

Нужны дальнейшие работы по изучению качества датчиков псевдослучайных чисел в задачах неопределенной размерности. Поскольку критиков датчиков обычно обвиняют в том, что они сами их не используют, отметим, что мы применяли этот инструментарий при изучении помех, создаваемых электровозами (см. монографию [5]), при изучении статистических критериев проверки однородности двух выборок (см. работу [22]).

**А нужна ли вообще асимптотика?** В настоящее время развивается актуальное направление прикладной статистики, связанное с интенсивным использованием вычислительной техники для изучения свойств статистических процедур. Как уже отмечалось, математические методы в статистике обычно позволяют получать лишь асимптотические результаты, и для переноса выводов на конечные объемы выборок приходится применять вычислительные методы. В Новосибирском государственном техническом университете разработан и успешно применяется оригинальный подход, основанный на интенсивном использовании современной вычислительной техники. Основная идея такова: в качестве альтернативы асимптотическим методам математической статистики используется анализ результатов статистического моделирования (порядка 2000 испытаний) выборок конкретных объемов (200, 500, 1000). При этом анализ предельных распределений заменяется на анализ распределений соответствующих статистик при указанных объемах выборок.

К достоинствам подхода относится возможность замены теоретических исследований расчетами. Разработанная программная система дает (в принципе) возможность численно изучить свойства любого статистического алгоритма для любого конкретного распределения результатов наблюдений и любого конкретного объема выборки. К недостаткам рассматриваемого подхода относится зависимость от свойств датчиков псевдослучайных чисел, а также - что более важно - неизвестность предельного распределения (и даже самого факта его существования), а потому невозможность обоснованного переноса полученных выводов на объемы выборок, отличные от исследованных. Поэтому с точки зрения теории математической статистики полученные рассматриваемым способом результаты следует рассматривать как правдоподобные (а не доказательные, как в классической математической статистике).

Кроме того, они принципиально неточные. Даже в наиболее благоприятных условиях отклонение (в метрике «супремум разности») смоделированного распределения, построенного по

2000 испытаниям, от теоретического предельного распределения может достигать $1{,}358 \times (1/2000)^{1/2} = 0{,}030$ (см. главу 1.2). Это означает, в частности, что процентные точки, соответствующие уровням значимости 0,05 и особенно 0,01, могут сильно отличаться от соответствующих процентных точек предельных распределений. Очевидно, следующий этап работ - изучение точности полученных в рассматриваемом подходе выводов, прежде всего приближений и процентных точек.

Однако сразу все не сделаешь. Поэтому новосибирцы совершенно правы, развивая новые компьютерные подходы к давним задачам прикладной статистики. Так, весьма полезными и интересными являются результаты, касающиеся непараметрических критериев согласия и построения оптимального группирования, в частности, при использовании критериев типа хи-квадрат.

Однако стоит сделать два замечания. В работе [23] сравниваются два плана контроля надежности технических изделий. Оказывается, что при объемах выборки, меньших 150, лучше первый план, а при объемах, больших 150 - второй. Значит, если бы по новосибирскому методу сравнивались эти планы при достаточно большом объеме выборки $n$=100, то лучшим был бы признан первый план, что неверно - наступит момент (объем выборки), когда лучшим станет второй план.

Другая относящаяся к делу ассоциация - из весьма содержательной монографии о прикладной математике [24]. Будем суммировать бесконечный ряд с членами $z_n = 1/n$ . Поскольку члены его убывают, то обычно используемые алгоритмы остановят вычисления на каком-то шагу. А сумма-то - бесконечна!

Кажется, что компьютер дал универсальную отмычку ко всем проблемам вообще и в области прикладной статистики в частности. Но это только кажется.

**Принцип уравнивания погрешностей** состоит в том, что погрешности различной природы должны вносит примерно одинаковый вклад в общую погрешность математической модели. Так, определение рационального объема выборки в статистике интервальных данных основано на уравнивании влияния метрологической и статистической погрешностей. Согласно подходу [5] выбор числа градаций в социологических анкетах целесообразно проводить на основе уравнивания погрешностей квантования и неопределенности в ответах респондентов. В классической модели управления запасами целесообразно уравнять влияние неточностей в определении параметров на отклонение целевой функции от оптимума. Из принципа уравнивания погрешностей следует, что относительные погрешности определения параметров модели должны совпадать. Погрешность, порожденная отклонением спроса от линейного, оценивается по данным об отпуске товаров. Это дает возможность оценить допустимые отклонения для других параметров. В частности, установить, что расхождения между методиками не являются существенными [5].

В терминах общей схемы устойчивости рассмотрим для простоты записи случай двух параметров. Пусть $Б = [0, \infty) \times [0,\infty)$ и $E(x, б) = E(x, (е, д))$, где $е > 0$ и $д > 0$ задают точность определения соответствующих параметров, так что $E(x,(\varepsilon_1,\delta_1)) \subseteq E(x,(\varepsilon_2,\delta_2))$ при $е_1 \le е_2$, $д_1 \le д_2$. Пусть е задано, а д исследователь может выбрать, причем известно, что уменьшение д связано с увеличением расходов. Как выбирать д? Представляется естественным «уравнять» отклонения, порожденные различными параметрами, т.е. определить д из условия

$$в(x, E(x, (е, д)) – в(x, E(x, (е, 0)) \approx в(x, E(x, (0, д)).$$

Если затраты и полезный эффект точно известны, то д можно определить путем решения соответствующей оптимизационной задачи. В противном случае соотношение (3) предлагается использовать в качестве эвристического правила.

## Литература

1. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В.Прохоров. – М.: Большая Российская энциклопедия, 1999. – 910с.
2. Гнеденко Б.В. Курс теории вероятностей: Учебник. 7-е изд., исправл. - М.: Эдиториал УРСС, 2001. 320 с.

3. Рао С.Р. Линейные статистические методы и их применения. – М.: Наука, 1968. 548 с.
4. Келли Дж. Общая топология. - М.: Наука, 1968. - 384 с.
5. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука, 1979. - 296 с.
6. Орлов А.И. Асимптотическое поведение статистик интегрального типа. – В сб.: Вероятностные процессы и их приложения. Межвузовский сборник. - М.: МИЭМ, 1989. С.118-123.
7. Вентцель Е.С. Теория вероятностей. – М.: Наука, 1964.- 576 с.
8. Биллингсли П. Сходимость вероятностных мер. – М.: Наука, 1977. - 352 с.
9. Орлов А.И. Задачи оптимизации и нечеткие переменные. – М.: Знание, 1980. - 64 с.
10. Лебег А. Об измерении величин. - М.: Учпедгиз, 1960. - 204 с.
11. Ефимов Н.В. Высшая геометрия. - М.: ГИФМЛ, 1961. - 580 с.
12. Орлов А.И. Основания теории нечетких множеств (обобщение аппарата Заде). Случайные толерантности. – В сб.: Алгоритмы многомерного статистического анализа и их применения. - М.: Изд-во ЦЭМИ АН СССР, 1975. - С.169-175.
13. Орлов А.И. Эконометрика. Учебник для вузов. Изд. 2-е, исправленное и дополненное. - М.: Изд-во "Экзамен", 2003. – 576 с.
14. Goodman I.R. Fuzzy sets as equivalence classes of random sets // Fuzzy Set and Possibility Theory: Recent Developments. - New York-Oxford-Toronto-Sydney-Paris-Frankfurt, Pergamon Press, 1982. - P.327-343. (Перевод на русский язык: Гудмэн И. Нечеткие множества как классы эквивалентности случайных множеств. - В сб.: Нечеткие множества и теория возможностей. Последние достижения. - М.: Радио и связь, 1986. - С. 241-264.)
15. Орлов А.И. Математика нечеткости. – Журнал «Наука и жизнь». 1982. No.7. С.60-67.
16. Поляк Б.Т., Щербаков П.С. Робастная устойчивость и управление. – М.: Наука, 2002. – 303 с.
17. Гнеденко Б.В., Колмогоров А.Н. Предельные распределения для сумм независимых случайных величин. - М.-Л.: ГИТТЛ, 1949. - 264 с.
18. Ибрагимов И.А., Хасьминский Р.З. Асимптотическая теория оценивания. - М.: Наука, 1979. 528 с.
19. Смирнов Н.В. Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках. // Бюллетень. МГУ им. М.В. Ломоносова. Сер. А. 1939. Т.2. № 2. С.3-14.
20. Гнеденко Б.В., Королюк В.С. О максимальном расхождении двух эмпирических распределений. // Доклады АН СССР. 1951. Т.80. № 4. С.525-528.
21. Методика. Проверка однородности двух выборок параметров продукции при оценке ее технического уровня и качества. - М.: Всесоюзный научно-исследовательский институт стандартизации Госстандарта СССР, 1987. - 116 с.
22. Камень Ю.Э., Камень Я.Э., Орлов А.И. . Реальные и номинальные уровни значимости в задачах проверки статистических гипотез // Заводская лаборатория. 1986. Т. 52. №. 12. С. 55-57.
23. Левин Б.Р., Демидович Н.О. Использование непараметрических методов при обработке результатов испытаний на надежность. // Надежность средств связи. - Киев: Техніка, 1976. - С.59-72.
24. Блехман И.И., Мышкис А.Д., Пановко Я.Г. Механика и прикладная математика: Логика и особенности приложений математики. - М.: Наука, 1983. - 328 с.

**Контрольные вопросы и задачи**

1. Почему в прикладной статистике необходимо использовать теоремы о наследовании сходимости?
2. Примените метод линеаризации для изучения распределения выборочной дисперсии (исходя из асимптотической нормальности при $n \to \infty$ среднего арифметического двумерных векторов $(X_k, (X_k)^2)$, $k = 1, 2, \dots, n$).
3. Как применяется в прикладной статистике принцип инвариантности?
4. Как с точки зрения нечетких множеств можно интерпретировать вероятность накрытия определенной точки случайным множеством?

5. На множестве $Y = \{y_1, y_2, y_3\}$ задано нечеткое множество $B$ с функцией принадлежности $м_B(y)$, причем $м_B(y_1) = 0{,}1$, $м_B(y_2) = 0{,}2$, $м_B(y_3) = 0{,}3$. Постройте случайное множество $A$ так, чтобы *Proj* $A = B$.
6. На множестве $Y = \{y_1, y_2, y_3\}$ задано нечеткое множество $B$ с функцией принадлежности $м_B(y)$, причем $м_B(y_1) = 0{,}2$, $м_B(y_2) = 0{,}1$, $м_B(y_3) = 0{,}5$. Постройте случайное множество $A$ так, чтобы *Proj* $A = B$.
7. На множестве $Y = \{y_1, y_2, y_3\}$ задано нечеткое множество $B$ с функцией принадлежности $м_B(y)$, причем $м_B(y_1) = 0{,}5$, $м_B(y_2) = 0{,}4$, $м_B(y_3) = 0{,}7$. Постройте случайное множество $A$ так, чтобы *Proj* $A = B$.
8. На множестве $Y = \{y_1, y_2, y_3\}$ задано нечеткое множество $B$ с функцией принадлежности $м_B(y)$, причем $м_B(y_1) = 0{,}3$, $м_B(y_2) = 0{,}2$, $м_B(y_3) = 0{,}1$. Постройте случайное множество $A$ так, чтобы *Proj* $A = B$.
9. В чем состоит основная идея принципа уравнивания погрешностей?

## Темы докладов, рефератов, исследовательских работ

1. Законы больших чисел и различные варианты Центральной предельной теоремы – основные результаты классической теории вероятностей.
**2.** Место теорем о наследовании сходимости и метода линеаризации в асимптотической прикладной статистике.
3. Принцип инвариантности для классических непараметрических статистик.
4. Обсудите суждение: «Мы мыслим нечетко» (см. [15]). Почему нечеткость мышления помогает взаимопониманию?
5. Взаимосвязь теории нечеткости и теории вероятностей.
6. Методы оценивания функции принадлежности.
7. Теория нечеткости и интервальная математика.
8. Описание данных для выборок, элементы которых – нечеткие множества.
9. Регрессионный анализ нечетких переменных (согласно [9]).
10. Кластерный анализ нечетких данных.
11. Непараметрические оценки плотности распределения вероятностей в пространстве нечетких множеств (согласно подходу главы 2.1).
12. Проблема устойчивости в математическом моделировании.

# Часть 2. Основные проблемы прикладной статистики

## 2.1. Описание данных

### 2.1.1. Модели порождения данных

**Детерминированный и модельно-вероятностный подходы.** В прикладной статистике есть два подхода к исходным данным – детерминированный и модельно-вероятностный. В первом из них данные рассматриваются сами по себе, без попыток связать их с какой-либо более общей ситуацией. Например, при анализе данных о производственной деятельности конкретного предприятия за конкретный период времени подсчитывается процент брака по конкретным технологическим процессам, число работников на различных должностях, объем реализованной продукции по месяцам. К этой же категории данных относятся различные виды отчетности – бухгалтерская, налоговая, статистическая (для органов Госкомстата РФ). Преимуществом детерминированного подхода является отсутствие каких-либо дополнительных предположений о данных. Недостаток состоит в невозможности обоснованного переноса выводов с конкретной ситуации на другие, ей аналогичные. Например, на другие периоды времени или на другие предприятия. При детерминированном подходе невозможно также оценить погрешность рассчитанных характеристик.

Чтобы выйти за пределы конкретной ситуации, необходимо использовать модельно-вероятностный подход, согласно которому основой алгоритмов расчетов является вероятностная модель порождения данных. При этом конкретные данные рассматриваются как реализации случайных величин, векторов, более общо – элементов, т.е как значения задающих их функций, определенных на вероятностном пространстве, в конкретной точке (элементарном событии щ).

Наиболее распространенная вероятностная модель порождения данных – это модель случайной выборки. Согласно этой модели данные $x_1, x_2, \dots, x_n$ рассматриваются как реализации независимых одинаково распределенных случайных элементов (величин, векторов, множеств и других объектов нечисловой природы) $X_1 = X_1(щ), X_2 = X_2(щ), \dots, X_n = X_n(щ)$, т.е. $x_1 = X_1(щ_0), x_2 = X_2(щ_0), \dots, x_n = X_n(щ_0)$ при некотором $щ_0$ из пространства элементарных событий Щ. Модель выборки обычно используется для описания результатов независимых наблюдений, измерений, анализов. опытов.

В некоторых случаях используют более специальные модели порождения данных. Например, при проведении испытаний на надежность используют план испытаний, согласно которому испытания прекращаются через время $T$. Это значит, что фиксируются только моменты отказа изделий, которые произошли до момента Т. Пусть $x_1, x_2, \dots, x_n$ – наработки на отказ $n$ изделий. Статистику доступны только значения $y_1, y_2, \dots, y_n$, где $y_j = x_j$ при $x_j < T$ и $y_j = T$ при $x_j \geq T$. Такая выборка, в которой часть описывающих реальное явление случайных величин заменена на граничное значение, называется цензурированной. Иногда используются и более сложные модели порождения данных. Например, если аппаратурой не фиксируются значения, меньшие некоторого порога, то выборка не только цензурирована, но и состоит из случайного числа элементов. Бывают и процедуры, когда минимальный и максимальный элементы выборки отбрасываются, а остальные предоставляются статистику, и т.д.

**Параметрические и непараметрические** м**одели случайной выборки**. Рассмотрим ситуацию, когда элементы выборки – числа. Модель описывается функцией распределения элементов выборки. Можно ли что-либо сказать об этой функции?

В учебных курсах по теории вероятностей и математической статистике обычно рассматривают различные параметрические семейства распределений числовых случайных величин. А именно, изучают семейства нормальных распределений, логарифмически нормальных, экспоненциальных, гамма-распределений, распределений Вейбулла-Гнеденко и др. Все они зависят от одного, двух или трех параметров. Поэтому для полного описания распределения достаточно

знать или оценить одно, два или три числа. Очень удобно. Поэтому широко развита и представлена в литературе параметрическая теория математической статистики, в которой предполагается, что распределения результатов наблюдений принадлежат тем или иным параметрическим семействам.

К сожалению, параметрические семейства существуют лишь в головах авторов учебников по теории вероятностей и математической статистике. В реальной жизни их нет. Поэтому прикладная статистика использует в основном непараметрические методы, в которых распределения результатов наблюдений могут иметь произвольный вид. В настоящем подразделе на примере нормального распределения подробно обсудим невозможность практического использования параметрических семейств для описания распределений конкретных анных.

В главе 2.3 разобраны параметрические методы отбраковки резко выделяющихся наблюдений и продемонстрирована невозможность практического использования ряда методов параметрической статистики, ошибочность выводов, к которым они приводят. В главе 3.1 рассмотрены непараметрические методы доверительного оценивания основных характеристик числовых случайных величин - математического ожидания, медианы, дисперсии, среднего квадратического отклонения, коэффициента вариации.

К настоящему времени непараметрические методы полностью покрывают область задач, которые ранее решались с помощью параметрической статистики. Поэтому можно порекомендовать использовать только непараметрическую статистику. Однако в литературе много внимания уделяется параметрическим методам, поэтому игнорировать в настоящем учебнике параметрическую статистику было признано нецелесообразным.

**Часто ли распределение результатов наблюдений является нормальным?** В эконометрических и экономико-математических моделях, применяемых, в частности, при изучении и оптимизации процессов маркетинга и менеджмента в целом, управления предприятием и регионом, точности и стабильности технологических процессов, в задачах надежности, обеспечения безопасности, в том числе экологической, функционирования технических устройств и объектов, разработки организационных схем часто применяют понятия и результаты теории вероятностей и математической статистики. При этом зачастую используют те или иные параметрические семейства распределений вероятностей. Наиболее популярно нормальное распределение. Используют также логарифмически нормальное распределение, экспоненциальное распределение, гамма-распределение, распределение Вейбулла-Гнеденко и т.д.

Очевидно, всегда необходимо проверять соответствие моделей реальности. Возникают два вопроса. Отличаются ли реальные распределения от используемых в модели? Насколько это отличие влияет на выводы?

Ниже на примере нормального распределения показано, что реальные распределения практически всегда отличаются от включенных в классические параметрические семейства. Имеющиеся отклонения от заданных семейств делают неверными выводы, основанные на использовании этих семейств. Например, выводы об отбраковке резко отличающихся наблюдений (выбросов).

Есть ли основания априори предполагать нормальность результатов измерений?

Иногда утверждают, что в случае, когда погрешность измерения (или иная случайная величина) определяется в результате совокупного действия многих малых факторов, то в силу Центральной Предельной Теоремы (ЦПТ) теории вероятностей эта величина хорошо приближается (по распределению) нормальной случайной величиной. Такое утверждение справедливо, если малые факторы действуют аддитивно и независимо друг от друга. Если же они действуют мультипликативно, то в силу той же ЦПТ аппроксимировать надо логарифмически нормальным распределением. В прикладных задачах обосновать аддитивность, а не мультипликативность действия малых факторов обычно не удается. Если же зависимость имеет общий характер, не приводится к аддитивному или мультипликативному виду, а также нет оснований принимать модели, дающие экспоненциальное, Вейбулла-Гнеденко, гамма или иные распределения, то о распределении итоговой случайной величины практически ничего не известно, кроме

внутриматематических свойств типа регулярности.

**Экспериментальное изучение распределений погрешностей**. При обработке конкретных данных иногда считают, что погрешности измерений имеют нормальное распределение. На предположении нормальности построены классические модели регрессионного, дисперсионного, факторного анализов, метрологические модели, которые еще продолжают встречаться как в отечественной нормативно-технической документации, так и в международных стандартах. На то же предположение опираются модели расчетов максимально достигаемых уровней тех или иных характеристик, применяемые при проектировании систем обеспечения безопасности функционирования экономических структур, технических устройств и объектов. Однако теоретических оснований для такого предположения нет. Необходимо экспериментально изучать распределения погрешностей.

Что же показывают результаты экспериментов? Сводка, данная в монографии [1], позволяет утверждать, что в большинстве случаев распределение погрешностей измерений отличается от нормального. Так, в Машинно-электротехническом институте (г. Варна в Болгарии) было исследовано распределение погрешностей градуировки шкал аналоговых электроизмерительных приборов. Изучались приборы, изготовленные в Чехословакии, СССР и Болгарии. Закон распределения погрешностей оказался одним и тем же. Он имеет плотность

$$f(x) = 0{,}534\exp(1-|x|^7).$$

Были проанализированы данные о параметрах 219 фактических распределений погрешностей, исследованных разными авторами, при измерении как электрических, так и не электрических величин самыми разнообразными (электрическими) приборами. В результате этого исследования оказалось, что 111 распределений, т.е. примерно 50% , принадлежат классу распределений с плотностью

$$f(x;\alpha,b.\sigma) = \frac{\alpha}{2\lambda\sigma\Gamma(1/\alpha)}\exp\left(-|\frac{x-b}{\lambda\sigma}|^{\alpha}\right),$$

где $\alpha$ - параметр степени (формы); $b$ - параметр сдвига; $\sigma$ - параметр масштаба; $\Gamma(\beta)$ - гамма-функция от аргумента $\beta$;

$$\lambda = \sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}$$

(см. [1, с. 56]); 63 распределения, т.е. 30%, имеют плотности с плоской вершиной и пологими длинными спадами и не могут быть описаны как нормальные или, например, экспоненциальные. Оставшиеся 45 распределений оказались двухмодальными.

В книге известного метролога проф. П. В. Новицкого [2] приведены результаты исследования законов распределения различного рода погрешностей измерения. Он изучил распределения погрешностей электромеханических приборов на кернах, электронных приборов для измерения температур и усилий, цифровых приборов с ручным уравновешиванием. Объем выборок экспериментальных данных для каждого экземпляра составлял 100-400 отсчетов. Оказалось, что 46 из 47 распределений значимо отличались от нормального. Исследована форма распределения погрешностей у 25 экземпляров цифровых вольтметров Щ-1411 в 10 точках диапазона. Результаты аналогичны. Дальнейшие сведения содержатся в монографии [1].

В лаборатории прикладной математики Тартуского государственного университета проанализировано 2500 выборок из архива реальных статистических данных. В 92% гипотезу нормальности пришлось отвергнуть.

Приведенные описания экспериментальных данных показывают, что погрешности измерений в большинстве случаев имеют распределения, отличные от нормальных. Это означает, в частности, что большинство применений критерия Стьюдента, классического регрессионного анализа и других статистических методов, основанных на нормальной теории, строго говоря, не является обоснованным. Поскольку неверна лежащая в их основе аксиома нормальности распределений соответствующих случайных величин.

Очевидно, для оправдания или обоснованного изменения существующей практики анализа статистических данных требуется изучить свойства процедур анализа данных при "незаконном" применении. Изучение процедур отбраковки показало, что они крайне неустойчивы к отклонениям от нормальности, а потому применять их для обработки реальных данных нецелесообразно (см. главу 2.3); поэтому нельзя утверждать, что произвольно взятая процедура устойчива к отклонениям от нормальности.

Иногда предлагают перед применением, например, критерия Стьюдента однородности двух выборок проверять нормальность. Хотя для этого имеется много критериев, но проверка нормальности - более сложная и трудоемкая статистическая процедура, чем проверка однородности (как с помощью статистик типа Стьюдента, так и с помощью непараметрических критериев). Для достаточно надежного установления нормальности требуется весьма большое число наблюдений. Так, чтобы гарантировать, что функция распределения результатов наблюдений отличается от некоторой нормальной не более, чем на 0,01 (при любом значении аргумента), требуется порядка 2500 наблюдений. В большинстве экономических, технических, медико-биологических и других прикладных исследований число наблюдений существенно меньше. Особенно это справедливо для данных, используемых при изучении проблем, связанных с обеспечением безопасности функционирования экономических структур и технических объектов.

**ЦПТ и нормальность.** Иногда пытаются использовать ЦПТ для приближения распределения погрешности к нормальному, включая в технологическую схему измерительного прибора специальные сумматоры. Оценим полезность этой меры. Пусть $Z_1$ , $Z_2$ ,..., $Z_k$ - независимые одинаково распределенные случайные величины с функцией распределения $H = H(x)$ такие, что $M(Z_1) = 0,\ D(Z_1) = 1,\ M\mid Z_1\mid^3 = \rho < +\infty.$ Рассмотрим

$$w = \frac{Z_1 + Z_2 + ... + Z_k}{\sqrt{k}}.$$

Показателем обеспечиваемой сумматором близости к нормальности является

$$C = \sup_H \sup_x \mid P(w < x) - \Phi(x) \mid .$$

Тогда

$$0{,}3989\frac{\rho}{\sqrt{k}} \le C \le 0{,}7975\frac{\rho}{\sqrt{k}}.$$

Правое неравенство в последнем соотношении вытекает из оценок константы в неравенстве Берри-Эссеена, полученном в книге [3, с.172], а левое - из примера в монографии [4, с.140-141]. Для нормального закона $\rho$ =1,6, для равномерного $\rho$ = 1,3, для двухточечного $\rho$ =1 (это - нижняя граница для $\rho$). Следовательно, для обеспечения расстояния (в метрике Колмогорова) до нормального распределения не более 0,01 для "неудачных" распределений необходимо не менее $k_0$ слагаемых, где

$$0{,}4\sqrt{k_0} < 0{,}01,\ \ k_0 > 1600.$$

В обычно используемых сумматорах слагаемых значительно меньше.

Сужая класс возможных распределений $H$, можно получить, как показано в монографии [5], более быструю сходимость, но теория здесь еще не смыкается с практикой. Кроме того, не ясно, обеспечивает ли близость распределения к нормальному (в определенной метрике) также и близость распределений статистик. Речь идет о сравнении распределения статистики, построенной по случайным величинам, полученным суммированием, к распределению статистики, соответствующей нормальным результатам наблюдений. Видимо, для каждой конкретной статистики необходимы специальные теоретические исследования, Именно к такому выводу приходит автор монографии [5]. В задачах отбраковки выбросов ответ: "Не обеспечивает" (см. ниже).

Отметим, что результат любого реального измерения записывается с помощью конечного числа десятичных знаков, обычно небольшого (2-5), так что любые реальные данные

целесообразно моделировать лишь с помощью дискретных случайных величин, принимающих сравнительно небольшое число значений. Нормальное распределение - лишь аппроксимация реального распределения. Так, например, данные конкретного исследования, приведенные в работе [6], принимают значения от 1,0 до 2,2, т.е. всего 13 возможных значений. Из принципа Дирихле следует, что в какой-то точке построенная по данным работы [6] функция распределения отличается от ближайшей функции нормального распределения не менее чем на 1/26, т.е. на 0,04. Кроме того, очевидно, что для нормального распределения случайной величины вероятность попасть в дискретное множество десятичных чисел с заданным числом знаков после запятой равна 0.

Из сказанного выше следует, что результаты измерений и вообще статистические данные имеют свойства, приводящие к тому, что моделировать их следует случайными величинами с распределениями, более или менее отличными от нормальных. В большинстве случаев распределения существенно отличаются от нормальных. В других ситуациях нормальные распределения могут, видимо, рассматриваться как некоторая аппроксимация. Но никогда нет полного совпадения. Отсюда вытекает необходимость изучения свойств классических статистических процедур в неклассических вероятностных моделях (подобно тому, как это сделано в главе 3.1 для критерия Стьюдента). А также целесообразность разработки устойчивых (учитывающих наличие отклонений от нормальности) и непараметрических, в том числе свободных от распределения процедур, их широкого внедрения в практику статистической обработки данных.

Опущенные здесь рассмотрения для других параметрических семейств приводят к аналогичным выводам. Итог можно сформулировать так. Распределения реальных данных практически никогда не входят в какое-либо конкретное параметрическое семейство. Реальные распределения всегда отличаются от тех, что включены в параметрические семейства. Отличия могут быть большие или маленькие, но они всегда есть.

### 2.1.2. Таблицы и выборочные характеристики

Исходные статистические данные могут быть достаточно обширными. В качестве примера приведем результаты экспертного опроса, проведенного Институтом высоких статистических технологий и эконометрики в 1994 г. (табл.1). В первом столбце приведены номера экспертов, в остальных четырех – четыре прогнозных значения, полученных от каждого эксперта. Отметим, что эксперт №28 не ответил на вопрос об инфляции. В таблицах реальных данных приходися сталкиваться с пропусками.

Таблица 1.

Прогнозы экспертов на 8 декабря 1994 г. (сделаны 19.10.1994)

| № п/п | Курс доллара США, руб. | Инфляция (%) за период прогноза | Цена батона белого хлеба, руб. | Цена 1 л молока, руб. |
|---|---|---|---|---|
| 1 | 4185 | 4,0 | 800 | 1305 |
| 2 | 4270 | 2,8 | 1028 | 1322 |
| 3 | 3200 | 17,0 | 760 | 755 |
| 4 | 4000 | 16,0 | 950 | 1000 |
| 5 | 3500 | 16,0 | 820 | 800 |
| 6 | 3800 | 5,0 | 1000 | 1000 |
| 7 | 3500 | 3,5 | 500 | 1500 |
| 8 | 3300 | 62,0 | 800 | 780 |
| 9 | 4100 | 54,0 | 900 | 899 |
| 10 | 3560 | 10,0 | 870 | 1050 |
| 11 | 4000 | 54,0 | 1000 | 1000 |

| | | | | |
|---|---|---|---|---|
| 12 | 5200 | 54,0 | 1500 | 1500 |
| 13 | 4000 | 9,0 | 830 | 1300 |
| 14 | 6000 | 54,0 | 2000 | 2000 |
| 15 | 4000 | 40,0 | 950 | 1200 |
| 16 | 3400 | 13,0 | 750 | 900 |
| 17 | 3500 | 15,0 | 1000 | 1250 |
| 18 | 4200 | 2,5 | 1000 | 1500 |
| 19 | 3560 | 200,0 | 940 | 1200 |
| 20 | 4300 | 6,0 | 950 | 1570 |
| 21 | 4000 | 3,0 | 1000 | 1100 |
| 22 | 4500 | 12,0 | 950 | 1100 |
| 23 | 4200 | 11,0 | 890 | 1100 |
| 24 | 3900 | 54,0 | 1000 | 1000 |
| 25 | 5500 | 62,0 | 1000 | 1400 |
| 26 | 5000 | 73,0 | 1000 | 1200 |
| 27 | 5600 | 54,0 | 1200 | 2000 |
| 28 | 3900 | - | 1500 | 1400 |
| 29 | 4200 | 38,0 | 950 | 1100 |
| 30 | 3680 | 38,0 | 850 | 1100 |
| 31 | 4000 | 2,0 | 840 | 1100 |
| 32 | 4600 | 46,0 | 1000 | 1100 |
| 33 | 4560 | 92,0 | 1300 | 1400 |

Описание данных - это первичное сжатие информации с целью сделать ее более обозримой, легкой для восприятия. Самый древний способ – это составление различных таблиц, вторичных по отношению к таблицам исходных данных.

Например, рассмотрим последний столбец табл.1. Для лучшего восприятия прогнозов экспертов о цене 1 л молока сгруппируем данные по интервалам, как это сделано в табл.2.

Таблица 2.
Прогнозируемая цена молока

| № п/п | Интервал, руб. | Число ответов |
|---|---|---|
| 1 | 700 – 799 | 2 |
| 2 | 800 – 899 | 2 |
| 3 | 900 – 999 | 1 |
| 4 | 1000 – 1099 | 5 |
| 5 | 1100 – 1199 | 7 |
| 6 | 1200 – 1299 | 4 |
| 7 | 1300 – 1399 | 3 |
| 8 | 1400 – 1499 | 3 |
| 9 | 1500 – 1599 | 4 |
| 10 | 2000 | 2 |
| | Всего | 33 |

Группировка данных в табл.2 по 10 интервалам может показаться слишком дробной. Нетрудно объединить градации и получить, например, табл.3.

Таблица 3.
Прогнозируемая цена молока (крупные градации)

| № п/п | Интервал, руб. | Число ответов |
|---|---|---|

| | | |
|---|---|---|
| 1 | 700 – 999 | 5 |
| 2 | 1000 – 1299 | 16 |
| 3 | 1300 – 1599 | 10 |
| 4 | 2000 | 2 |
| 5 | Всего | 33 |

Сколько использовать градаций (т.е. строк в таблице)? Общих рекомендаций дать нельзя. Ответ зависит от цели статистического исследования, от структуры конкретных данных.

Табличный материал может быть выражен в виде различных диаграмм, в том числе круговых и столбчатых. Несколько десятков лет назад были популярны гистограммы – столбчатые диаграммы, для которых интервалы группирования имеют одинаковую длину.

В настоящее время гистограммы рассматривают как устаревшие инструменты статистического анализа. Для описания массива данных рекомендуется использовать вариационные ряды, эмпирические функции распределения (см. главу 1.2) и – особенно настоятельно – непараметрические оценки плотности (см. подраздел 2.1.6). Кроме того, целесообразно рассчитывать и приводить в документации в разделе «Описание данных» выборочные характеристики:

- выборочное среднее арифметическое;
- выборочную дисперсию;
- выборочное среднее квадратическое отклонение;
- коэффициент вариации
- медиану;
- минимум (первый член вариационного ряда);
- максимум (последний член вариационного ряда);
- размах
- моду и амплитуду моды;
- верхний квартиль;
- нижний квартиль;
- межквартильное расстояние.

Определения всех этих выборочных характеристик даны выше в главе 1.2. В настоящем подразделе сведены вместе наиболее распространенные приемы описания числовых данных.

### 2.1.3. Шкалы измерения, инвариантные алгоритмы и средние величины

**Инвариантные алгоритмы и средние величины.** Основное требование к алгоритмам анализа данных формулируется в теории измерений (см. главу 1.1) так: *выводы, сделанные на основе данных, измеренных в шкале определенного типа, не должны меняться при допустимом преобразовании шкалы измерения этих данных.* Другими словами, выводы должны быть *инвариантны* по отношению к допустимым преобразованиям шкалы.

Таким образом, одна из основных целей теории измерений - борьба с субъективизмом исследователя при приписывании численных значений реальным объектам. Так, расстояния можно измерять в аршинах, метрах, микронах, милях, парсеках и других единицах измерения. Массу (вес) - в пудах, килограммах, фунтах и др. Цены на товары и услуги можно указывать в юанях, рублях, тенге, гривнах, латах, кронах, марках, долларах США и других валютах (при фиксированных курсах пересчета). Подчеркнем очень важное, хотя и вполне очевидное обстоятельство: выбор единиц измерения зависит от исследователя, т.е. субъективен. *Статистические выводы могут быть адекватны реальности только тогда, когда они не зависят от того, какую единицу измерения предпочтет исследователь, т.е. когда они инвариантны относительно допустимого преобразования шкалы.*

Оказывается, сформулированное условие является достаточно сильным. Из многих алгоритмов анализа статистических данных ему удовлетворяют лишь некоторые. Покажем это на примере сравнения средних величин.

Пусть $X_1, X_2, ..., X_n$ - выборка объема $n$. Часто используют среднее арифметическое

$$X_{ср} = \frac{X_1 + X_2 + ... + X_n}{n}.$$

Использование среднего арифметического настолько привычно, что второе слово в термине часто опускают. И говорят о средней зарплате, среднем доходе и других средних для конкретных экономических данных, подразумевая под "средним" среднее арифметическое. Такая традиция может приводить к ошибочным выводам. Покажем это на примере расчета средней заработной платы (среднего дохода) работников условного предприятия (табл.4).

Таблица 4.
Численность работников различных категорий, их заработная плата и суммарные доходы (в условных единицах).

| № п/п | Категория работников | Число работников | Заработная плата | Суммарные доходы |
|---|---|---|---|---|
| 1 | Низкоквалифицированные рабочие | 40 | 100 | 4000 |
| 2 | Высококвалифицированные рабочие | 30 | 200 | 6000 |
| 3 | Инженеры и служащие | 25 | 300 | 7500 |
| 4 | Менеджеры | 4 | 1000 | 4000 |
| 5 | Генеральный директор (владелец) | 1 | 18500 | 18500 |
| 6 | Всего | 100 | | 40000 |

Первые три строки в табл.4 вряд ли требуют пояснений. Менеджеры - это директора по направлениям, а именно, по производству (главный инженер), по финансам, по маркетингу и сбыту, по персоналу (по кадрам). Владелец сам руководит предприятием в качестве генерального директора. В столбце "заработная плата" указаны доходы одного работника соответствующей категории, а в столбце "суммарные доходы" - доходы всех работников соответствующей категории.

Фонд оплаты труда составляет 40000 единиц, работников всего 100, следовательно, средняя заработная плата составляет 40000/100 = 400 единиц. Однако эта средняя арифметическая величина явно не соответствует интуитивному представлению о "средней зарплате". Из 100 работников лишь 5 имеют заработную плату, ее превышающую, а зарплата остальных 95 существенно меньше средней арифметической. Причина очевидна - заработная плата одного человека - генерального директора - превышает заработную плату 95 работников - низкоквалифицированных и высококвалифицированных рабочих, инженеров и служащих.

Ситуация напоминает описанную в известном рассказе о больнице, в которой 10 больных, из них у 9 температура 40 $^0$C, а один уже отмучился, лежит в морге с температурой 0$^0$C. Между тем средняя температура по больнице равна 36$^0$C - лучше не бывает!

Сказанное показывает, что среднее арифметическое можно использовать лишь для достаточно однородных совокупностей (без больших выбросов в ту или иную сторону). А какие средние целесообразно использовать для описания заработной платы? Вполне естественно использовать медиану. Для данных табл.4 медиана - среднее арифметическое 50-го и 51-го работника, если их заработные платы расположены в порядке неубывания. Сначала идут зарплаты 40 низкоквалифицированных рабочих, а затем - с 41-го до 70-го работника - заработные платы высококвалифицированных рабочих. Следовательно, медиана попадает именно на них и равна 200. У 50-ти работников заработная плата не превосходит 200, и у 50-ти - не менее 200, поэтому

медиана показывает "центр", около которого группируется основная масса исследуемых величин. Еще одна средняя величина - мода, наиболее часто встречающееся значение. В рассматриваемом случае это заработная плата низкоквалифицированных рабочих, т.е. 100. Таким образом, для описания зарплаты имеем три средние величины - моду (100 единиц), медиану (200 единиц) и среднее арифметическое (400 единиц). Для наблюдающихся в реальной жизни распределений доходов и заработной платы справедлива та же закономерность: мода меньше медианы, а медиана меньше среднего арифметического.

Для чего в технических, экономических, медицинских и иных исследованиях используются средние величины? Обычно для того, чтобы заменить совокупность чисел одним числом, чтобы сравнивать совокупности с помощью средних.

Пусть, например, $Y_1, Y_2,...,Y_n$ - совокупность оценок экспертов, "выставленных" одному объекту экспертизы (например, одному из вариантов стратегического развития фирмы), $Z_1, Z_2,...,Z_n$ - второму (другому варианту такого развития). Как сравнивать эти совокупности? Очевидно, самый простой способ - по средним значениям.

А как вычислять средние? Известны различные виды средних величин: среднее арифметическое, медиана, мода, среднее геометрическое, среднее гармоническое, среднее квадратическое. Напомним, что общее понятие средней величины введено французским математиком первой половины XIX в. академиком О. Коши. Оно таково: средней величиной является любая функция $f(X_1, X_2,...,X_n)$ такая, что при всех возможных значениях аргументов значение этой функции не меньше, чем минимальное из чисел $X_1, X_2,...,X_n$, и не больше, чем максимальное из этих чисел. Все перечисленные выше виды средних являются средними по Коши.

При допустимом преобразовании шкалы значение средней величины, очевидно, меняется. Но выводы о том, для какой совокупности среднее больше, а для какой - меньше, не должны меняться (в соответствии с требованием инвариантности выводов, принятом как основное требование в теории измерений). Сформулируем соответствующую математическую задачу поиска вида средних величин, результат сравнения которых устойчив относительно допустимых преобразований шкалы.

Пусть $f(X_1, X_2,...,X_n)$ - среднее по Коши. Пусть среднее по первой совокупности меньше среднего по второй совокупности:

$$f(Y_1, Y_2,...,Y_n) < f(Z_1, Z_2,...,Z_n).$$

Тогда согласно теории измерений для устойчивости результата сравнения средних необходимо, чтобы для любого допустимого преобразования $g$ из группы допустимых преобразований в соответствующей шкале было справедливо также неравенство

$$f(g(Y_1), g(Y_2),...,g(Y_n)) < f(g(Z_1), g(Z_2),...,g(Z_n)),$$

т.е. среднее преобразованных значений из первой совокупности также было меньше среднего преобразованных значений для второй совокупности. Причем сформулированное условие должно быть верно для любых двух совокупностей $Y_1, Y_2,...,Y_n$ и $Z_1, Z_2,...,Z_n$. И, напомним, для любого допустимого преобразования. Средние величины, удовлетворяющие сформулированному условию, назовем допустимыми (в соответствующей шкале). Согласно теории измерений только такими средними можно пользоваться при анализе мнений экспертов и иных данных, измеренных в рассматриваемой шкале.

С помощью математической теории, развитой в монографии [7], удается описать вид допустимых средних в основных шкалах. Сразу ясно, что для данных, измеренных в шкале наименований, в качестве среднего годится только мода.

**Средние величины в порядковой шкале.** Рассмотрим обработку, для определенности, мнений экспертов, измеренных в порядковой шкале. Справедливо следующее утверждение.

*Теорема 1.* Из всех средних по Коши допустимыми средними в порядковой шкале являются только члены вариационного ряда (порядковые статистики).

Теорема 1 справедлива при условии, что среднее $f(X_1, X_2,...,X_n)$ является непрерывной (по совокупности переменных) и симметрической функцией. Последнее означает, что при перестановке

аргументов значение функции $f(X_1, X_2,...,X_n)$ не меняется. Это условие является вполне естественным, ибо среднюю величину мы находим для *совокупности (множества)*, а не для *последовательности*. Множество не меняется в зависимости от того, в какой последовательности мы перечисляем его элементы.

Согласно теореме 1 в качестве среднего для данных, измеренных в порядковой шкале, можно использовать, в частности, медиану (при нечетном объеме выборки). При четном же объеме следует применять один из двух центральных членов вариационного ряда - как их иногда называют, левую медиану или правую медиану. Моду тоже можно использовать - она всегда является членом вариационного ряда. Но никогда нельзя рассчитывать среднее арифметическое, среднее геометрическое и т.д.

Приведем численный пример, показывающий некорректность использования среднего арифметического $f(X_1, X_2) = (X_1 + X_2)/2$ в порядковой шкале. Пусть $Y_1= 1$, $Y_2 = 11$, $Z_1= 6$, $Z_2= 8$. Тогда $f(Y_1, Y_2) = 6$, что меньше, чем $f(Z_1, Z_2) = 7$. Пусть строго возрастающее преобразование $g$ таково, что $g(1) = 1$, $g(6) = 6$, $g(8) = 8$, $g(11) = 99$. Таких преобразований много. Например, можно положить $g(x) = x$ при $x$, не превосходящих 8, и $g(x) = 99(x-8)/3 + 8$ для $x$, больших 8. Тогда $f(g(Y_1), g(Y_2)) = 50$, что больше, чем $f(g(Z_1), g(Z_2)) = 7$. Как видим, в результате допустимого, т.е. строго возрастающего преобразования шкалы упорядоченность средних величин изменилась.

Таким образом, теория измерений выносит жесткий приговор среднему арифметическому - использовать его с порядковой шкале нельзя. Однако же те, кто не знает теории измерений, используют его. Всегда ли они ошибаются? Оказывается, можно в какой-то мере реабилитировать среднее арифметическое, если перейти к вероятностной постановке и к тому же удовлетвориться результатами для больших объемов выборок. В монографии [7] получено также следующее утверждение.

*Теорема 2.* Пусть $Y_1, Y_2, \dots, Y_m$ - независимые одинаково распределенные случайные величины с функцией распределения $F(x)$, а $Z_1, Z_2, \dots, Z_n$ - независимые одинаково распределенные случайные величины с функцией распределения $H(x)$, *причем выборки* $Y_1, Y_2, \dots, Y_m$ и $Z_1, Z_2, \dots, Z_n$ независимы между собой и $MY_1 > MZ_1$ . Для того, чтобы вероятность события

$$\{\omega : \frac{g(Y_1) + g(Y_2) + ... + g(Y_m)}{m} > \frac{g(Z_1) + g(Z_2) + ... + g(Z_n)}{n}\}$$

стремилась к 1 при $\min(m,n) \to \infty$ для любой строго возрастающей непрерывной функции $g$, удовлетворяющей условию

$$\overline{\lim_{|x|\to\infty}} |\frac{g(x)}{x}| < \infty,$$

необходимо и достаточно, чтобы при всех $x$ выполнялось неравенство $F(x) \geq H(x)$, причем существовало число $x_0$ , *для которого* $F(x_0) < H(x_0)$.

*Примечание.* Условие с верхним пределом носит чисто внутриматематический характер. Фактически функция $g$ - произвольное допустимое преобразование в порядковой шкале.

Согласно теореме 2 средним арифметическим можно пользоваться и в порядковой шкале, если сравниваются выборки из двух распределений, удовлетворяющих приведенному в теореме неравенству. Проще говоря, одна из функций распределения должна всегда лежать над другой. Функции распределения не могут пересекаться, им разрешается только касаться друг друга. Это условие выполнено, например, если функции распределения отличаются только сдвигом, т.е.

$$F(x) = H(x+b)$$

при некотором $b$. Последнее условие выполняется, если два значения некоторой величины измеряются с помощью одного и того же средства измерения, у которого распределение погрешностей не меняется при переходе от измерения одного значения рассматриваемой величины к измерению другого.

**Средние по Колмогорову.** Естественная система аксиом (требований к средним величинам) приводит к так называемым ассоциативным средним. Их общий вид нашел в 1930 г.

А.Н.Колмогоров [8]. Теперь их называют «средними по Колмогорову». Они являются обобщением нескольких из перечисленных выше средних.

Для чисел $X_1, X_2,...,X_n$ среднее по Колмогорову вычисляется по формуле

$$G\{(F(X_1) + F(X_2)+... + F(X_n))/n\},$$

где $F$ - строго монотонная функция (т.е. строго возрастающая или строго убывающая), $G$ - функция, обратная к $F$. Среди средних по Колмогорову - много хорошо известных персонажей. Так, если $F(x) = x$, то среднее по Колмогорову - это среднее арифметическое, если $F(x) = \ln x$, то среднее геометрическое, если $F(x) = 1/x$, то среднее гармоническое, если $F(x) = x^2$, то среднее квадратическое, и т.д. (в последних трех случаях усредняются положительные величины). Среднее по Колмогорову - частный случай среднего по Коши. С другой стороны, такие популярные средние, как медиана и мода, нельзя представить в виде средних по Колмогорову. В монографии [7] доказаны следующие утверждения.

*Теорема 3.* При справедливости некоторых внутриматематических условий регулярности в шкале интервалов из всех средних п Колмогоров допустимым является только среднее арифметическое.

Таким образом, среднее геометрическое или среднее квадратическое температур (в шкале Цельсия), потенциальных энергий или координат точек не имеют смысла. В качестве среднего надо применять среднее арифметическое. А также можно использовать медиану или моду.

*Теорема 4.* При справедливости некоторых внутриматематических условий регулярности в шкале отношений из всех средних по Колмогорову допустимыми являются только степенные средние с $F(x) = x^c$, $c \neq 0$, и среднее геометрическое.

*Замечание.* Среднее геометрическое является пределом степенных средних при $c \to 0$.

Есть ли средние по Колмогорову, которыми нельзя пользоваться в шкале отношений? Конечно, есть. Например, с $F(x) = e^x$.

Аналогично средним величинам могут быть изучены и другие статистические характеристики - показатели разброса, связи, расстояния и др. (см., например, [7]). Нетрудно показать, например, что коэффициент корреляции не меняется при любом допустимом преобразовании в шкале интервалов, как и отношение дисперсий. Дисперсия не меняется в шкале разностей, коэффициент вариации - в шкале отношений, и т.д.

Приведенные выше результаты о средних величинах широко применяются, причем не только в экономике, менеджменте, теории экспертных оценок или социологии, но и в инженерном деле, например, для анализа методов агрегирования датчиков в АСУ ТП доменных печей. Велико прикладное значение теории измерений в задачах стандартизации и управления качеством, в частности, в квалиметрии. Здесь есть и интересные теоретические результаты. Так, например, любое изменение коэффициентов весомости единичных показателей качества продукции приводит к изменению упорядочения изделий по средневзвешенному показателю (эта теорема доказана проф. В.В. Подиновским).

При подготовке и принятии решений необходимо использовать только инвариантные алгоритмы обработки данных. В настоящем подразделе показано, что требование инвариантности выделяет из многих алгоритмов усреднения лишь некоторые, соответствующие используемым шкалам измерения. Инвариантные алгоритмы в общем случае рассматриваются в математической теории измерений [9]. Нацеленное на прикладные исследования изложение теории измерений дается в монографиях [7, 10].

### 2.1.4. Вероятностные модели порождения нечисловых данных

Рассмотрим основные вероятностные модели порождения нечисловых данных. А именно, дихотомических данных, результатов парных сравнений, бинарных отношений, рангов, объектов общей природы. Обсудим различные варианты вероятностных моделей и их практическое использование (см. также обзор [11]).

**Дихотомические данные.** Рассмотрим базовую вероятностную модель дихотомических данных - *бернуллиевский вектор* (в терминологии энциклопедии [12] - *люсиан*), т.е. конечную последовательность $X = (X_1, X_2, ..., X_k)$ независимых испытаний Бернулли $X_i$, для которых $P(X_i = 1) = p_i$ и $P(X_i = 0) = p_i$, $i = 1,2,...,k,$ причем вероятности $p_i$ могут быть различны.

Бернуллиевские вектора часто применяются при практическом использовании эконометрических методов. Так, они использованы в монографии [7] для описания равномерно распределенных случайных толерантностей. Как известно, толерантность на множестве из $m$ элементов можно задать симметричной матрицей $\| \delta_{ij} \|$ из 0 и 1, на главной диагонали которой стоят 1. Тогда случайная толерантность описывается распределением $m(m-1)/2$ дихотомических случайных величин $\delta_{ij}$, $1 \le i \prec j \le m,$ а для равномерно распределенной (на множестве всех толерантностей) толерантности эти случайные величины, как можно доказать, оказываются независимыми и принимают значения 0 и 1 с равными вероятностями 1/2. Записав элементы $\delta_{ij}$ задающей такую толерантность матрицы в строку, получим бернуллиевский вектор с $k=m(m-1)/2$ и $p_i = 1/2,$ $i = 1,2,...,k.$

В связи с оцениванием по статистическим данным функции принадлежности нечеткой толерантности в 1970-е годы была построена теория случайных толерантностей с такими независимыми $\delta_{ij}$, что вероятности $P(\delta_{ij} = 1) = p_{ij}$ произвольны [7]. Случайные множества с независимыми элементами использовались как общий язык для описания парных сравнений и случайных толерантностей. В некоторых публикациях термин "люсиан" применялся как сокращение для выражения "случайные множества с независимыми элементами".

Был выявлен ряд областей, в которых полезен математический аппарат решения различных статистических задач, связанных с бернуллиевскими векторами. Перечислим эти области, включая ранее названные: анализ случайных толерантностей; случайные множества с независимыми элементами; обработка результатов независимых парных сравнений; статистические методы анализа точности и стабильности технологического процесса, а также анализ и синтез планов статистического приемочного контроля (по альтернативным, т.е. дихотомическим, признакам); обработка маркетинговых и социологических анкет (с закрытыми вопросами типа "да" - "нет"); обработка социально-психологических и медицинских данных, в частности, ответов на психологические тесты типа MMPI (используемых в задачах управления персоналом), топографических карт (применяемых для анализа и прогноза зон поражения при технологических авариях, распространении коррозии, распространении экологически вредных загрязнений в других ситуациях) и т.д.

Теорию бернуллиевских векторов можно выразить в терминах любой из этих теоретических и прикладных областей. Однако терминология одной из этих областей "режет слух" и приводит к недоразумениям в другой из них. Поэтому целесообразно использовать термин "бернуллиевский вектор" в указанном выше значении, не связанном ни с какой из перечисленных областей приложения этой теории (в ряде публикаций в том же значении использовался термин "люсиан").

Распределение бернуллиевского вектора $X$ полностью описывается вектором $P = (p_1, p_2, ..., p_k)$,т.е. нечетким подмножеством множества $\{1,2,...,k\}$. Действительно, для любого детерминированного вектора $x = (x_1, x_2, ..., x_k)$ из 0 и 1 имеем

$$P(X = x) = \prod_{1 \le j \le k} h(x_j, p_j),$$

где $h(x,p)=p$ при $x = 1$ и $h(x,p)=1 - p$ при $x=0$.

Теперь можно уточнить способы использования люсианов в прикладной статистике. Бернуллиевскими векторами можно моделировать: результаты статистического контроля (0 - годное изделие, 1 - дефектное); результаты маркетинговых и социологических опросов (0 - опрашиваемый выбрал первую из двух подсказок, 1 - вторую); распределение посторонних

включений в материале (0 - нет включения в определенном объеме материала, 1 - есть); результаты испытаний и анализов (0 - нет нарушений требований нормативно-технической документации, 1 - есть такие нарушения); процессы распространения, например, пожаров (0 - нет загорания, 1 - есть; подробнее см. [7, с.215-223]); технологические процессы (0 - процесс находится в границах допуска, 1 - вышел из них); ответы экспертов (опрашиваемых) о сходстве объектов (проектов, образцов) и т.д.

**Парные сравнения.** Общую модель парных сравнений опишем согласно монографии Г. Дэвида [13, с.9]. Предположим, что $t$ объектов $A_1, A_2, ..., A_t$ сравниваются попарно каждым из $n$ экспертов. Всего возможных пар для сравнения имеется $s = t(t-1)/2$. Эксперт с номером $\gamma$ делает $r_\gamma$ повторных сравнений для каждой из $s$ возможностей. Пусть $X(i, j, \gamma, \delta)$, $i,j=1,2,...,t$, $i \neq j$, $\gamma=1,2,...,n$; $\delta=1,2,...,r_\gamma$, - случайная величина, принимающая значение 1 или 0 в зависимости от того, предпочитает ли эксперт $\gamma$ объект $A_i$ или объект $A_j$ в $\delta$-м сравнении двух объектов. Предполагается, что все сравнения проводятся независимо друг от друга, так что случайные величины $X(i, j, \gamma, \delta)$ независимы в совокупности, если не считать того, что $X(i, j, \gamma, \delta) + X(j, i, \gamma, \delta) = 1$. Положим

$$P(X(i, j, \gamma, \delta) = 1) = \pi(i, j, \gamma, \delta).$$

Ясно, что описанная модель парных сравнений представляет собой частный случай бернуллиевского вектора. В этой модели число наблюдений равно числу неизвестных параметров, поэтому для получения статистических выводов необходимо наложить априорные условия на $\pi(i, j, \gamma, \delta)$, например [13, с.9]:

$\pi(i, j, \gamma, \delta) = \pi(i, j, \gamma)$ (нет эффекта от повторений);

$\pi(i, j, \gamma, \delta) = \pi(i, j)$ (нет эффекта от повторений и от экспертов).

Теорию независимых парных сравнений целесообразно разделить на две части - непараметрическую, в которой статистические задачи ставятся непосредственно в терминах $\pi(i, j, \gamma, \delta)$, и параметрическую, в которой вероятности $\pi(i, j, \gamma, \delta)$ выражаются через меньшее число иных параметров. Ряд результатов непараметрической теории парных сравнений непосредственно вытекает из теории бернуллиевских векторов.

В параметрической теории парных сравнений наиболее популярна так называемая линейная модель [13, с.11], в которой предполагается , что каждому объекту $A_i$ можно сопоставить некоторую "ценность" $V_i$ так, что вероятность предпочтения $\pi(i, j)$ (т.е. предполагается дополнительно, что эффект от повторений и от экспертов отсутствует) выражается следующим образом:

$$\pi(i, j) = H(V_i - V_j), \qquad (1)$$

где *H(x)* - функция распределения, симметричная относительно 0, т.е.

$$H(-x) = 1 - H(x) \qquad (2)$$

при всех *x*.

Широко применяются модели Терстоуна - Мостеллера и Брэдли - Терри, в которых *H(x)* - соответственно функции нормального и логистического распределений. Поскольку функция *Ф(x)* стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1 и функция

$$\Psi(x) = e^x (1 + e^x)^{-1}$$

стандартного логистического распределения удовлетворяют (см., например, [14]) соотношению

$$\sup_{x \in R^1} |\Phi(x) - \Psi(1{,}7x)| < 0{,}01,$$

то для обоснованного выбора по статистическим данным между моделями Терстоуна-Мостеллера и Брэдли-Терри необходимо не менее тысячи наблюдений.

Соотношение (1) вытекает из следующей модели поведения эксперта: он измерят "ценность" $V_i$ и $V_j$ объектов $A_i$ и $A_j$, но с ошибками $\varepsilon_i$ и $\varepsilon_j$ соответственно, а затем сравнивает свои оценки ценности объектов $y_i = V_i + \varepsilon_i$ и $y_j = V_j + \varepsilon_j$. Если $y_i > y_j$, то он предпочитает $A_i$, в противном случае - $A_j$. Тогда

$$\pi(i, j) = P(\varepsilon_i - \varepsilon_j < V_i - V_j) = H(V_i - V_j). \qquad (3)$$

Обычно предполагают, что субъективные ошибки эксперта $\varepsilon_i$ и $\varepsilon_j$ независимы и имеют одно и то же непрерывное распределение. Тогда функция распределения $H(x)$ из соотношения (3) непрерывна и удовлетворяет функциональному уравнению (2).

Существует много разновидностей моделей парных сравнений, постоянно предполагаются новые. В качестве примера опишем модель парных сравнений, основанную не на процедуре упорядочения, а на определении сходства объектов. Пусть каждому объекту $A_i$ соответствует точка $a_i$ в $r$-мерном евклидовом пространстве $R^r$. Эксперт "измеряет" $a_i$ и $a_j$ с ошибками $\varepsilon_i$ и $\varepsilon_j$ соответственно и в случае, если евклидово расстояние между $a_i + \varepsilon_i$ и $a_j + \varepsilon_j$ меньше 1, заявляет о сходстве объектов $A_i$ и $A_j$, в противном случае - об их различии. Предполагается, что ошибки $\varepsilon_i$ и $\varepsilon_j$ независимы и имеют одно и то же распределение, например, круговое нормальное распределение с нулевым математическим ожиданием и дисперсией координат $\sigma^2$. Целью статистической обработки является определение по результатам парных сравнений оценок параметров $a_1, a_2, \ldots, a_r$, и $\sigma^2$, а также проверка согласия опытных данных с моделью.

Рассмотренные модели парных сравнений могут быть обобщены в различных направлениях. Так, можно ввести понятие "ничья "- ситуации, когда эксперт оценивает объекты одинаково. Модели с учетом "ничьих" предполагают, что эксперт может отказаться от выбора одного из объектов и заявить об их эквивалентности, т. е. число возможных ответов увеличивается с 2 до 3. В моделях множественных сравнений эксперту представляется не два объекта , а три или большее число

Модели, учитывающие "ничьи", строятся обычно с помощью используемых в психофизике "порогов чувствительности": если $| y_i - y_j | \le r$ (где $r$ - порог чувствительности), то объекты $A_i$ и $A_j$ эксперт объявляет неразличимыми. Приведем пример модели с "ничьими", основанной на другом принципе. Пусть каждому объекту $A_i$ соответствует точка $a_i$ в $r$-мерном линейном пространстве. Как и прежде, эксперт "измеряет " объектные точки " $a_i$ и $a_j$ с ошибками $\varepsilon_i$ и $\varepsilon_j$ соответственно, т.е. принимает решение на основе $y_i = a_i + \varepsilon_i$ и $y_j = a_j + \varepsilon_j$. Если все координаты $y_i$ больше соответствующих координат $y_j$ , то $A_i$ предпочитается $A_j$. Соответственно, если каждая координата $y_i$ меньше координаты $y_j$ с тем же номером, то эксперт считает наилучшим объект $A_j$. Во всех остальных случаях эксперт объявляет о ничейной ситуации. Эта модель при $r = 1$ переходит в описанную выше линейную модель. Она связана с принципом Парето в теории группового выбора и предусматривает выбор оптимального по Парето объекта, если он существует (роль согласуемых критериев играют процедуры сравнения значений отдельных координат), и отказ от выбора, если такого объекта нет.

Можно строить модели, учитывающие порядок предъявления объектов при сравнении, зависимость результата сравнения от результатов предшествующих сравнений. Опишем одну из подобных моделей.

Пусть эксперт сравнивает три объекта - *A, B, C,* причем сначала сравниваются $A$ и $B$, потом - $B$ и $C$ и, наконец, $A$ и $C$. Для определенности пусть $A > B$ будет означать, что $A$ более предпочтителен, чем $B$. Пусть при предъявлении двух объектов

$$P(A > B) = \pi_{AB}, P(B > C) = \pi_{BC}, P(A > C) = \pi_{AC}.$$

Теперь пусть пара *B, C* предъявляется после пары *A, B*. Естественно предположить, что высокая оценка $B$ в первом сравнении повышает вероятность предпочтения $B$ и во втором, и, наоборот,

отрицательное мнение о $B$ в первом сравнении сохраняется и при проведении второго сравнения. Это предположение проще всего учесть в модели следующим образом:

$$P(B > C \mid B > A) = \pi_{BC} + \delta, \quad P(B > C \mid A > B) = \pi_{BC} - \delta,$$

где $\delta$ - некоторое положительное число, показывающее степень влияния первого сравнения на второе. По аналогичным причинам вероятности исхода третьего сравнения в зависимости от результатов первых двух можно описать так:

$$P(A > C \mid A > B, B > C) = \pi_{AC} + 2\delta, \quad P(A > C \mid A > B, B < C) = \pi_{AC},$$

$$P(A > C \mid A < B, B > C) = \pi_{AC}, \quad P(A > C \mid A < B, B < C) = \pi_{AC} - 2\delta.$$

Статистическая задача состоит в определении параметров $\pi_{AB}$, $\pi_{BC}$, $\pi_{AC}$ и $\delta$ по результатам сравнений, проведенных $n$ экспертами, и в проверке адекватности модели.

Ясно, что можно рассматривать и другие модели, в частности, учитывающие тягу экспертов к транзитивности ответов. Очевидно, что проблемы построения моделей парных сравнений относятся не к прикладной статистике, а к тем прикладным областям, для решения задач которых развиваются методы парных сравнений, например, к экономике предприятия, стратегическому менеджменту, производственной психологии, изучению поведения потребителей, экспертным оценкам и т. д.

Метод парных сравнений был введен в 1860 г. Г. Т. Фехнером для решения задач психофизики. Расскажем об этом несколько подробнее. Как известно, основателем психофизики по праву считается Густав Теодор Фехнер (1801 - 1887), а год выхода в свет его фундаментальной работы "Элементы психофизики" (1860) - датой рождения новой науки. В этой работе широко применялся предложенный Г.Т. Фехнером метод парных сравнений (обсуждение событий тех лет с современных позиций дано в монографии [13, с.14-16]).

С точки зрения математической статистики приведенные выше модели не представляют большого теоретического интереса: оценки параметров находятся обычно методом максимального правдоподобия, а проверка согласия проводится по критерию отношения правдоподобия или асимптотически эквивалентными ему критериями типа хи-квадрат [13]. Вычислительные процедуры обычно сложны и плохо исследованы; их можно упростить и одновременно повысить обоснованность, перейдя от оценок максимального правдоподобия к одношаговым оценкам (см. главу 2.2).

Отметим некоторые сложности при обосновании возможности использовании линейных моделей типа (1) - (3). Вероятностно-статистическая теория достаточно проста, когда предполагается, что каждому отдельному сравнению двух объектов соответствуют свои собственные ошибки экспертов, причем все ошибки независимы в совокупности. Однако это предположение отнюдь не очевидно с содержательной точки зрения. В качестве примера рассмотрим три объекта *A, B* и *C*, которые сравнивают попарно: *A* и *B*, *B* и *C*, *A* и *C*. В соответствии со сказанным, в рассмотрение вводят 6 ошибок одного и того же эксперта: $\varepsilon_A$ и $\varepsilon_B$ в первом сравнении, $\varepsilon_B'$ и $\varepsilon_C$-во втором, $\varepsilon_A'$ и $\varepsilon_C'$- в третьем, причем все эти 6 случайных величин независимы в совокупности. Между тем естественно думать, что мнения эксперта об одном и том же объекте связаны между собой, т. е. $\varepsilon_A$ и $\varepsilon_A'$ зависимы, равно как $\varepsilon_B$ и $\varepsilon_B'$, а также $\varepsilon_C$ и $\varepsilon_C'$. Более того, если принять, что точка зрения эксперта полностью определена для него самого, то следует положить $\varepsilon_A = \varepsilon_A'$ и соответственно $\varepsilon_B = \varepsilon_B'$ и $\varepsilon_C = \varepsilon_C'$. При этом, напомним, случайные величины $\varepsilon_A$, $\varepsilon_B$ и др. интерпретируется как отклонения мнений отдельных экспертов от истины. Видимо, ошибку эксперта целесообразно считать состоящей из двух слагаемых, а именно: отклонения от истины, вызванного внутренними особенностями эксперта (систематическая погрешность) и колебания мнения эксперта в связи с очередным парным сравнением (случайная погрешность). Игнорирование систематической погрешности облегчает развитие математико-

статистической теории, а ее учет приводит к необходимости изучения зависимых парных сравнений.

При обработке результатов парных сравнений первый этап - проверка согласованности. Понятие согласованности уточняется различными способами, но все они имеют один и тот же смысл проверки однородности обрабатываемого материала, т.е. того, что целесообразно агрегировать мнения отдельных экспертов, объединить данные и совместно их обрабатывать. При отсутствии однородности данные разбиваются на группы (классы, кластеры, таксоны) с целью обеспечения однородности внутри отдельных групп. Естественно, согласованность целесообразно проверять, вводя возможно меньше гипотез о структуре данных. Следовательно, целесообразно пользоваться для этого непараметрической теорией парных сравнений, основанной на теории бернуллиевских векторов.

Хорошо известно, что модели парных сравнений с успехом применяются в экспертных и экспериментальных процедурах упорядочивания и выбора. В частности, для анализа голосований, турниров, выбора наилучшего объекта (проекта, образца, кандидатуры); в планировании и анализе сравнительных экспериментов и испытаний; в органолептической экспертизе (в частности, дегустации); при изучении поведения потребителей; визуальной колоритмии, определении индивидуальных рейтингов и вообще изучении предпочтений при выборе и т. д. (подробнее см. [7, 13]).

**Бинарные отношения.** Теорию ранговой корреляции (см. главу 3.2) можно рассматривать как теорию статистического анализа случайных ранжировок, равномерно распределенных на множестве всех ранжировок. Так, при обработке данных классического психофизического эксперимента по упорядочению кубиков соответственно их весу, подробно описанного в работе [15], оказалась адекватной следующая т.н. *Т*-модель ранжирования.

Пусть имеется $t$ объектов $A_1, A_2, ..., A_t$, причем каждому объекту $A_i$ соответствует число $a_i$, описывающее его положение на шкале изучаемого признака. Испытуемый упорядочивает объекты так, как если бы оценивал соответствующие им значения с ошибками, т.е. находил $y_i = a_i + \varepsilon_i$, $i$=1,2,…$n$, где $\varepsilon_i$ - ошибка при рассмотрении $i$-го объекта, а затем располагал бы объекты в том порядке, в каком располагаются $y_1, y_2, ..., y_t$. В этом случае вероятность появления упорядочения $A_{i1}, A_{i2}, ...A_{it}$ есть $P(y_{i1} < y_{i2} < ... < y_{it})$, а ранги $R_1, R_2, ...R_t$ объектов являются рангами случайных величин $y_1, y_2, ..., y_t.$, полученными при их упорядочении в порядке возрастания. Кроме того, для простоты расчетов в модели предполагается, что ошибки испытуемого $\varepsilon_1, \varepsilon_2, ..., \varepsilon_t$ независимы и имеют нормальное распределение с математическим ожиданием 0 и дисперсией $\sigma^2$.

Как уже отмечалось в главе 1.1, бинарное отношение на множестве из $t$ элементов полностью описывается матрицей из 0 и 1 порядка $t \times t$. Поэтому задать распределение случайного бинарного отношения - это то же самое, что задать распределение вероятностей на множестве всех матриц описанного вида, состоящем из $2^{(t^2)}$ элементов. Пространства ранжировок, разбиений, толерантностей зачастую удобно считать подпространствами пространства всех бинарных отношений, тогда распределения вероятностей на них - частные случаи описанного выше распределения, выделенные тем, что вероятности принадлежности соответствующим подпространствам равны 1. Распределение произвольного бинарного отношения описывается $2^{(t^2)}$ - 1 параметрами, распределение случайной ранжировки (без связей) - ($t!$ - 1) параметрами, а описанная выше *Т*-модель ранжирования - ($t$ + 1) параметром. При $t$ = 4 эти числа равны соответственно 65535, 23 и 5. Первое из этих чисел показывает практическую невозможность использования в вероятностно-статистических моделях произвольных бинарных отношений, поскольку по имеющимся данным невозможно оценить столь большое число параметров. Приходится ограничиваться теми или иными семействами бинарных отношений - ранжировками, разбиениями, толерантностями и др. Модель произвольной случайной ранжировки при $t$ = 5

описывается 119 параметрами, при $t$ = 6 - уже 719 параметрами, при $t$ = 7 число параметром достигает 5049, что уже явно за возможностями оценивания. В то же время $T$-модель ранжирования при $t$ = 7 описывается всего 8-ю параметрами, а потому может быть кандидатом для практического использования.

Что естественно предположить относительно распределения случайного элемента со значениями в том или ином пространстве бинарных отношений? Зачастую целесообразно считать, что распределение имеет некий центр, попадание в который наиболее вероятно, а по мере удаления от центра вероятности убывают. Это соответствует естественной модели измерения с ошибкой; в классическом одномерном случае результат подобного измерения обычно описывается унимодальной симметричной плотностью, монотонно возрастающей слева от модального значения, в котором плотность максимальна, и монотонно убывающей справа от него. Чтобы ввести понятие монотонного распределения в пространстве бинарных отношений, будем исходить из метрики в этом пространстве. Воспользовавшись тем, что бинарные отношения $C$ и $D$ однозначно описываются матрицами $\| c_{ij} \|$ и $\| d_{ij} \|$ порядка $t \times t$ соответственно, рассмотрим расстояние (в несколько другой терминологии - метрику) в пространстве бинарных отношений

$$d(C,D) = \sum_{1 \le i,j \le t} | c_{ij} - d_{ij} |. \qquad (4)$$

Метрика (4) в различных пространствах бинарных отношений - ранжировок, разбиений, толерантностей - может быть введена с помощью соответствующих систем аксиом (см. главу 1.1). В настоящее время метрику (4) обычно называют расстоянием Кемени в честь американского исследователя Джона Кемени, впервые получившего эту метрику исходя из предложенной им системы аксиом для расстояния между упорядочениями (ранжировками).

В статистике нечисловых данных используются и иные метрики, отличающиеся от расстояния Кемени. Более того, для использования понятия монотонного распределения, о котором сейчас идет речь, нет необходимости требовать выполнения неравенства треугольника, а достаточно, чтобы $d(C,D)$ можно было рассматривать как показатель различия. Под показателем различия понимаем такую функцию $d(C,D)$ двух бинарных отношений $C$ и $D$, что $d(C,D)$= 0 при $C = D$ и увеличение $d(C,D)$ интерпретируется как возрастание различия между $C$ и $D$.

*Определение 1*. Распределение бинарного отношения $X$ называется монотонным с центром в $C_0$ относительно расстояния (показателя различия) $d$, если из $d(C,C_0) < d(D,C_0)$ следует, что $P(X=C) > P(X=D)$.

Это определение впервые введено в монографии [7, с.196]. Оно может использоваться в любых пространствах бинарных отношений и, более того, в любых пространствах из конечного числа элементов, лишь бы в них была введена функция $d(C,D)$ - показатель различия элементов $C$ и $D$ этого пространства. Монотонное распределение унимодально, мода находится в $C_0$.

*Определение 2*. Распределение бинарного отношения $X$ называется симметричным относительно расстояния $d$ с центром в $C_0$, если существует такая функция $f : R^1_+ \to [0,1]$, что

$$P(X = C) = f(d(C,C_0)). \qquad (5)$$

Если распределение $X$ монотонно и таково, что из $d(C,C_0) = d(D,C_0)$ следует $P(X=C) = P(X=D)$, то оно симметрично. Если функция $f$ в формуле (5) монотонно строго убывает, то соответствующее распределение монотонно в смысле определения 1.

Поскольку толерантность на множестве из $t$ элементов задается $0{,}5t(t - 1)$ элементами матрицы из 0 и 1 порядка $t \times t$, лежащими выше главной диагонали, то распределение на множестве толерантностей задается в общем случае $2^{0,5t(t-1)}$ параметрами. Естественно выделить семейство распределений, соответствующее независимым элементам матрицы. Оно задается бернуллиевским вектором (люсианом) с $0{,}5t(t - 1)$ параметрами (выше бернуллиевские вектора рассмотрены подробнее). Математическая техника, необходимая для изучения толерантностей с независимыми элементами, существенно проще, чем в случае ранжировок и разбиений. Здесь легко отказаться от условия равномерности распределения. Этому условию соответствует $p_{ij} \equiv 1/2$, в то

время как статистические методы анализа люсианов, развитые в статистике нечисловых данных (см., например, работы [7, 16, 17]) не налагают никаких существенных ограничений на $p_{ij}$ .

Как уже отмечалось, при обработке мнений экспертов сначала проверяют согласованность. В частности, если мнения экспертов описываются монотонными распределениями, то для согласованности необходимо совпадение центров этих распределений. К сожалению, рассмотренные выше классические методы проверки согласованности для ранжировок, основанные на коэффициентах ранговой корреляции и конкордации, позволяют лишь отвергнуть гипотезу о равнораспределенности. Но не установить, можно ли считать, что центры соответствующих экспертам распределений совпадают или же, например, существует две группы экспертов, каждая со своим центром. Теория случайных толерантностей лишена этого недостатка. Отсюда вытекают следующие практические рекомендации.

Пусть цель обработки экспертных данных состоит в получении ранжировки, отражающей групповое мнение. Однако согласно рекомендуемой процедуре экспертного опроса пусть эксперты не упорядочивают объекты, а проводят парные сравнения, сравнивая каждый из рассматриваемых объектов со всеми остальными, причем ровно один раз. Тогда ответ эксперта - толерантность, но, вообще говоря, не ранжировка, поскольку в ответах эксперта может нарушаться транзитивность.

Возможны два пути обработки данных. Первый - превратить ответ эксперта в ранжировку (тем или иным способом "спроектировав" его на пространство ранжировок), а затем проверять согласованность ранжировок с помощью известных критериев. При этом от толерантности перейти к ранжировке можно, например, так. Будем выбирать ближайшую (в смысле применяемого расстояния) матрицу к матрице ответов эксперта из всех, соответствующих ранжировкам без связей.

Второй путь - проверить согласованность случайных толерантностей, а групповое мнение искать с помощью медианы Кемени (подробнее см. ниже) непосредственно по исходным данным, т.е. по толерантностям. Групповое мнение при этом может быть найдено в пространстве ранжировок. Второй путь мы считаем более предпочтительным, поскольку при этом обеспечивается более адекватная проверка согласованности и исключается процедура укладывания мнения эксперта в «прокрустово ложе» ранжировки (эта процедура может приводить как к потере информации, так и к принципиально неверным выводам, вызванным искажениями мнений экспертов).

Области применения статистики бинарных отношений многообразны: ранговая корреляция - оценка величины связи между переменными, измеренными в порядковой шкале; анализ экспертных или экспериментальных упорядочений; анализ разбиений технико-экономических показателей на группы сходных между собой; обработка данных о сходстве (взаимозаменяемости); статистический анализ классификаций; математические вопросы теории менеджмента и др.

**Случайные множества.** Будем рассматривать случайные подмножества некоторого множества $Q$. Если $Q$ состоит из конечного числа элементов, то считаем, что случайное подмножество $S$ - это случайный элемент со значениями в $2^Q$ - множестве всех подмножеств множества $Q$, состоящем из $2^{card(Q)}$ элементов. Чтобы удовлетворить математиков, считаем, что все подмножества $Q$ измеримы. Тогда распределение случайного подмножества $S = S(\omega)$ множества $Q$ - это

$$P_S(A) = P(S = A) = \mathrm{P}(\{\omega : S(\omega) = A\}), A \subseteq Q. \qquad (6)$$

В формуле (6) предполагается, что $S : \Omega \to 2^Q$, где $(\Omega, F, \mathrm{P})$ - вероятностное пространство (здесь $\Omega$ - пространство элементарных событий, $F$ - $\sigma$-алгебра случайных событий, P - вероятностная мера на $F$), на котором определен случайный элемент $S(\omega)$. Через распределение $P_S(A)$ выражаются вероятности различных событий, связанных с $S$. Так, чтобы найти вероятность накрытия фиксированного элемента $q$ случайным множеством $S$, достаточно вычислить

$$P(q \in S) = \mathrm{P}(\{\omega : q \in S(\omega)\}) = \sum_{A: q \in A, A \subseteq 2^Q} P(S = A),$$

где суммирование идет по всем подмножествам $A$ множества $Q$, содержащим $q$. Пусть $Q=\{q_1, q_2, \ldots, q_k\}$. Рассмотрим случайные величины, определяемые по случайному множеству $S$ следующим образом

$$\chi_i(\omega) = \begin{cases} 1, q_i \in S(\omega), \\ 0, q_i \notin S(\omega). \end{cases}$$

*Определение 3*. Случайное множество $S$ называется случайным множеством с независимыми элементами, если случайные величины $\chi_i(\omega), i = 1,2,...,k,$ независимы (в совокупности).

Последовательность случайных величин $\chi_1, \chi_2, ..., \chi_k$ --бернуллиевский вектор с $X_i = \chi_i$ и $p_i = P(q_i \in S(\omega)), i = 1,2,...,k.$ Из сказанного выше следует, что распределение случайного множества с независимыми элементами задается формулой

$$P(S = A) = \prod_{q_i \in A} p_i \prod_{q_i \in Q \setminus A} (1 - p_i),$$

т.е. такие распределения образуют $k = card(Q)$ - мерное параметрическое семейство, входящее в $(2^{card(Q)} - 1)$ - одномерное семейство всех распределений случайных подмножеств множества $Q$.

При исследовании случайных подмножеств произвольного множества $Q$ будем рассматривать их как случайные величины со значениями в некотором пространстве подмножеств множества $Q$, например, в пространстве замкнутых подмножеств $2^Q$ множества $Q$.

Представляющими интерес лишь для математиков способами введения измеримой структуры в $2^Q$ интересоваться не будем. Отсутствие специального интереса к проблеме измеримости связано с тем, что при вероятностно-статистическом моделировании и обработке на ЭВМ все случайные подмножества рассматриваются как конечные (т.е. подмножества конечного множества).

Случайные множества находят разнообразные применения в многообразных проблемах эконометрики и математической экономики, в том числе в задачах управлении запасами и ресурсами (см. об этом главу 5 в монографии [7]), в задачах менеджмента и, в частности, маркетинга, в экспертных оценках, например, при анализе мнений голосующих или опрашиваемых, каждый из которых отмечает несколько пунктов из списка и т.д. Кроме того, случайные множества применяются в гранулометрии, при изучении пористых сред и объектов сложной природы в таких областях, как металлография, петрография, биология, в частности, математическая морфология, в изучении структуры веществ и материалов, в исследовании процессов распространения, в том числе просачивания, распространения пожаров, экологических загрязнений, при районировании, в изучении областей поражения, например, поражения металла коррозией и сердечной мышцы при инфаркте миокарда, и т.д., и т.п. Можно вспомнить о компьютерной томографии, о наглядном представлении сложной информации на экране компьютера, об изучении распространения рекламной информации, о картах Кохонена (популярный метод представления информации при применении нейросетей) и т.д.

**Ранговые методы.** В главе 1.1 установлено, что любой адекватный алгоритм в порядковой шкале является функцией от некоторой матрицы $C$. Пусть никакие два из результатов наблюдений $x_1, x_2, \ldots, x_n$ не совпадают, а $r_1, r_2, \ldots, r_n$ - их ранги. Тогда элементы матрицы $C$ и ранги результатов наблюдений связаны взаимно однозначным соответствием:

$$r_i = 1 + \sum_{1 \le j \le n} (1 - c_{ij}),$$

а $c_{ij}$ через ранги выражаются так: $c_i = 1$, если $r_i < r_j$, и $c_{ij} = 0$ в противном случае.

Сказанное означает, что при обработке данных, измеренных в порядковой шкале, могут применяться только ранговые статистические методы. Отметим, что часто используемое в непараметрической статистике преобразование $Y = F(X)$ (здесь $F(x)$ - непрерывная функция распределения случайной величины $X$, причем $F$ предполагается произвольной) фактически означает переход к порядковой шкале, поскольку статистические выводы при этом инвариантны относительно допустимых преобразований в порядковой шкале.

Разумеется, ранговые статистические методы могут применяться не только при обработке данных, измеренных в порядковой шкале. Так, для проверки независимости двух количественных признаков в случае, когда нет уверенности в нормальности соответствующего двумерного распределения, целесообразно пользоваться коэффициентами ранговой корреляции Кендалла или Спирмена.

В настоящее время с помощью непараметрических и прежде всего ранговых методов можно решать все те задачи эконометрики и прикладной статистики, что и с помощью параметрических методов, в частности, основанных на предположении нормальности. Однако параметрические методы вошли в массовое сознание исследователей и инженеров и мешают широкому внедрению более обоснованной и прогрессивной ранговой статистики. Так, при проверке однородности двух выборок вместо критерия Стьюдента целесообразно использовать ранговые методы (см. главу 3.1), но пока это делается редко.

**Объекты общей природы.** Вероятностная модель объекта нечисловой природы в общем случае - случайный элемент со значениями в пространстве произвольного вида, а модель выборки таких объектов - совокупность независимых одинаково распределенных случайных элементов. Именно такая модель была использована для обработки наблюдений, каждое из которых - нечеткое множество [18].

Из-за имеющего разнобоя в терминологии приведем математические определения из справочника по теории вероятностей академика РАН Ю.В Прохорова и проф. Ю.А. Розанова [19].

Пусть $(\mathrm{X},\mathrm{B})$-некоторое измеримое пространство; $(F,\mathrm{B})$-измеримая функция $\xi=\xi(\omega)$ на пространстве элементарных событий $(\Omega,F,\mathrm{P})$ (где P - вероятностная мера на $\sigma$-алгебре $F$ - измеримых подмножеств $\Omega$, называемых событиями) со значениями в $(\mathrm{X},\mathrm{B})$ называется случайной величиной (чаще этот математический объект называют случайным элементом, оставляя термин "случайная величина" за частным случаем, когда X - числовая прямая) в фазовом пространстве $(\mathrm{X},\mathrm{B})$. Распределением вероятностей этой случайной величины $\xi$ называется функция $P_\xi=P_\xi(B)$ на $\sigma$-алгебре B фазового пространства, определенная как

$$P_\xi=P\{\xi\in B\}\quad (B\in\mathrm{B}) \qquad (7)$$

(распределение вероятностей $P_\xi$ представляет собой вероятностную меру в фазовом пространстве $(\mathrm{X},\mathrm{B})$) [19, с. 132].

Пусть $\xi_1,\xi_2,...,\xi_n$ - случайные величины на пространстве случайных событий $(\Omega,F,\mathrm{P})$ в соответствующих фазовых пространствах $(\mathrm{X}_k,\mathrm{B}_k)$. Совместным распределением вероятностей этих величин называется функция $P_{\xi_1,\xi_2,...,\xi_n}=P_{\xi_1,\xi_2,...,\xi_n}(B_1,B_2,...,B_n)$, определенная на множествах $B_1\in\mathrm{B}_1,\ B_2\in\mathrm{B}_2,\ \ldots,B_n\in\mathrm{B}_n$ как

$$P_{\xi_1,\xi_2,...,\xi_n}(B_1,B_2,...,B_n)=P_{\xi_1,\xi_2,...,\xi_n}(\xi_1\in B_1,\xi_2\in B_2,...,\xi_n\in B_n). \qquad (8)$$

Распределение вероятностей $P_{\xi_1,\xi_2,...,\xi_n}$ как функция на полукольце множеств вида $B_1\times B_2\times...\times B_n,\ B_1\in\mathrm{B}_1,B_2\in\mathrm{B}_2,...,B_n\in\mathrm{B}_n,$ в произведении пространств $\mathrm{X}_1,\mathrm{X}_2,...,\mathrm{X}_n$ представляет собой функцию распределения. Случайные величины $\xi_1,\xi_2,...,\xi_n$ называются независимыми, если при любых $B_1, B_2,...,B_n$ (см. [19, с.133])

$$P_{\xi_1,\xi_2,...,\xi_n}(B_1,B_2,...,B_n)=P_{\xi_1}(B_1)P_{\xi_2}(B_2)...P_{\xi_n}(B_n). \qquad (9)$$

Предположим, что совместное распределение вероятностей $P_{\xi,\eta}(A,B)$ случайных величин $\xi$ и $\eta$ абсолютно непрерывно относительно некоторой меры $Q$ на произведении пространств $\mathrm{X}\times\mathrm{Y}$, являющейся произведением мер $Q_\mathrm{X}$ и $Q_\mathrm{Y}$, т.е.:

$$P_{\xi,\eta}(A,B)=\int\limits_{A\times B}p(x,y)Q(dx,dy) \qquad (10)$$

для любых $A \in \mathrm{A}$ и $B \in \mathrm{B}$, где *p(x,y)* - соответствующая плотность распределения вероятностей [19, с.145].

В формуле (10) предполагается, что $\xi = \xi(\omega)$ и $\eta = \eta(\omega)$ - случайные величины на одном и том же пространстве элементарных событий $\Omega$ со значениями в фазовых пространствах $(\mathrm{X}, \mathrm{A})$ и $(\mathrm{Y}, \mathrm{B})$. Существование плотности *p(x,y)* вытекает из абсолютной непрерывности $P_{\xi,\eta}(A, B)$ относительно $Q$ в соответствии с теоремой Радона - Никодима.

Условное распределение вероятностей $P_{\xi}(A \mid \eta), A \in \mathrm{A},$ может быть выбрано одинаковым для всех $\omega \in \Omega,$ при которых случайная величина $\eta = \eta(\omega)$ сохраняет одно и то же значение: $\eta(\omega) = y.$ При почти каждом $y \in \mathrm{Y}$ (относительно распределения $P_{\eta}$ в фазовом пространстве $(\mathrm{Y}, \mathrm{B})$) условное распределение вероятностей $P_{\xi}(A \mid y) = P_{\omega,\xi}(A),$ где $\omega \in \{\eta = y\}$ и $A \in \mathrm{A},$ будет абсолютно непрерывно относительно меры $Q_{\mathrm{X}}$:

$$Q_{\mathrm{X}}(A) = \int\limits_{A \times \mathrm{X}} Q(dx, dy).$$

Причем соответствующая плотность условного распределения вероятностей будет иметь вид (см. [19, с.145-146]):

$$p_{\xi}(x \mid y) = \frac{P_{\xi}(dx \mid y)}{Q_{\mathrm{X}}(dx)} = \frac{p(x, y)}{\int\limits_{\mathrm{X}} p(x, y) Q_{\mathrm{X}}(dx)}. \qquad (11)$$

При построении вероятностных моделей реальных явлений важны вероятностные пространства из конечного числа элементарных событий. Для них перечисленные выше общие понятия становятся более прозрачными, в частности, снимаются вопросы измеримости (все подмножества конечного множества обычно считаются измеримыми). Вместо плотностей и условных плотностей рассматриваются вероятности и условные вероятности. Отметим, что вероятности можно рассматривать как плотности относительно меры, приписывающей каждому элементу пространства элементарных событий вес 1, т.е. считающей меры

$$Q(A) = Card(A)$$

(мера каждого множества равна числу его элементов). В целом ясно, что определения основных понятий теории вероятностей в общем ситуации практически не отличаются от таковых в элементарных курсах, во всяком случае с идейной точки зрения.

За последние тридцать лет в прикладной статистике сформировалась новая область - статистика нечисловых данных, она же - статистика объектов нечисловой природы. К настоящему времени она развита не менее, чем ранее выделенные статистика случайных величин, многомерный статистический анализ, статистика временных рядов и случайных процессов. Краткая сводка основных постановок и результатов прикладной статистики в пространствах нечисловой природы даны ниже в настоящей главе и в главе 3.4.

Теория, построенная для результатов наблюдений, лежащих в пространствах общей природы, является центральным стержнем в статистике нечисловой природы. В ее рамках удалось разработать и изучить методы оценивания параметров и характеристик, проверки гипотез (в частности, с помощью статистик интегрального типа), параметрической и непараметрической регрессии (восстановления зависимостей), непараметрического оценивания плотности, дискриминантного и кластерного анализов и т.д.

Вероятностно-статистические методы, развитые для результатов наблюдений из пространств произвольного вида, позволяют единообразно проводить анализ данных из любого конкретного пространства. Так, в монографии [7] они применены к конечным случайным множествам, в работе [18] - к нечетким множествам. С их помощью установлено поведение обобщенного мнения экспертной комиссии (медианы Кемени) при увеличении числа экспертов, когда ответы экспертов лежат в том или ином пространстве бинарных отношений. Методы классификации могут быть основаны на непараметрических оценках плотности распределения

вероятностей в пространстве общей природы. Такие методы были применены для медицинской диагностики в пространстве разнотипных данных, когда часть координат вектора измерена по количественным шкалам, а часть - по качественным, и т.д.

### 2.1.5. Средние и законы больших чисел

Законы больших чисел состоят в том, что эмпирические средние сходятся к теоретическим. В классическом варианте: выборочное среднее арифметическое при определенных условиях сходится по вероятности при росте числа слагаемых к математическому ожиданию. На основе законов больших чисел обычно доказывают состоятельность различных статистических оценок. В целом эта тематика занимает заметное место в теории вероятностей и математической статистике.

Однако математический аппарат при этом основан на свойствах сумм случайных величин (векторов, элементов линейных пространств). Следовательно, он не пригоден для изучения вероятностных и статистических проблем, связанных со случайными объектами нечисловой природы. Это такие объекты, как бинарные отношения, нечеткие множества, вообще элементы пространств без векторной структуры. Объекты нечисловой природы все чаще встречаются в прикладных исследованиях. Много конкретных примеров приведено выше в настоящей главе. Поэтому представляется полезным получение законов больших чисел в пространствах нечисловой природы. Необходимо решить следующие задачи.

А) Определить понятие эмпирического среднего.

Б) Определить понятие теоретического среднего.

В) Ввести понятие сходимости эмпирических средних к теоретическому.

Г) Доказать при тех или иных комплексах условий сходимость эмпирических средних к теоретическому.

Д) Обобщив это доказательство, получить метод обоснования состоятельности различных статистических оценок.

Е) Дать применения полученных результатов при решении конкретных задач.

Ввиду принципиальной важности рассматриваемых результатов приводим доказательство закона больших чисел, а также результаты компьютерного анализа множества эмпирических средних.

**Определения средних величин.** Пусть $X$ - пространство произвольной природы, $x_1, x_2, x_3,\dots,x_n$ - его элементы. Чтобы ввести эмпирическое среднее для $x_1, x_2, x_3,\dots,x_n$ будем использовать действительнозначную (т.е. с числовыми значениями) функцию $f(x,y)$ двух переменных со значениями в $X$. В стандартных математических обозначениях: $f : X^2 \to R^1$. Величина $f(x,y)$ интерпретируется как показатель различия между $x$ и $y$: чем $f(x,y)$ больше, тем $x$ и $y$ сильнее различаются. В качестве $f$ можно использовать расстояние в $X$, квадрат расстояния и т.п.

*Определение 1.* Средней величиной для совокупности $x_1, x_2, x_3,\dots,x_n$ (относительно меры различия $f$), обозначаемой любым из трех способов:

$$x_{cp} = E_n(f) = E_n(x_1, x_2, x_3,\dots,x_n; f),$$

называем решение оптимизационной задачи

$$\sum_{i=1}^{n} f(x_i, y) \to \min, \quad y \in X. \qquad (1)$$

Это определение согласуется с классическим: если $X = R^1$, $f(x,y) = (x - y)^2$, то $x_{cp}$ - выборочное среднее арифметическое. Если же $X = R^1$, $f(x,y) = |x - y|$, то при $n = 2k+1$ имеем $x_{cp} = x(k+1)$, при $n= 2k$ эмпирическое среднее является отрезком $[x(k), x(k+1)]$. Здесь через $x(i)$ обозначен $i$-ый член вариационного ряда, построенного по $x_1, x_2, x_3,\dots,x_n$, т.е. $i$-я порядковая статистика. Таким образом, при $X = R^1$, $f(x,y) = |x - y|$ решение задачи (1) дает естественное определение выборочной медианы. Правда, несколько отличающееся от определения, предлагаемого в курсах "Общей

теории статистики", в котором при $n = 2k$ медианой называют полусумму двух центральных членов вариационного ряда $(x(k) + x(k+1))/2$. Иногда $x(k)$ называют левой медианой, а $x(k+1)$ - правой медианой [7].

Решением задачи (1) является множество $E_n(f)$, которое может быть пустым, состоять из одного или многих элементов. Выше приведен пример, когда решением является отрезок. Если $X = R^1 \setminus \{x_0\}$, $f(x,y) = (x - y)^2$, а среднее арифметическое выборки равно $x_0$, то $E_n(f)$ пусто.

При моделировании реальных ситуаций часто можно принять, что $X$ состоит из конечного числа элементов. Тогда множество $E_n(f)$ непусто - минимум на конечном множестве всегда достигается.

Понятия случайного элемента $x = x(\omega)$ со значениями в $X$, его распределения, независимости случайных элементов используем согласно предыдущему пункту настоящей главы, т.е. каноническому справочнику Ю.В. Прохорова и Ю.А. Розанова [19]. Будем считать, что функция $f$ измерима относительно $\sigma$-алгебры, участвующей в определении случайного элемента $x = x(\omega)$. Тогда $f(x(\omega), y)$ при фиксированном $y$ является действительнозначной случайной величиной. Предположим, что она имеет математическое ожидание.

*Определение 2.* Теоретическим средним $E(x,f)$ (другими словами, математическим ожиданием) случайного элемента $x = x(\omega)$ относительно меры различия $f$ называется решение оптимизационной задачи

$$Mf(x(\omega), y) \to \min, y \in X.$$

Это определение, как и для эмпирических средних, согласуется с классическим. Если $X = R^1$, $f(x,y) = (x - y)^2$, то $E(x,f) = M(x(\text{щ}))$ - обычное математическое ожидание. При этом $Mf(x(\omega), y)$ - дисперсия случайной величины $x = x(\omega)$. Если же $X = R^1$, $f(x,y) = |x - y|$, то $E(x,f) = [a,b]$, где $a = \sup\{t: F(t) \le 0{,}5\}$, $b = \inf\{t: F(t) \ge 0{,}5\}$, где $F(t)$ - функция распределения случайной величины $x = x(\omega)$. Если график $F(t)$ имеет плоский участок на уровне $F(t) = 0{,}5$, то медиана - теоретическое среднее в смысле определения 2 - является отрезком. В классическом случае обычно говорят, что каждый элемент отрезка $[a; b]$ является одним из возможных значений медианы. Поскольку наличие указанного плоского участка - исключительный случай, то обычно решением задачи (2) является множество из одного элемента $a = b$ - классическая медиана распределения случайной величины $x = x(\omega)$.

Теоретическое среднее $E(x, f)$ можно определить лишь тогда, когда $Mf(x(\omega), y)$ существует при всех $y \in X$. Оно может быть пустым множеством, например, если $X = R^1 \setminus \{x_0\}$, $f(x,y) = (x - y)^2$, $x_0 = M(x(\text{щ}))$. И то, и другое исключается, если $X$ конечно. Однако и для конечных $X$ теоретическое среднее может состоять не из одного, а из многих элементов. Отметим, однако, что в множестве всех распределений вероятностей на $X$ подмножество тех распределений, для которых $E(x,f)$ состоит более чем из одного элемента, имеет коразмерность 1, поэтому основной является ситуация, когда множество $E(x,f)$ содержит единственный элемент [7].

**Существование средних величин.** Под существованием средних величин будем понимать непустоту множеств решений соответствующих оптимизационных задач.

Если $X$ состоит из конечного числа элементов, то минимум в задачах (1) и (2) берется по конечному множеству. А потому, как уже отмечалось, эмпирические и теоретические средние существуют.

Ввиду важности обсуждаемой темы приведем доказательства. Для строгого математического изложения нам понадобятся термины из раздела математики под названием "общая топология". Топологические термины и результаты будем использовать в соответствии с классической монографией [20]. Так, топологическое пространство называется бикомпактным в том и только в том случае, когда из каждого его открытого покрытия можно выбрать конечное подпокрытие [20, с.183].

*Теорема 1.* Пусть $X$ - бикомпактное пространство, функция $f$ непрерывна на $X^2$ (в топологии произведения). Тогда эмпирическое и теоретическое средние существуют.

*Доказательство.* Функция $f(x_i, y)$ от $y$ непрерывна, сумма непрерывных функций непрерывна, непрерывная функция на бикомпакте достигает своего минимума, откуда и следует заключение теоремы относительно эмпирического среднего.

Перейдем к теоретическому среднему. По теореме Тихонова [20, с.194] из бикомпактности $X$ вытекает бикомпактность $X^2$. Для каждой точки $(x, y)$ из $X^2$ рассмотрим $\varepsilon / 2$ - окрестность в $X^2$ в смысле показателя различия $f$, т.е. множество

$$U(x, y) = \{(x', y') : | f(x, y) - f(x', y') | < \varepsilon / 2\}.$$

Поскольку $f$ непрерывна, то множества $U(x,y)$ открыты в рассматриваемой топологии в $X^2$. По теореме Уоллеса [20, с.193] существуют открытые (в $X$) множества $V(x)$ и $W(y)$, содержащие $x$ и $y$ соответственно и такие, что их декартово произведение $V(x)$Ч$W(y)$ целиком содержится внутри $U(x, y)$.

Рассмотрим покрытие $X^2$ открытыми множествами $V(x)$Ч$W(y)$. Из бикомпактности $X^2$ вытекает существование конечного подпокрытия $\{V(x_i)$Ч$W(y_i)$, $i = 1, 2, \ldots, m\}$. Для каждого $x$ из $X$ рассмотрим все декартовы произведения $V(x_i)$Ч$W(y_i)$, куда входит точка $(x, y)$ при каком-либо $y$. Таких декартовых произведений и их первых множителей $V(x_i)$ конечное число. Возьмем пересечение таких первых множителей $V(x_i)$ и обозначим его $Z(x)$. Это пересечение открыто, как пересечение конечного числа открытых множеств, и содержит точку $x$. Из покрытия бикомпактного пространства $X$ открытыми множествами $Z(x)$ выберем открытое подпокрытие $Z_1, Z_2, \ldots, Z_k$.

Покажем, что если $x_1'$ и $x_2'$ принадлежат одному и тому же $Z_j$ при некотором $j$, то

$$\sup\{| f(x_1', y) - f(x_2', y) |, y \in X\} < \varepsilon. \qquad (3)$$

Пусть $Z_j = Z(x_0)$ при некотором $x_0$. Пусть $V(x_i)$Ч$W(y_i)$, $i \in I$, - совокупность всех тех исходных декартовых произведений из системы $\{V(x_i)$Ч$W(y_i)$, $i = 1, 2, \ldots, m\}$, куда входят точки $(x_0, y)$ при различных $y$. Покажем, что их объединение содержит также точки $(x_1', y)$ и $(x_2', y)$ при всех $y$. Действительно, если $(x_0, y)$ входит в $V(x_i)$Ч$W(y_i)$, то $y$ входит в $W(y_i)$, а $x_1'$ и $x_2'$ вместе с $x_0$ входят в $V(x_i)$, поскольку $x_1'$, $x_2'$ и $x_0$ входят в $Z(x_0)$. Таким образом, $(x_1', y)$ и $(x_2', y)$ принадлежат $V(x_i)$Ч$W(y_i)$, а потому согласно определению $V(x_i)$Ч$W(y_i)$

$$| f(x_1', y) - f(x_i, y_i) | < \varepsilon / 2, \quad | f(x_2', y) - f(x_i, y_i) | < \varepsilon / 2,$$

откуда и следует неравенство (3).

Поскольку $X^2$ - бикомпактное пространство, то функция $f$ ограничена на $X^2$, а потому существует математическое ожидание $Mf(x(\omega), y)$ для любого случайного элемента $x(\omega)$, удовлетворяющего приведенным выше условиям согласования топологии, связанной с $f$, и измеримости, связанной с $x(\omega)$. Если $x_1$ и $x_2$ принадлежат одному открытому множеству $Zj$, то

$$| Mf(x_1, y) - Mf(x_2, y) | < \varepsilon,$$

а потому функция

$$g(y) = Mf(x(\omega), y) \qquad (4)$$

непрерывна на $X$. Поскольку непрерывная функция на бикомпактном множестве достигает своего минимума, т.е. существуют такие точки $z$, на которых $g(z) = \inf\{g(y), y \in X\}$, то теорема 1 доказана.

В ряде интересных для приложений ситуаций $X$ не является бикомпактным пространством. Например, если $X = R^1$. В этих случаях приходится наложить на показатель различия $f$ некоторые ограничения, например, так, как это сделано в теореме 2.

*Теорема 2.* Пусть $X$ - топологическое пространство, непрерывная (в топологии произведения) функция $f: X^2 \to R^2$ неотрицательна, симметрична (т.е. $f(x,y) = f(y,x)$ для любых $x$ и $y$ из $X$), существует число $D > 0$ такое, что при всех $x, y, z$ из $X$

$$f(x,y) \le D\{f(x,z) + f(z,y)\}. \qquad (5)$$

Пусть в $X$ существует точка $x_0$ такая, что при любом положительном $R$ множество $\{x: f(x, x_0) \le R\}$ является бикомпактным. Пусть для случайного элемента $x(\omega)$, согласованного с топологией в рассмотренном выше смысле, существует $g(x_0) = Mf(x(\omega), x_0)$.

Тогда существуют (т.е. непусты) математическое ожидание $E(x,f)$ и эмпирические средние $E_n(f)$.

*Замечание.* Условие (5) - некоторое обобщение неравенства треугольника. Например, если $g$ - метрика в $X$, а $f = g^p$ при некотором натуральном $p$, то для $f$ выполнено соотношение (5) с $D = 2^p$.

*Доказательство.* Рассмотрим функцию $g(y)$, определенную формулой (4). Имеем

$$f(x(\omega), y) \le D\{f(x(\omega), x_0) + f(x_0, y)\}. \qquad (6)$$

Поскольку по условию теоремы $g(x_0)$ существует, а потому конечно, то из оценки (6) следует существование и конечность $g(y)$ при всех $y$ из $X$. Докажем непрерывность этой функции.

Рассмотрим шар (в смысле меры различия $f$) радиуса $R$ с центром в $x_0$:

$$K(R) = \{x : f(x, x_0) \le R\}, \quad R > 0.$$

В соответствии с условием теоремы $K(R)$ как подпространство топологического пространства $X$ является бикомпактным. Рассмотрим произвольную точку $x$ из $X$. Справедливо разложение

$$f(x(\omega), y) = f(x(\omega), y)\chi(x(\omega) \in K(R)) + f(x(\omega), y)\chi(x(\omega) \notin K(R)),$$

где $\chi(C)$ - индикатор множества $C$. Следовательно,

$$g(y) = Mf(x(\omega), y)\chi(x(\omega) \in K(R)) + Mf(x(\omega), y)\chi(x(\omega) \notin K(R)). \qquad (7)$$

Рассмотрим второе слагаемое в (7). В силу (5)

$$f(x(\omega), y)\chi(x(\omega) \notin K(R)) \le D\{f(x(\omega), x_0)\chi(x(\omega) \notin K(R)) + f(x_0, y)\chi(x(\omega) \notin K(R))\}. \qquad (8)$$

Возьмем математическое ожидание от обеих частей (8):

$$Mf(x(\omega), y)\chi(x(\omega) \notin K(R)) \le D\int_R^{+\infty} tdP\{f(x(\omega), x_0) \le t\} + Df(x_0, y)P(x(\omega) \notin K(R)). \qquad (9)$$

В правой части (9) оба слагаемых стремятся к 0 при безграничном возрастании $R$: первое - в силу того, что

$$g(x_0) = Mf(x(\omega), x_0) = \int_0^{+\infty} tdP(f(x(\omega), x_0) \le t) < \infty,$$

второе - в силу того, что распределение случайного элемента $x(\omega)$ сосредоточено на $X$ и

$$X \setminus \bigcup_{R>0} K(R) = \varnothing.$$

Пусть $U(x)$ - такая окрестность $x$ (т.е. открытое множество, содержащее $x$), для которой

$$\sup\{f(y, x), y \in U(x)\} < +\infty.$$

Имеем

$$f(y, x_0) \le D(f(x_0, x) + f(x, y)). \qquad (10)$$

В силу (9) и (10) при безграничном возрастании $R$

$$Mf(x(\omega), y)\chi(x(\omega) \notin K(R)) \to 0 \qquad (11)$$

равномерно по $y \in U(x)$. Пусть $R(0)$ таково, что левая часть (11) меньше $\varepsilon > 0$ при $R > R(0)$ и, кроме того, $y \in U(x) \subseteq K(R(0))$. Тогда при $R > R(0)$

$$|g(y) - g(x)| \le |Mf(x(\omega), y)\chi(x(\omega) \in K(R)) - Mf(x(\omega), x)\chi(x(\omega) \in K(R))| + 2\varepsilon. \qquad (12)$$

Нас интересует поведение выражения в правой части формулы (12) при $y \in U(x)$. Рассмотрим $f_1$ - сужение функции $f$ на замыкание декартова произведения множеств $U(x) \times K(R)$, и случайный элемент $x_1(\omega) = x(\omega)\chi(x(\omega) \in K(R))$. Тогда

$$Mf(x(\omega), y)\chi(x(\omega) \in K(R)) = Mf_1(x_1(\omega), y)$$

при $y \in U(x)$, а непрерывность функции $g_1(y) = Mf_1(x_1(\omega), y)$ была доказана в теореме 1. Последнее означает, что существует окрестность $U_1(x)$ точки $x$ такая, что

$$| Mf_1(x_1(\omega), y) - Mf_1(x_1(\omega), x) | < \varepsilon \qquad (13)$$

при $y \in U_1(x)$. Из (12) и (13) вытекает, что при $y \in U(x) \bigcap U_1(x)$

$$| g(y) - g(x) | < 3\varepsilon,$$

что и доказывает непрерывность функции $g(x)$.

Докажем существование математического ожидания $E(x,f)$. Пусть $R(0)$ таково, что

$$P(x(\omega) \in K(R(0))) > 1/2. \qquad (14)$$

Пусть $H$ - некоторая константа, значение которой будет выбрано позже. Рассмотрим точку $x$ из множества $K(HR(0))^C$ - дополнения $K(HR(0))$, т.е. из внешности шара радиуса $HR(0)$ с центром в $x_0$. Пусть $x(\omega) \in K(R(0))$. Тогда имеем

$$f(x_0, x) \le D\{f(x_0, x(\omega)) + f(x(\omega), x)\},$$

откуда

$$f(x(\omega), x) \ge \frac{1}{D} f(x_0, x) - f(x_0, x(\omega)) \ge \frac{HR(0)}{D} - R(0). \qquad (15)$$

Выбирая $H$ достаточно большим, получим с учетом условия (14), что при $x \in K(HR(0))^C$ справедливо неравенство

$$Mf(x(\omega), x) \ge \frac{1}{2}\left(\frac{HR(0)}{D} - R(0)\right). \qquad (16)$$

Можно выбрать $H$ так, чтобы правая часть (16) превосходила $g(x_0) = Mf(x(\omega), x_0)$.

Сказанное означает, что Argmin $g(x)$ достаточно искать внутри бикомпактного множества $K(HR(0))$. Из непрерывности функции $g$ вытекает, что ее минимум достигается на указанном бикомпактном множестве, а потому - и на всем $X$. Существование (непустота) теоретического среднего $E(x,f)$ доказана.

Докажем существование эмпирического среднего $E_n(f)$. Есть искушение проводить его дословно так же, как и доказательство существования математического ожидания $E(x,f)$, лишь с заменой 1/2 в формуле (16) на частоту попадания элементов выборки $x_i$ в шар $K(R(0))$. Эта частота, очевидно, стремится к вероятности попадания случайного элемента $x = x(\omega)$ в $K(R(0))$, большей 1/2 в соответствии с (14). Однако это рассуждение показывает лишь, что вероятность непустоты $E_n(f)$ стремится к 1 при безграничном росте объема выборки. Точнее, оно показывает, что

$$\lim_{n \to \infty} P\{E_n(f) \ne \varnothing \wedge E_n(f) \subseteq K(HR(0))\} = 1.$$

Поэтому пойдем другим путем, не опирающимся к тому же на вероятностную модель выборки. Положим

$$R(1) = \max\{f(x_i, x_0), i = 1,2,..., n\}. \qquad (17)$$

Если $x$ входит в дополнение шара $K(HR(1))$, то аналогично (15) имеем

$$f(x_i, x_0) \ge \frac{HR(1)}{D} - R(1). \qquad (18)$$

При достаточно большом $H$ из (17) и (18) следует, что

$$\sum_{i=1}^{n} f(x_i, x_0) \le nR(1) < \sum_{i=1}^{n} f(x_i, x), \quad x \in \{K(HR(1))\}^C.$$

Следовательно, Argmin достаточно искать на $K(HR(1))$. Заключение теоремы 2 следует из того, что на бикомпактном пространстве $K(HR(1))$ минимизируется непрерывная функция.

Теорема 2 полностью доказана.

**О формулировках законов больших чисел.** Пусть $x, x_1, x_2, x_3, ..., x_n$ - независимые одинаково распределенные случайные элементы со значениями в $X$. Закон больших чисел - это

утверждение о сходимости эмпирических средних к теоретическому среднему (математическому ожиданию) при росте объема выборки *n*, т.е. утверждение о том, что

$$E_n(f) = E_n(x_1, x_2, x_3, ..., x_n; f) \to E(x, f) \qquad (19)$$

при $n \to \infty$. Однако и слева, и справа в формуле (19) стоят, вообще говоря, множества. Поэтому понятие сходимости в (19) требует обсуждения и определения.

В силу классического закона больших чисел при $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i, y) \to Mf(x, y) \qquad (20)$$

в смысле сходимости по вероятности, если правая часть существует (теорема А.Я. Хинчина, 1923 г.).

Если пространство *X* состоит из конечного числа элементов, то из соотношения (20) легко вытекает (см., например, [7, с.192-193]), что

$$\lim_{n\to\infty} P\{E_n(f) \subseteq E(x, f)\} = 1. \qquad (21)$$

Другими словами, $E_n(f)$ является состоятельной оценкой $E(x, f)$.

Если $E(x, f)$ состоит из одного элемента, $E(x, f) = \{x_0\}$, то соотношение (21) переходит в следующее:

$$\lim_{n\to\infty} P\{E_n(f) = \{x_0\}\} = 1. \qquad (22)$$

Однако с прикладной точки зрения доказательство соотношений (21) - (22) не дает достаточно уверенности в возможности использования $E_n(f)$ в качестве оценки *E(x,f).* Причина в том, что в процессе доказательства объем выборки предполагается настолько большим, что при всех $y \in X$ одновременно левые части соотношений (20) сосредотачиваются в непересекающихся окрестностях правых частей.

*Замечание.* Если в соотношении (20) рассмотреть сходимость с вероятностью 1, то аналогично (21) получим т.н. усиленный закон больших чисел [7, с.193-194]. Согласно этой теореме с вероятностью 1 эмпирическое среднее $E_n(f)$ входит в теоретическое среднее *E(x,f),* начиная с некоторого объема выборки *n*, вообще говоря, случайного, $n = n(\omega)$. Мы не будем останавливаться на сходимости с вероятностью 1, поскольку в соответствующих постановках, подробно разобранных в монографии [7], нет принципиальных отличий от случая сходимости по вероятности.

Если *X* не является конечным, например, $X = R^1$, то соотношения (21) и (22) неверны. Поэтому необходимо искать иные формулировки закона больших чисел. В классическом случае сходимости выборочного среднего арифметического к математическому ожиданию, т.е. $\bar{x} \to M(x)$, можно записать закон больших чисел так: для любого $\varepsilon > 0$ справедливо предельное соотношение

$$\lim_{n\to\infty} P\{\bar{x} \in (M(x) - \varepsilon;\ M(x) + \varepsilon)\} = 1. \qquad (23)$$

В этом соотношении в отличие от (21) речь идет о попадании эмпирического среднего $E_n(f) = \bar{x}$ не непосредственно внутрь теоретического среднего *E(x,f),* а в некоторую *окрестность* теоретического среднего.

Обобщим эту формулировку. Как задать окрестность теоретического среднего в пространстве произвольной природы? Естественно взять его окрестность, определенную с помощью какой-либо метрики. Однако полезно обеспечить на ее дополнении до *X отделенность* множества значений *Mf*(*x*( $\omega$ ),*y*) как функции *y* от минимума этой функции на всем *X*.

Поэтому мы сочли целесообразным определить такую окрестность с помощью самой функции *Mf*(*x*( $\omega$ ),*y*).

*Определение 3.* Для любого $\varepsilon > 0$ назовем $\varepsilon$-пяткой функции *g(x)* множество

$$K_\varepsilon(g) = \{x : g(x) < \inf\{d(y), y \in X\} + \varepsilon, x \in X\}.$$

Таким образом, в $\varepsilon$-пятку входят все те $x$, для которых значение $g(x)$ либо минимально, либо отличается от минимального (или от инфимума – точной нижней грани) не более чем на $\varepsilon$. Так, для $X = R^1$ и функции $g(x) = x^2$ минимум равен 0, а $\varepsilon$-пятка имеет вид интервала $(-\sqrt{\varepsilon};\sqrt{\varepsilon})$. В формулировке (23) классического закона больших чисел утверждается, что при любом $\varepsilon>0$ вероятность попадания среднего арифметического в $\sqrt{\varepsilon}$-пятку математического ожидания стремится к 1. Поскольку $\varepsilon > 0$ произвольно, то вместо $\sqrt{\varepsilon}$-пятки можно говорить о $\varepsilon$-пятке, т.е. перейти от (23) к эквивалентной записи

$$\lim_{n\to\infty} P\{\bar{x} \in K_\varepsilon(M(x(\omega)-x)^2\} = 1. \qquad (24)$$

Соотношение (24) допускает непосредственное обобщение на общий случай пространств произвольной природы.

СХЕМА ЗАКОНА БОЛЬШИХ ЧИСЕЛ. Пусть $x, x_1, x_2, x_3, ..., x_n$ - независимые одинаково распределенные случайные элементы со значениями в пространстве произвольной природы $X$ с показателем различия $f$: $X^2 \to R^1$. *Пусть выполнены некоторые математические условия регулярности.* Тогда для любого $\varepsilon > 0$ справедливо предельное соотношение

$$\lim_{n\to\infty} P\{E_n(f) \subseteq K_\varepsilon(E(x,f))\} = 1. \qquad (25)$$

Аналогичным образом может быть сформулирована и общая идея усиленного закона больших чисел. Ниже приведены две конкретные формулировки "условий регулярности".

**Законы больших чисел.** Начнем с рассмотрения естественного обобщения конечного множества - бикомпактного пространства $X$.

*Теорема 3*. В условиях теоремы 1 справедливо соотношение (25).

*Доказательство.* Воспользуемся построенным при доказательстве теоремы 1 конечным открытым покрытием $\{Z_1, Z_2, ..., Z_k\}$ пространства $X$ таким, что для него выполнено соотношение (3). Построим на его основе разбиение $X$ на непересекающиеся множества $W_1, W_2, ..., W_m$ (объединение элементов разбиения $W_1, W_2, ..., W_m$ составляет $X$). Это можно сделать итеративно. На первом шаге из $Z_1$ следует вычесть $Z_2, ..., Z_k$ - это и будет $W_1$. Затем в качестве нового пространства надо рассмотреть разность $X$ и $W_1$, а покрытием его будет $\{Z_2, ..., Z_k\}$. И так до $k$-го шага, когда последнее из рассмотренных покрытий будет состоять из единственного открытого множества $Z_k$. Остается из построенной последовательности $W_1, W_2, ..., W_k$ вычеркнуть пустые множества, которые могли быть получены при осуществлении описанной процедуры (поэтому, вообще говоря, $m$ может быть меньше $k$).

В каждом из элементов разбиения $W_1, W_2, ..., W_m$ выберем по одной точке, которые назовем центрами разбиения и соответственно обозначим $w_1, w_2, ..., w_m$. Это и есть то конечное множество, которым можно аппроксимировать бикомпактное пространство $X$. Пусть $y$ входит в $W_j$. Тогда из соотношения (3) вытекает, что

$$\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i, y) - \frac{1}{n}\sum_{i=1}^{n} f(x_i, w_j)\right| < \varepsilon. \qquad (26)$$

Перейдем к доказательству соотношения (25). Возьмем произвольное $\delta>0$. Рассмотрим некоторую точку $b$ из $E(x,f)$. Доказательство будет основано на том, что с вероятностью, стремящейся к 1, для любого $y$ вне $K_\delta(E(x,f))$ выполнено неравенство

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i, y) > \frac{1}{n}\sum_{i=1}^{n} f(x_i, b). \qquad (27)$$

Для обоснования этого неравенства рассмотрим все элементы разбиения $W_1, W_2, ..., W_m$, имеющие непустое пересечение с внешностью $\delta$-пятки $K_\delta(E(x,f))$. Из неравенства (26) следует, что для любого $y$ вне $K_\delta(E(x,f))$ левая часть неравенства (27) не меньше

$$\min_j \left( \frac{1}{n} \sum_{i=1}^{n} f(x_i, w_j) \right) - \varepsilon, \qquad (28)$$

где минимум берется по центрам всех элементов разбиения, имеющим непустое пересечение с внешностью $\delta$-пятки. Возьмем теперь в каждом таком разбиении точку $v_i$, лежащую вне д-пятки $K_\delta(E(x,f))$. Тогда из неравенств (3) и (28) следует, что левая часть неравенства (27) не меньше

$$\min_j \left( \frac{1}{n} \sum_{i=1}^{n} f(x_i, v_j) \right) - 2\varepsilon. \qquad (29)$$

В силу закона больших чисел для действительнозначных случайных величин каждая из участвующих в соотношениях (27) и (29) средних арифметических имеет своими пределами соответствующие математические ожидания, причем в соотношении (29) эти пределы не менее

$$Mf(x(\omega), b) + \delta - 2\varepsilon,$$

поскольку точки $v_i$ лежат вне $\delta$-пятки $K_\delta(E(x,f))$. Следовательно, при

$$\delta - 2\varepsilon > 0$$

и достаточно большом $n$, обеспечивающем необходимую близость рассматриваемого конечного числа средних арифметических к их математическим ожиданиям, справедливо неравенство (27).

Из неравенства (27) следует, что пересечение $E_n(f)$ с внешностью $K_\delta(E(x,f))$ пусто. При этом точка $b$ может входить в $E_n(f)$, а может и не входить. Во втором случае $E_n(f)$ состоит из иных точек, входящих в $K_\delta(E(x,f))$. Теорема 3 доказана.

Если $X$ не является бикомпактным пространством, то необходимо суметь оценить рассматриваемые суммы "на периферии", вне бикомпактного ядра, которое обычно выделяется естественным путем. Один из возможных комплексов условий сформулирован выше в теореме 2.

*Теорема 4.* В условиях теоремы 2 справедлив закон больших чисел, т.е. соотношение (25).

*Доказательство.* Будем использовать обозначения, введенные в теореме 2 и при ее доказательстве. Пусть $r$ и $R$, $r < R$, - положительные числа. Рассмотрим точку $x$ в шаре $K(r)$ и точку $y$ вне шара $K(R)$. Поскольку

$$f(x_0, y) \le D\{f(x_0, x) + f(x, y)\},$$

то

$$f(x,y) \ge \frac{1}{D} f(x_0, y) - f(x_0, x) \ge \frac{R}{D} - r. \qquad (30)$$

Положим

$$g_n(x) = g_n(x, \omega) = \frac{1}{n} \sum_{i=1}^{n} f(x_i, x).$$

Сравним $g_n(x_0)$ и $g_n(y)$. Выборку $x_1, x_2, x_3, ..., x_n$ разобьем на две части. В первую часть включим те элементы выборки, которые входят в $K(r)$, во вторую - все остальные (т.е. лежащие вне $K(r)$). Множество индексов элементов первой части обозначим $I = I(n,r)$. Тогда в силу неотрицательности $f$ имеем

$$g_n(y) \ge \frac{1}{n} \sum_{i \in I} f(x_i, y),$$

а в силу неравенства (30)

$$\sum_{i \in I} f(x_i, y) \ge \left( \frac{R}{D} - r \right) Card I(n,r),$$

где $Card\ I(n,r)$ - число элементов в множестве индексов $I(n,r)$. Следовательно,

$$g_n(y) \ge \frac{1}{n} \left( \frac{R}{D} - r \right) J, \qquad (31)$$

где $J = Card\ I(n,r)$ - биномиальная случайная величина $B(n,p)$ с вероятностью успеха $p = P\{x_i(\omega) \in K(r)\}$. По теореме Хинчина для $g_n(x_0)$ справедлив (классический) закон больших чисел. Пусть $\varepsilon > 0$. Выберем $n_1 = n_1(\varepsilon)$ так, чтобы при $n > n_1$ было выполнено соотношение

$$P\{g_n(x_0) - g(x_0) > \varepsilon\} < \varepsilon, \qquad (32)$$

где $g(x_0) = Mf(x_1, x_0)$. Выберем $r$ так, чтобы вероятность успеха $p > 0{,}6$. По теореме Бернулли можно выбрать $n_2 = n_2(\varepsilon)$ так, чтобы при $n > n_2$

$$P\{J > 0{,}5n\} > 1 - \varepsilon. \qquad (33)$$

Выберем $R$ так, чтобы

$$\frac{1}{2}\left(\frac{R}{D} - r\right) > g(x_0) + \varepsilon.$$

Тогда

$$K_\varepsilon(g) \subseteq K(R) \qquad (34)$$

и согласно (31), (32) и (33) при $n > n_3 = \max(n_1, n_2)$ с вероятностью не менее $1 - \varepsilon$ имеем

$$g_n(y) > g_n(x_0) \qquad (35)$$

для любого $y$ вне $K(R)$. Из (34) следует, что минимизировать $g_n$ достаточно внутри бикомпактного шара $K(R)$, при этом $E_n(f)$ не пусто и

$$E_n(f) \subseteq K(R) \qquad (36)$$

с вероятностью не менее $1\text{-}2\varepsilon$.

Пусть $g_n^{'}$ и $g^{'}$ - сужения $g_n$ и $g(x) = Mf(x(\omega), x)$ соответственно на $K(R)$ как функций от $x$. В силу (34) справедливо равенство $K_\varepsilon(g^{'}) = K_\varepsilon(f)$. Согласно доказанной выше теореме 3 найдется $n_4 = n_4(\omega)$ такое, что

$$P(K_0(g_n^{'}) \subseteq K_\varepsilon(g)) > 1 - \varepsilon.$$

Согласно (36) с вероятностью не менее $1 - 2\varepsilon$

$$K_0(g_n^{'}) = E_n(f)$$

при $n > n_3$. Следовательно, при $n > n_5(\varepsilon) = \max(n_3, n_4)$ имеем

$$P(E_n(f) \subseteq K_\varepsilon(g)) > 1 - 3\varepsilon,$$

что и завершает доказательство теоремы 4.

Справедливы и иные варианты законов больших чисел, полученные, в частности, в статье [21].

**Медиана Кемени и экспертные оценки.** Рассмотрим на основе развитой выше теории частный случай пространств нечисловой природы - пространство бинарных отношений на конечном множестве $Q = \{q_1, q_2, ..., q_k\}$ и его подпространства. Как известно, каждое бинарное отношение $A$ можно описать матрицей $\|a(i,j)\|$ из 0 и 1, причем $a(i,j) = 1$ тогда и только тогда $q_i$ и $q_j$ находятся в отношении $A$, и $a(i,j) = 0$ в противном случае.

*Определение 4.* Расстоянием Кемени между бинарными отношениями $A$ и $B$, описываемыми матрицами $\|a(i,j)\|$ и $\|b(i,j)\|$ соответственно, называется

$$d(A, B) = \sum_{i,j=1}^{k} |\, a(i, j) - b(i, j)\,|.$$

*Замечание.* Иногда в определение расстояния Кемени вводят множитель, зависящий от $k$.

*Определение 5.* Медианой Кемени для выборки, состоящей из бинарных отношений, называется эмпирическое среднее, построенное с помощью расстояния Кемени.

Поскольку число бинарных отношений на конечном множестве конечно, то эмпирические и теоретические средние для произвольных показателей различия существуют и справедливы законы больших чисел, описанные формулами (21) и (22) выше.

Бинарные отношения, в частности, упорядочения, часто используются для описания мнений экспертов. Тогда расстояние Кемени измеряет близость мнений экспертов, а медиана Кемени позволяет находить итоговое усредненное мнение комиссии экспертов. Расчет медианы Кемени обычно включают в информационное обеспечение систем принятия решений с использованием оценок экспертов. Речь идет, например, о математическом обеспечении автоматизированного рабочего места "Математика в экспертизе" (АРМ "МАТЭК"), предназначенного, в частности, для использования при проведении экспертиз в задачах экологического страхования. Поэтому представляет большой практический интерес численное изучение свойств медианы Кемени при конечном объеме выборки. Такое изучение дополняет описанную выше асимптотическую теорию, в которой объем выборки предполагается безгранично возрастающим ($n \to \infty$).

**Компьютерное изучение свойств медианы Кемени при конечных объемах выборок.** С помощью специально разработанной программной системы В.Н. Жихаревым был проведен ряд серий численных экспериментов по изучению свойств выборочных медиан Кемени. Представление о полученных результатах дается приводимой ниже табл.5, взятой из статьи [22]. В каждой серии методом статистических испытаний определенное число раз моделировался случайный и независимый выбор экспертных ранжировок, а затем находились все медианы Кемени для смоделированного набора мнений экспертов. При этом в сериях 1-5 распределение ответа эксперта предполагалось равномерным на множестве всех ранжировок. В серии 6 это распределение являлось монотонным относительно расстояния Кемени с некоторым центром (о понятии монотонности см. выше), т.е. вероятность выбора определенной ранжировки убывала с увеличением расстояния Кемени этой ранжировки от центра. Таким образом, серии 1-5 соответствуют ситуации, когда у экспертов нет почвы для согласия, нет группировки их мнений относительно некоторого единого среднего группового мнения, в то время как в серии 6 есть единое мнение - описанный выше центр, к которому тяготеют ответы экспертов.

Результаты, приведенные в табл.5, можно комментировать разными способами. Неожиданным явилось большое число элементов в выборочной медиане Кемени - как среднее, так и особенно максимальное. Одновременно обращает на себя внимание убывание этих чисел при росте числа экспертов и особенно при переходе к ситуации реального существования группового мнения (серия 6). Достаточно часто один из ответов экспертов входит в медиану Кемени (т.е. пересечение множества ответов экспертов и медианы Кемени непусто), а диаметр медианы как множества в пространстве ранжировок заметно меньше диаметра множества ответов экспертов. По этим показателям - наилучшее положение в серии 6. Грубо говоря, всяческие "патологии" в поведении медианы Кемени наиболее резко проявляются в ситуации, когда ее применение не имеет содержательного обоснования, т.е. когда у экспертов нет основы для согласия, их ответы равномерно распределены на множестве ранжировок.

Таблица 5.

Вычислительный эксперимент по изучению свойств медианы Кемени

| Номер серии | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Число испытаний | 100 | 1000 | 50 | 50 | 1000 | 1000 |
| Количество объектов | 5 | 5 | 7 | 7 | 5 | 5 |
| Количество экспертов | 10 | 30 | 10 | 30 | 10 | 10 |
| Частота непустого пересечения | 0,85 | 0,58 | 0,52 | 0,2 | 0,786 | 0,911 |
| Среднее отношение диаметров | 0.283 | 0,124 | 0,191 | 0,0892 | 0,202 | 0.0437 |
| Средняя мощность медианы | 5,04 | 2,41 | 6,4 | 2,88 | 3,51 | 1,35 |
| Максимальная. мощность медианы | 30 | 14 | 19 | 11 | 40 | 12 |

Увеличение числа испытаний в 10 раз при переходе от серии 1 к серии 5 не очень сильно повлияло на приведенные в таблице характеристики, поэтому представляется, что суть дела выявляется при числе испытаний (в методе Монте-Карло), равном 100 или даже 50. Увеличение числа объектов или экспертов увеличивает число элементов в рассматриваемом пространстве ранжировок, а потому уменьшается частота попадания какого-либо из мнений экспертов внутрь медианы Кемени. А также отношение диаметра медианы к диаметру множества экспертов и число элементов медианы Кемени (среднее и максимальное). Можно сказать, что увеличение числа объектов или экспертов уменьшает степень дискретности задачи, приближает ее к непрерывному случаю, а потому уменьшает выраженность различных "патологий".

Есть много интересных результатов, которые здесь не рассматриваем. Они связаны, в частности, со сравнением медианы Кемени с другими методами усреднения мнений экспертов, например, с нахождением итогового упорядочения по методу средних рангов [10]. А также с использованием малых окрестностей ответов экспертов для поиска входящих в медиану ранжировок, с теоретической и численной оценкой скорости сходимости в законах больших чисел.

### 2.1.6. Непараметрические оценки плотности

Эмпирическая функция распределения – это состоятельная непараметрическая оценка функции распределения числовой случайной величины. А как оценить плотность? Если продифференцировать эмпирическую функцию распределения, то получим бесконечности в точках, соответствующих элементам выборки, и 0 во всех остальных. Ясно, что это не оценка плотности.

Как же действовать? Каждому элементу выборки соответствует в эмпирическом распределении вероятность $1/n$, где $n$ – объем выборки. Целесообразно эту вероятность не помещать в одну точку, а «размазать» вокруг нее, построив «холмик». Если «холмики» налегают друг на друга, то получаем положительную плотность на всей прямой. Чтобы получить состоятельную оценку плотности, необходимо выбирать ширину «холмика» в зависимости от объема выборки. При этом число «холмиков», покрывающих фиксированную точку, должно безгранично расти. Но одновременно доле таких «холмиков» следует убывать, поскольку покрывающие «холмики» должны быть порождены лишь ближайшими членами вариационного ряда.

Реализация описанной идеи привела к различным вариантам непараметрических оценок плотности. Основополагающей является работа Н.В.Смирнова 1951 г. [23]. Вначале рассматривались непараметрические оценки плотности распределения числовых случайных величин и конечномерных случайных векторов. В 1980-х годах удалось сконструировать такие оценки в пространствах произвольной природы [24], а затем и для конкретных видов нечисловых данных [25].

Сначала рассмотрим непараметрические оценки плотности в наиболее общей ситуации. В статистике нечисловых данных выделяют общую теорию и статистику в конкретных пространствах нечисловой природы (например, статистику ранжировок). В общей теории есть два основных сюжета. Один связан со средними величинами и асимптотическим поведением решений экстремальных статистических задач, второй - с непараметрическими оценками плотности. Первый сюжет только что рассмотрен, второму посвящена заключительная часть настоящей главы.

Понятие плотности в пространстве произвольной природы $X$ требует специального обсуждения. В пространстве $X$ должна быть выделена некоторая специальная мера $\mu$, относительно которой будут рассматриваться плотности, соответствующие другим мерам, например, мере $\nu$, задающей распределение вероятностей некоторого случайного элемента $\xi$. В таком случае $\nu(A) = \mathrm{P}(\xi \in A)$ для любого случайного события $A$. Плотность $f(x)$, соответствующая мере $\nu$ - это такая функция, что

$$\nu(A) = \int_A f(x) d\mu$$

для любого случайного события $A$. Для случайных величин и векторов мера $\mu$ - это объем множества $A$, в математических терминах - мера Лебега. Для дискретных случайных величин и элементов со значениями в конечном множестве $X$ в качестве меры $\mu$ естественно использовать считающую меру, которая событию $A$ ставит в соответствие число его элементов. Используют также нормированную случайную меру, когда число точек в множестве $A$ делят на число точек во всем пространстве $X$. В случае считающей меры значение плотности в точке $x$ совпадает с вероятностью попасть в точку $x$, т.е. $f(x) = P(o = x)$. Таким образом, с рассматриваемой точки зрения стирается грань между понятиями «плотность вероятности» и «вероятность (попасть в точку)».

Как могут быть использованы непараметрические оценки плотности распределения вероятностей в пространствах нечисловой природы? Например, для решения задач классификации (диагностики, распознавания образов - см. главу 3.2). Зная плотности распределения классов, можно решать основные задачи диагностики - как задачи выделения кластеров, так и задачи отнесения вновь поступающего объекта к одному из диагностических классов. В задачах кластер-анализа можно находить моды плотности и принимать их за центры кластеров или за начальные точки итерационных методов типа $k$-средних или динамических сгущений. В задачах собственно диагностики (дискриминации, распознавания образов с учителем) можно принимать решения о диагностике объектов на основе отношения плотностей, соответствующих классам. При неизвестных плотностях представляется естественным использовать их состоятельные оценки.

Методы оценивания плотности вероятности в пространствах общего вида предложены и первоначально изучены в работе [24]. В частности, в задачах диагностики объектов нечисловой природы предлагаем использовать непараметрические ядерные оценки плотности типа Парзена - Розенблатта (этот вид оценок и его название впервые были введены в статье [24] ). Они имеют вид:

$$f_n(x) = \frac{1}{\eta_n(h_n, x)} \sum_{1 \le i \le n} K\left(\frac{d(x_i, x)}{h_n}\right),$$

где $K$: $R^1_+ \to R^1$ - так называемая ядерная функция, $x_1, x_2, \ldots, x_n \in X$ - выборка, по которой оценивается плотность, $d(x_i, x)$ - показатель различия (метрика, расстояние, мера близости) между элементом выборки $x_i$ и точкой $x$, в которой оценивается плотность, последовательность $h_n$ показателей размытости такова, что $h_n \to 0$ и $nh_n \to \infty$ при $n \to \infty$, а $\eta_n(h_n, x)$ - нормирующий множитель, обеспечивающий выполнение условия нормировки (интеграл по всему пространству от непараметрической оценки плотности $f_n(x)$ по мере $\mu$ должен равняться 1). Ранее американские исследователи Парзен и Розенблатт использовали подобные статистики в случае $X = R^1$ с $d(x_i, x) = |x_i - x|$.

Введенные описанным образом ядерные оценки плотности - частный случай так называемых линейных оценок, также впервые предложенных в работе [24]. В теоретическом плане они выделяются тем, что удается получать результаты такого же типа, что в классическом одномерном случае, но, разумеется, с помощью совсем иного математического аппарата.

**Свойства непараметрических ядерных оценок плотности.** Рассмотрим выборку со значениями в некотором пространстве произвольного вида. В этом пространстве предполагаются заданными показатель различия $d$ и мера $\mu$. Одна из основных идей рассматриваемого подхода состоит в том, чтобы согласовать их между собой. А именно, на их основе построим новый показатель различия $d_1$ , так называемый "естественный", в терминах которого проще формулируются свойства непараметрической оценки плотности. Для этого рассмотрим шары $L_t(x) = \{y \in X : d(y,x) \le t\}$ радиуса $t \ge 0$ и их меры $F_x(t) = \mu(L_t(x))$. Предположим, что $F_x(t)$ как функция $t$ при фиксированном $x$ непрерывна и строго возрастает. Введем функцию $d_1(x,y)=$

$F_x(d(x,y))$. Это - монотонное преобразование показателя различия или расстояния, а потому $d_1(x,y)$ - также показатель различия (даже если $d$ - *метрика,* для $d_1$ неравенство треугольника может быть не выполнено). Другими словами, $d_1(x,y)$, как и $d(x,y)$, можно рассматривать как показатель различия (меру близости) между $x$ и $y$.

Для вновь введенного показателя различия $d_1(x,y)$ введем соответствующие шары $L_{1t}(x) = \{y \in X : d_1(y,x) \le t\}$. Поскольку обратная функция $F^{-1}{}_x(t)$ определена однозначно, то

$$L_{1t}(x) = \{y \in X : d_1(y,x) \le F_x^{-1}(t)\} = L_T(x),$$

где $T = F^{-1}{}_x(t)$. Следовательно, справедлива цепочка равенств $F^1{}_x(t) = \mu(L_{1t}(x)) = \mu(L_T(x)) = F_x(F^{-1}{}_x(t)) = t$ (для всех тех значений параметра $t$, для которых определены все участвующие в записи математические объекты).

Переход от $d$ к $d_1$ напоминает классическое преобразование, использованное Н.В. Смирновым при изучении непараметрических критериев согласия и однородности, а именно, преобразование $\eta = F(\xi)$, переводящее случайную величину $\xi$ с непрерывной функцией распределения $F(x)$ в случайную величину $\eta$, равномерно распределенную на отрезке [0,1]. Оба рассматриваемых преобразования существенно упрощают дальнейшие рассмотрения. Преобразование $d_1 = F_x(d)$ зависит от точки $x$, что не влияет на дальнейшие рассуждения, поскольку ограничиваемся изучением сходимости в отдельно взятой точке.

Функцию $d_1(x,y)$, для которой мера шара радиуса $t$ равна $t$, называем в соответствии с работой [24] «естественным показателем различия» или «естественной метрикой». В случае конечномерного пространства $R^k$ и евклидовой метрики $d$ имеем $d_1(x,y) = c_k d^k(x,y)$, где $c_k$ - объем шара единичного радиуса в $R^k$.

Поскольку можно записать, что

$$K\left(\frac{d(x_i,x)}{h_n}\right) = K_1\left(\frac{d_1(x_i,x)}{h_n}\right),$$

где

$$K_1(u) = K\left(\frac{F_x^{-1}(uh_n)}{h_n}\right),$$

то переход от одного показателя различия к другому, т.е. от $d$ к $d_1$, соответствует переходу от одной ядерной функции к другой, т.е. от $K$ к $K_1$. Выгода от такого перехода заключается в том, что утверждения о поведении непараметрических оценок плотности приобретают более простую формулировку.

*Теорема 5.* Пусть $d$ - естественная метрика, плотность $f$ непрерывна в точке $x$ и ограничена на всем пространстве $X$, причем $f(x)>0$, ядерная функция $K(u)$ удовлетворяет простым условиям регулярности

$$\int_0^1 K(u)du = 1, \int_0^\infty (|K(u)| + K^2(u))du < \infty.$$

Тогда $\eta_n(h_n, x) = nh_n$, оценка $f_n(x)$ является состоятельной, т.е. $f_n(x) \to f(x)$ по вероятности при $n \to \infty$ и, кроме того,

$$\lim_{n\to\infty}(nh_n Df_n(x)) = f(x)\int_0^{+\infty} K^2(u)du.$$

Теорема 5 доказывается методами, развитыми в работе [24]. Однако остается открытым вопрос о скорости сходимости ядерных оценок, в частности, о поведении величины $\alpha_n = M(f_n(x) - f(x))^2$ - среднего квадрата ошибки, и об оптимальном выборе показателей размытости $h_n$. Для того, чтобы продвинуться в решении этого вопроса, введем новые понятия. Для случайного элемента $X(\omega)$ со значениями в $X$ рассмотрим т.н. круговое распределение $G(x,t) = P\{d(X(\omega), x) \le t\}$ и круговую плотность $g(x,t) = G'_t(x,t)$.

*Теорема 6.* Пусть ядерная функция $K(u)$ непрерывна и финитна, т.е. существует число $E$ такое, что $K(u)=0$ *при* $u>E$. Пусть круговая плотность является достаточно гладкой, т.е. допускает разложение

$$g(x,t) = f(x) + tg_t'(x,0) + \frac{t^2}{2} g_{tt}''(x,0) + \frac{t^3}{3!} g_{ttt}'''(x,0) + ... + \frac{t^k}{k!} g_{t^{(k)}}^{(k)}(x,0) + o(h_n^k)$$

при некотором $k$, причем остаточный член равномерно ограничен на $[0,hE]$. Пусть

$$\int_0^E u^i K(u)du = 0, i = 1,2,...,k-1.$$

Тогда

$$\alpha_n = [Mf_n(x) - f(x)]^2 + Df_n(x) =$$

$$= h_n^{2k}\left(\int_0^E u^k K(u)du\right)^2 (g_{t^{(k)}}^k(x,0))^2 + \frac{f(x)}{nh_n}\int_0^E K^2(u)du + o\left(h_n^{2k} + \frac{1}{nh_n}\right).$$

Доказательство теоремы 6 проводится с помощью разработанной в статистике объектов нечисловой природы математической техники, образцы которой представлены, в частности, в работе [24]. Если коэффициенты при основных членах в правой части последней формулы не равны 0, то величина $\alpha_n$ достигает минимума, равного $\alpha_n = O\left(n^{-1+\frac{1}{2k+1}}\right)$, при $h_n = n^{-\frac{1}{2k+1}}$. Эти выводы совпадают с классическими результатами, полученными ранее рядом авторов для весьма частного случая прямой $X = R^1$ (см., например, монографию [26, с.316]). Заметим, что для уменьшения смещения оценки приходится применять знакопеременные ядра $K(u)$.

**Непараметрические оценки плотности в конечных пространствах** [25]. В случае пространств из конечного числа элементов естественных метрик не существует. Однако можно получить аналоги теорем 5 и 6, переходя к пределу не только по объему выборки $n$, но и по новому параметру дискретности $m$.

Рассмотрим некоторую последовательность $X_m$, $m = 1, 2, \ldots$, конечных пространств. Пусть в $X_m$ заданы показатели различия $d_m$. Будем использовать нормированные считающие меры $\mu_m$, ставящие в соответствие каждому подмножеству $A$ долю элементов всего пространства $X_m$, входящих в $A$. Как и ранее, рассмотрим как функцию $t$ объем шара радиуса $t$, т.е.

$$F_{mx}(t) = \mu_m(\{y \in X_m : d_m(x,y) \le t\}).$$

Введем аналог естественного показателя различия $d_{1m}(x,y) = F_{mx}(d_m(x,y))$. Наконец, рассмотрим аналоги преобразования Смирнова $F_{mx}^1(t) = \mu_m(\{y \in X_m : d_{1m}(x,y) \le t\})$. Функции $F_{mx}^1(t)$, в отличие от ситуации предыдущего раздела, уже не совпадают тождественно с $t$, они кусочно-постоянны и имеют скачки в некоторых точках $t_i$, $i = 1, 2, \ldots$, причем в этих точках $F_{mx}^1(t_i) = t_i$.

*Теорема 7.* Пусть точки скачков равномерно сближаются, т.е. $\max(t_i - t_{i-1}) \to 0$ при $m \to \infty$ (другими словами, $\sup |F_{mx}^1(t) - t| \to 0$ при $m \to \infty$). Тогда существует последовательность параметров дискретности $m_n$ такая, что при предельном переходе $n \to \infty, m \to \infty, m \ge m_n$ справедливы заключения теорем 5 и 6.

*Пример 1.* Пространство $X_m = 2^{\sigma(m)}$ всех подмножеств конечного множества $\sigma(m)$ из $m$ элементов допускает (см. главу 1.1 и монографию [7]) аксиоматическое введение метрики $d(A,B) = card(A \Delta B)/2^m$, где $\Delta$ - символ симметрической разности множеств. Рассмотрим непараметрическую ядерную оценку плотности типа Парзена - Розенблатта

$$f_{nm}(A) = \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{1}{h_n}\Phi\left(\frac{2card(A\Delta X_i) - m}{\sqrt{m}}\right)\right),$$

где Ф(·) - функция нормального стандартного распределения. Можно показать, что эта оценка удовлетворяет условиям теоремы 7 с $m_n = (\ln n)^6$.

*Пример 2.* Рассмотрим пространство функций $f : Y_r \to Z_q$, определенных на конечном множестве $Y_r = \{1/r,\ 2/r,...,(r-1)/r,\ 1\}$, со значениями в конечном множестве $Z_q = \{0,\ 1/q, 2/q,...,\ (q-1)/q,\ 1\}$. Это пространство можно интерпретировать как пространство нечетких множеств (см. главу 1.1), а именно, $Y_r$ - носитель нечеткого множества, а $Z_q$ - множество значений функции принадлежности. Очевидно, число элементов пространства $X_m$ равно $(q+1)^r$. Будем использовать расстояние $d(f,g) = \sup |f(y) - g(y)|$. Непараметрическая оценка плотности имеет вид:

$$f_{nm}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{[2\sup_y |x(y) - x_i(y)| + 1/q]^r}{h_n(1+1/q)^r}\right).$$

Если $r = n^\alpha,\ q = n^\beta$, то при $\beta > \alpha$ выполнены условия теоремы 7, а потому справедливы теоремы 5 и 6.

*Пример 3.* Рассматривая пространства ранжировок $m$ объектов, в качестве расстояния $d(A,B)$ между ранжировками $A$ и $B$ примем минимальное число инверсий, необходимых для перехода от А к В. Тогда $\max(t_i - t_{i-1})$ не стремится к 0 при $m \to \infty$, условия теоремы 7 не выполнены.

*Пример 4.* В прикладных работах наиболее распространенный пример объектов нечисловой природы – вектор разнотипных данных: реальный объект описывается вектором, часть координат которого - значения количественных признаков, а часть - качественных (номинальных и порядковых). Для пространств разнотипных признаков, т.е. декартовых произведений непрерывных и дискретных пространств, возможны различные постановки. Пусть, например, число градаций качественных признаков остается постоянным. Тогда непараметрическая оценка плотности сводится к произведению двух величин - частоты попадания в точку в пространстве качественных признаков и классической оценки типа Парзена-Розенблатта в пространстве количественных переменных. В общем случае расстояние $d(x,y)$ можно, например, рассматривать как сумму трех расстояний. А именно, евклидова расстояния $d_1$ между количественными факторами, расстояния $d_2$ между номинальными признаками ($d_2(x,y) = 0$, если $x = y$, и $d_2(x,y) = 1$, если $x \neq y$) и расстояния $d_3$ между порядковыми переменными (если $x$ и $y$ - номера градаций, то $d_3(x,y) = |x - y|$). Наличие количественных факторов приводит к непрерывности и строгому возрастанию функции $F_{mx}(t)$, а потому для непараметрических оценок плотности в пространствах разнотипных признаков верны теоремы 5 - 6.

Программная реализация описания числовых данных с помощью непараметрических оценок плотности включена в ряд программных продуктов по прикладной статистике, в частности, в пакет программ анализа данных ППАНД [27].

## Литература

1. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. - Л.: Энергоатомиздат, 1985. - 248 с.
2. Новицкий П.В. Основы информационной теории измерительных устройств. -Л.: Энергия, 1968. - 248 с.
3. Боровков А.А. Теория вероятностей. - М.: Наука, 1976. - 352 с.
4. Петров В.В. Суммы независимых случайных величин. - М.: Наука, 1972. - 416 с.

5. Золотарев В.М. Современная теория суммирования независимых случайных величин. - М.: Наука, 1986. - 416 с.
6. Егорова Л.А., Харитонов Ю.С., Соколовская Л.В.//Заводская лаборатория. - 1976. Т.42, №10. С. 1237.
7. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука,1979. – 296 с.
8. Колмогоров А.Н. Избранные труды: Математика и механика. - М.: Наука, 1985. С. 136-138.
9. Пфанцагль И. Теория измерений. - М.: Мир, 1976. - 165 с.
10. Орлов А.И. Эконометрика. Учебник для вузов. Изд. 2-е, исправленное и дополненное. - М.: Изд-во "Экзамен", 2003. – 576 с.
11. Орлов А.И. Вероятностные модели конкретных видов объектов нечисловой природы. – Журнал «Заводская лаборатория». 1995. Т.61. No.5. С.43-51.
12. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В. Прохоров. - М.: Большая Российская энциклопедия, 1999. - 910 с.
13. Дэвид Г. Метод парных сравнений. - М.: Статистика, 1978.- 144 с.
14. Орлов А.И. Логистическое распределение. – В сб.: Математическая энциклопедия. Т.3. - М.: Советская энциклопедия, 1982. - С.414.
15. Тюрин Ю.Н., Василевич А.П., Андрукович П.Ф. Статистические модели ранжирования. - В сб.: Статистические методы анализа экспертных оценок. - М.: Наука, 1977. - С.30-58.
16. Орлов А.И. Случайные множества с независимыми элементами (люсианы) и их применения. – В сб.: Алгоритмическое и программное обеспечение прикладного статистического анализа. Ученые записки по статистике, т.36. - М.: Наука, 1980. - С. 287-308.
17. Орлов А.И. Парные сравнения в асимптотике Колмогорова. – В сб.: Экспертные оценки в задачах управления. - М.: Изд-во Института проблем управления АН СССР, 1982. - С. 58-66.
18. Орлов А.И. Задачи оптимизации и нечеткие переменные. - М.: Знание, 1980. – 64 с.
19. Прохоров Ю.В., Розанов Ю.А. Теория вероятностей. (Основные понятия. Предельные теоремы. Случайные процессы.) - М.: Наука, 1973.- 496 с.
20. Келли Дж. Общая топология. - М.: Наука, 1968. - 384 с.
21. Орлов А.И. Асимптотика решений экстремальных статистических задач. – В сб.: Анализ нечисловых данных в системных исследованиях. Сборник трудов. Вып.10. - М.: Всесоюзный научно-исследовательский институт системных исследований, 1982. - С. 4-12.
22. Жихарев В.Н., Орлов А.И. Законы больших чисел и состоятельность статистических оценок в пространствах произвольной природы. – В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. – Пермь: Изд-во Пермского государственного университета, 1998. С.65-84.
23. Смирнов Н.В. О приближении плотностей распределения случайных величин. – Ученые записки МГПИ им. В.П.Потемкина. 1951. Т.XVI. Вып.3. С. 69-96.
24. Орлов А.И. Непараметрические оценки плотности в топологических пространствах. – В сб.: Прикладная статистика. Ученые записки по статистике, т.45. - М.: Наука, 1983. - С. 12-40.
25. Орлов А.И. Ядерные оценки плотности в пространствах произвольной природы. – В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. - Пермь: Пермский госуниверситет, 1996, с.68-75.
26. Ибрагимов И.А., Хасьминский Р.З. Асимптотическая теория оценивания. - М.: Наука, 1979. - 528 с.
27. Пакет программ анализа данных "ППАНД". Учебное пособие / Орлов А.И., Легостаева И.Л. и еще 9 соавторов. - М.: Сотрудничающий центр ВОЗ по профессиональной гигиене, 1990. - 93 с.

## Контрольные вопросы и задачи

1. Часто ли результаты измерений имеют нормальное распределение?

2. По выборке фактических данных о величине годового дохода (в тыс. долл.), взятых на конец 1970-х гг. (США), постройте вариационный ряд, гистограмму (группируя данные по 6-ти равным интервалам); определить выборочные среднее арифметическое, медиану и моду:

2,0; 13,4; 2,2; 6,7; 11,1; 10,0; 2,6; 12,9; 10,5; 9,2; 11,1;
14,0; 26,0; 17,5; 7,2; 18,7; 9,9; 7,6; 11,7;11,3; 6,5.

3. Дано распределение по градациям (табл.6) почасовой зарплаты 303 рабочих, занятых в промышленности ($f_i$- число рабочих, имеющих почасовую зарплату $x_i$). Постройте эмпирическую функцию распределения, найти выборочные медиану, моду и среднее арифметическое.

Таблица 6.

Распределение рабочих по ставкам почасовой оплаты

| $x_i$ | 2,5 | 2,6 | 2,7 | 2,8 | 2,9 | 3,0 | 3,1 | 3,2 | 3,3 | 3,4 |
|---|---|---|---|---|---|---|---|---|---|---|
| $f_i$ | 10 | 25 | 41 | 74 | 58 | 34 | 17 | 14 | 11 | 3 |

4. Какие средние величины целесообразно использовать при расчете средней заработной платы (или среднего дохода)?

5. Постройте пример, показывающий некорректность использования среднего арифметического $f(X_1, X_2) = (X_1 + X_2)/2$ в порядковой шкале, используя допустимое преобразование $g(x) = x^2$ (при положительных усредняемых величинах $x$).

6. Постройте пример, показывающий некорректность использования среднего геометрического в порядковой шкале. Другими словами, приведите пример чисел $x_1$, $x_2$, $y_1$, $y_2$ и строго возрастающего преобразования $f$: $R^1 \to R^1$ таких, что

$$(x_1x_2)^{1/2} < (y_1y_2)^{1/2}, \quad [f(x_1)f(x_2)]^{1/2} > [f(y_1)f(y_2)]^{1/2}.$$

7. Приведите пример чисел $x_1$, $x_2$, $y_1$, $y_2$ и строго возрастающего преобразования $f$: $R^1 \to R^1$ таких, что

$$[(x_1)^2 + (x_2)^2]^{1/2} < [(y_1)^2 + (y_2)^2]^{1/2},$$
$$[(f(x_1))^2 + (f(x_2))^2]^{1/2} > [(f(y_1))^2 + (f(y_2))^2]^{1/2}.$$

8. Какая математическая модель используется для описания случайного множества?

9. Как соотносятся эмпирические и теоретические средние для числовых данных и в пространствах произвольной природы?

10. Почему описание числовых данных с помощью непараметрических оценок плотности предпочтительнее их описания с помощью гистограмм?

## Темы докладов, рефератов, исследовательских работ

1. Проведите описание данных, приведенных в табл.1 (подраздел 2.1.2). Постройте таблицы (типа табл. 2 и 3 там же), рассчитайте выборочные характеристики.
2. Показатели разброса, связи, показатели различия (в том числе метрики) в порядковой шкале.
3. Ранговые методы математической статистики как инвариантные методы анализа порядковых данных.
4. Показатели разброса, связи, показатели различия (в том числе метрики) в шкале интервалов.
5. Показатели разброса, связи, показатели различия (в том числе метрики) в шкале отношений.
6. Теорема В.В. Подиновского: любое изменение коэффициентов весомости единичных показателей качества продукции приводит к изменению упорядочения изделий по средневзвешенному показателю.
7. Вероятностные модели бинарных отношений.
8. Вероятностные модели парных сравнений.
9. Средние и законы больших чисел в пространстве упорядочений.
10. Непараметрические оценки плотности в непрерывных и дискретных пространствах.

## 2.2. Оценивание

### 2.2.1. Методы оценивания параметров

В прикладной статистике используются разнообразные параметрические модели. Термин «параметрический» означает, то вероятностно-статистическая модель полностью описывается конечномерным вектором фиксированной размерности. Причем эта размерность не зависит от объема выборки.

Рассмотрим выборку $x_1, x_2, \ldots, x_n$ из распределения с плотностью $f(x;и_0)$, где $f(x;и_0)$-элемент параметрического семейства плотностей распределения вероятностей $\{f(x;и), иєИ\}$. Здесь И – заранее известное $k$-мерное пространство параметров, являющееся подмножеством евклидова пространства $R^k$, а конкретное значение параметра $и_0$ статистику неизвестно. Обычно в прикладной статистике применяются параметрические семейства с $k$ = 1,2,3 (см. главу 1.2). В статистике нечисловых данных вместо плотности часто рассматриваются вероятности попадания в точки. Напомним, что в параметрических задачах оценивания принимают вероятностную модель, согласно которой результаты наблюдений $x_1, x_2, \ldots, x_n$ рассматривают как реализации $n$ независимых случайных величин.

Задача оценивания состоит в том, чтобы оценить неизвестное статистику значение параметра $и_0$ наилучшим (в каком-либо смысле) образом.

*Пример 1*. В статистических задачах стандартизации и управления качеством используют семейство гамма-распределений. Плотность гамма-распределения имеет вид

$$f(x;a,b,c)=\begin{cases}\dfrac{1}{\Gamma(a)}(x-c)^{a-1}b^{-a}\exp\left[-\dfrac{x-c}{b}\right], & x\ge c,\\ 0, & x<c.\end{cases} \qquad (1)$$

Плотность вероятности в формуле (1) определяется тремя параметрами $a, b, c$, где $a>2$, $b>0$. При этом $a$ является параметром формы, $b$ - параметром масштаба и $c$ - параметром сдвига. Множитель $1/\Gamma(a)$ является нормировочным, он введен, чтобы

$$\int_{-\infty}^{+\infty} f(x;a,b,c)dx=1.$$

Здесь $\Gamma(a)$ - одна из используемых в математике специальных функций, так называемая "гамма-функция", по которой названо и распределение, задаваемое формулой (1),

$$\Gamma(a)=\int_0^{+\infty} x^{a-1}e^{-x}dx.$$

Подробные решения задач оценивания параметров для гамма-распределения содержатся в разработанном нами государственном стандарте ГОСТ 11,011-83 «Прикладная статистика. Правила определения оценок и доверительных границ для параметров гамма-распределения» [1]. В настоящее время эта публикация используется в качестве методического материала для инженерно-технических работников промышленных предприятий и прикладных научно-исследовательских институтов.

Поскольку гамма-распределение зависит от трех параметров, то имеется $2^3 - 1 = 7$ вариантов постановок задач оценивания. Они описаны в табл.1.

Таблица 1.

Постановки задач оценивания для параметров гамма-распределения

| № п/п | Параметр формы | Параметр масштаба | Параметр сдвига |
|---|---|---|---|
| 1 | Известен | Оценивается | Известен |
| 2 | Оценивается | Известен | Известен |
| 3 | Известен | Известен | Оценивается |
| 4 | Оценивается | Оценивается | Известен |
| 5 | Известен | Оценивается | Оценивается |
| 6 | Оценивается | Известен | Оценивается |
| 7 | Оценивается | Оценивается | Оценивается |

В табл.2 приведены реальные данные о наработке резцов до предельного состояния, в часах. Упорядоченная выборка (вариационный ряд) объема $n = 50$ взята из государственного стандарта [1]. Проверка согласия данных о наработке резцов с семейством гамма-распределений проведена в главе 2.3. Именно эти данные будут служить исходным материалом для демонстрации тех или иных методов оценивания параметров.

Таблица 2.

Наработка резцов до предельного состояния, ч

| № п/п | Наработка, ч | № п/п | Наработка, ч | № п/п | Наработка, ч |
|---|---|---|---|---|---|
| 1 | 9 | 18 | 47,5 | 35 | 63 |
| 2 | 17,5 | 19 | 48 | 36 | 64,5 |
| 3 | 21 | 20 | 50 | 37 | 65 |
| 4 | 26,5 | 21 | 51 | 38 | 67,5 |
| 5 | 27,5 | 22 | 53,5 | 39 | 68,5 |
| 6 | 31 | 23 | 55 | 40 | 70 |
| 7 | 32,5 | 24 | 56 | 41 | 72,5 |
| 8 | 34 | 25 | 56 | 42 | 77,5 |
| 9 | 36 | 26 | 56,5 | 43 | 81 |
| 10 | 36,5 | 27 | 57,5 | 44 | 82,5 |
| 11 | 39 | 28 | 58 | 45 | 90 |
| 12 | 40 | 29 | 59 | 46 | 96 |
| 13 | 41 | 30 | 59 | 47 | 101,5 |
| 14 | 42,5 | 31 | 60 | 48 | 117,5 |
| 15 | 43 | 32 | 61 | 49 | 127,5 |
| 16 | 45 | 33 | 61,5 | 50 | 130 |
| 17 | 46 | 34 | 62 | | |

Выбор «наилучших» оценок в определенной параметрической модели прикладной статистики – научно-исследовательская работа, растянутая во времени. Выделим два этапа. *Этап асимптотики*: оценки строятся и сравниваются по их свойствам при безграничном росте объема выборки. На этом этапе рассматривают такие характеристики оценок, как состоятельность, асимптотическая эффективность и др. *Этап конечных объемов выборки:* оценки сравниваются, скажем, при $n = 10$. Ясно, что исследование начинается с этапа асимптотики: чтобы сравнивать оценки, надо сначала их построить и быть уверенными, что они не являются абсурдными (такую уверенность дает доказательство состоятельности).

С какой оценки начинать? Одним из наиболее известных и простых в употреблении методов является метод моментов. Название связано с тем, что этот метод опирается на использование выборочных моментов

$$M_{nm} = \frac{1}{n}\sum_{i=1}^{n} x_i^m, \quad m = 1,2,...$$

где $x_1, x_2, ..., x_n$ – выборка, т.е. набор независимых одинаково распределенных случайных величин с числовыми значениями.

В прикладной статистике метод анализа данных называется *методом моментов*, если он использует статистику

$$Y_n = g(M_{n1}, M_{n2}, ..., M_{nq}), \quad (2)$$

где $g$: $R^q \to R^k$ – некоторая функция (здесь $k$ – число неизвестных числовых параметров). Чаще всего термин «метод моментов» используют, когда речь идет об оценивании параметров. В этом случае обычно предполагают, что плотность вероятности распределения элементов выборки $f(x)$ входит в заранее известное статистику параметрическое семейство $\{f(x;и), иєИ\}$, т.е. $f(x) = f(x;и_0)$ при некотором $и_0$. Здесь И – заранее заданное $k$-мерное пространство параметров, являющееся подмножеством евклидова пространства $R^k$, а конкретное значение параметра $и_0$ статистику неизвестно, его и следует оценить. Известно также, что неизвестный

параметр определяется с помощью известной статистику функции через начальные моменты элементов выборки:

$$\theta_0 = g(a_1, a_2, ..., a_q), \quad a_m = M(x_i^m), \quad m = 1,2,... \quad (3)$$

В методе моментов в качестве оценки $и_0$ используют статистику $Y_n$ вида (2), которая отличается от формулы (2) тем, что теоретические моменты заменены выборочными.

Статистики $Y_n$ вида (2) применяются не только для оценивания параметров, но и для непараметрического оценивания характеристик случайной величины, таких, как коэффициент вариации, и для проверки гипотез. Во всех случаях применения статистики $Y_n$ вида (2) говорят о методе моментов.

Распределение вектора $Y_n$ во всех практически важных случаях является асимптотически нормальным. Это утверждение опирается на следующий общий факт.

Пусть случайный вектор $Z_n$ є $R^q$ асимптотически нормален с математическим ожиданием $z_\infty$ и ковариационной матрицей $||c_{ij}||/n$, а функция $h$: $R^q \to R^1$ достаточно гладкая. Тогда случайная величина $h(Z_n)$ асимптотически нормальна с математическим ожиданием $h(z_\infty)$ и дисперсией

$$\sigma^2 = \frac{1}{n}\sum_{r=1}^{q}\sum_{s=1}^{q}\frac{\partial h}{\partial x_r}\frac{\partial h}{\partial x_s}c_{rs}. \quad (4)$$

Этот способ нахождения предельного распределения известен как д-метод Рао [2], метод линеаризации [3]. Последний термин и будем использовать. Условия регулярности, накладываемые на распределение случайной величины $Z_n$ и функцию $h$, при которых метод линеаризации обоснован, хорошо известны (см. [4], [2, с.337-339], а также главу 1.4 настоящего учебника).

Для получения асимптотического распределения статистики $Y_n$ вида (2) можно применить метод линеаризации к асимптотически нормальному вектору выборочных моментов ($M_{n1}$, $M_{n2}$, …, $M_{nq}$) и функции $g$ из формулы (2).

В силу многомерной центральной предельной теоремы (см. главу 1.4) указанная асимптотическая нормальность имеет место, если, например,

$$M|x_i|^{2q+1} < +\infty.$$

Это условие выполнено, в частности, для результатов измерений, распределения которых сосредоточены на ограниченных сверху и снизу интервалах.

При реализации намеченного плана для применения формулы (4) необходимо использовать асимптотические дисперсии и ковариации выборочных моментов, т.е. величины, обозначенные в формуле (4) как $c_{rs}$ . Эти величины имеют вид [2, с.388]:

$$c_{rr} = \mu_{2r} - \mu_r^2 - 2r\mu_{r-1}\mu_{r+1} + r^2\mu_{r-1}^2\mu_2,$$

$$c_{rs} = \mu_{r+s} - \mu_r\mu_s + rs\mu_2\mu_{r-1}\mu_{s-1} - r\mu_{r-1}\mu_{s+1} - s\mu_{r+1}\mu_{s-1}, \quad r,s = 1,2,..., \quad \mu_0 = 0. \quad (5)$$

Здесь $м_г$ – теоретический центральный момент порядка $r$, т.е.

$$\mu_r = M(x_i - M(x_i))^r, \quad r = 1,2,...$$

Таким образом, для получения асимптотического распределения случайной величины $Y_n$ вида (2) достаточно знать теоретические центральные моменты результатов наблюдений и вид функции $g$. Отметим, что асимптотическим смещением оценок в рассматриваемом случае можно пренебречь, поскольку его вклад в средний квадрат ошибки статистической оценки – бесконечно малая величина более высокого порядка по сравнению с асимптотической дисперсией.

Однако моменты неизвестны. Их приходится оценивать. В соответствии с теоремами о наследовании сходимости для нахождения асимптотического распределения функции от выборочных моментов можно воспользоваться не теоретическими моментами, а их состоятельными оценками. Эти оценки можно получить разными способами. Можно непосредственно применить формулы (5), заменив теоретические моменты выборочными. Можно выразить моменты через параметры рассматриваемого распределения. Можно применять более сложные процедуры, например, на основе непараметрических устойчивых

(робастных) оценок моментов типа урезанных средних Пуанкаре и др. (в первой в России книге по общей теории устойчивости [5] проблематика робастных оценок рассмотрена в главе 2).

Для оценивания параметров гамма-распределения воспользуемся известной формулой [6, с.184-185], согласно которой для случайной величины $X$, имеющей гамма-распределение с параметрами формы $a$, масштаба $b$ =1 и сдвига c=0,

$$M(X^m) = \frac{\Gamma(a+m)}{\Gamma(a)} = a(a+1)...(a+m-1), \quad m = 1,2,... \quad (6)$$

Следовательно, $M(X) = a$, $M(X^2) = a(a+1)$, $D(X) = M(X^2) - (M(X))^2 = a(a+1) - a^2 = a$. Найдем третий центральный момент $M(X - M(X))^3$. Справедливо равенство

$$M(X - M(X))^3 = M(X^3) - 3\, M(X^2)\, M(X) + 3\, M(\mathrm{X})\, (M(X))^2 - (M(X))^3.$$

Из равенства (6) вытекает, что

$$M(X - M(X))^3 = a(a+1)(a+2) - 3\, a\, (a+1)\, a + 3\, a\, a^2 - a^3 = 2a.$$

Если $Y$ – случайная величина, имеющая гамма-распределение с произвольными параметрами формы $a$, масштаба $b$ и сдвига $c$, то $Y = bX + c$. Следовательно, $M(Y) = ab+c$, $D(Y) = ab^2$, $M(Y - M(Y))^3 = 2\, a\, b^3$.

*Пример 2.* Оценивание методом моментов параметров гамма-распределения в случае трех неизвестных параметров (строка 7 таблицы 1).

В соответствии с проведенными выше рассуждениями для оценивания трех параметров достаточно использовать три выборочных момента – выборочное среднее арифметическое

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n},$$

выборочную дисперсию

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

и выборочный третий центральный момент

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3.$$

Приравнивая теоретические моменты, выраженные через параметры распределения, и выборочные моменты, получаем систему уравнений метода моментов:

$$ab + c = \bar{x}, \quad ab^2 = s^2, \quad 2ab^3 = m_3.$$

Решая эту систему, находим оценки метода моментов. Подставляя второе уравнение в третье, получаем оценку метода моментов для параметра сдвига:

$$2s^2 b = m_3, \quad b^* = \frac{1}{2}\frac{m_3}{s^2}.$$

Подставляя эту оценку во второе уравнение, находим оценку метода моментов для параметра формы:

$$a(b^*)^2 = a\left(\frac{1}{2}\frac{m_3}{s^2}\right)^2 = \frac{a}{4}\frac{m_3^2}{s^4} = s^2, \quad a^* = 4\frac{s^6}{m_3^2}.$$

Наконец, из первого уравнения находим оценку для параметра сдвига:

$$c^* = \bar{x} - a^* b^* = \bar{x} - 4\frac{s^6}{m_3^2}\frac{1}{2}\frac{m_3}{s^2} = \bar{x} - 2\frac{s^4}{m_3}.$$

Для реальных данных [1], приведенных выше в табл.2, выборочное среднее арифметическое $\bar{x}$ = 57,88, выборочная дисперсия $s^2$ = 663,00, выборочный третий центральный момент $m_3$ = 14927,91. Согласно только что полученным формулам оценки метода моментов таковы: $a^*$ = 5,23; $b^*$ = 11,26, $c^*$ = - 1,01.

Оценки параметров гамма-распределения, полученные методом моментов, являются функциями от выборочных моментов. В соответствии со сказанным выше они являются асимптотически нормальными случайными величинами. Их распределения аппроксимируются нормальными распределениями, математические ожидания которых равны соответствующим параметрам, а дисперсии находятся с помощью формулы (4) с учетом формул (5) и (6). В табл.3

приведены оценки метода моментов и их асимптотические дисперсии при различных вариантах сочетания известных и неизвестных параметров гамма-распределения.

Таблица 3.

Оценки метода моментов и их асимптотические дисперсии

| № п/п | Описание вероятностной модели | | | Оцени-ваемый пара-метр | Вид оценки | Асимптотическая дисперсия оценки |
|---|---|---|---|---|---|---|
| | *a* | *b* | *c* | | | |
| 1 | - | - | + | *a* | $\frac{(\bar{x})^2}{s^2}$ | $\frac{2a(a+1)}{n}$ |
| 2 | - | - | + | *b* | $\frac{s^2}{\bar{x}}$ | $\frac{b^2}{n}\left(2+\frac{3}{a}\right)$ |
| 3 | - | - | - | *a* | $4\frac{s^6}{m_3^2}$ | $\frac{6a}{n}(a^2+6a+5)$ |
| 4 | - | - | - | *b* | $\frac{1}{2}\frac{m_3}{s^2}$ | $\frac{b^2}{2an}(6a^2+25a+24)$ |
| 5 | - | - | - | *c* | $\bar{x}-2\frac{s^4}{m_3}$ | $\frac{ab^2}{n}(3a^2+13a+10)$ |
| 6 | + | - | - | *b* | $\frac{s}{\sqrt{a}}$ | $\frac{b}{2n}(a+3)$ |
| 7 | + | - | - | *c* | $\bar{x}-s\sqrt{a}$ | $\frac{ab^2}{2n}(a+1)$ |
| 8 | - | + | - | *A* | $\frac{s^2}{b^2}$ | $\frac{2a}{n}(a+3)$ |
| 9 | - | + | - | *c* | $\bar{x}-\frac{s^2}{b}$ | $\frac{ab^2}{n}(2a+3)$ |
| 10 | + | + | - | *c* | $\bar{x}-ab$ | $\frac{ab^2}{n}$ |

*Примечание.* При описании вероятностной модели известные статистику параметры отмечены плюсами, оцениваемые – минусами.

Все оценки метода моментов, приведенные в табл.3, включены в государственный стандарт [1]. Они охватывают все постановки задач оценивания параметров гамма-распределения (см. табл.1) , кроме тех, когда неизвестен только один параметр – *a* или *b*. Для этих исключительных случаев в [1] разработаны специальные методы оценивания.

Поскольку асимптотическое распределение оценок метода моментов известно, то не представляет труда формулировка правил проверки статистических гипотез относительно значений параметров распределений, а также построение доверительных границ для параметров. Например, в вероятностной модели, когда все три параметра неизвестны, в соответствии с третьей строкой таблицы 3 нижняя доверительная граница для параметра *a*, соответствующая доверительной вероятности г = 0,95, в асимптотике имеет вид

$$a_H = a^* - 1{,}96\left\{\frac{6a^*}{n}([a^*]^2+6a^*+5)\right\}^{1/2},$$

а верхняя доверительная граница для той же доверительной вероятности такова

$$a_B = a^* + 1{,}96\left\{\frac{6a^*}{n}([a^*]^2+6a^*+5)\right\}^{1/2},$$

где $a^*$ - оценка метода моментов параметра формы (табл.3).

Метод моментов является универсальным. Однако получаемые с его помощью оценки лишь в редких случаях обладают оптимальными свойствами. Поэтому в прикладной статистике применяют и другие виды оценок.

В работах, предназначенных для первоначального знакомства с математической статистикой, обычно рассматривают оценки максимального правдоподобия (сокращенно ОМП):

$$\theta_0(n) = \theta_0(n; x_1, x_2, ..., x_n) = Arg \min_{\theta \in \Theta} \prod_{i=1}^{n} f(x_i; \theta). \quad (7)$$

Таким образом, сначала строится плотность распределения вероятностей, соответствующая выборке. Поскольку элементы выборки независимы, то эта плотность представляется в виде произведения плотностей для отдельных элементов выборки. Совместная плотность рассматривается в точке, соответствующей наблюденным значениям. Это выражение как функция от параметра (при заданных элементах выборки) называется функцией правдоподобия. Затем тем или иным способом ищется значение параметра, при котором значение совместной плотности максимально. Это и есть оценка максимального правдоподобия.

Хорошо известно, что оценки максимального правдоподобия входят в класс наилучших асимптотически нормальных оценок (определение дано ниже). Однако при конечных объемах выборки в ряде задач ОМП недопустимы, т.к. они хуже ( дисперсия и средний квадрат ошибки больше), чем другие оценки, в частности, несмещенные [6]. Именно поэтому в ГОСТ 11.010-81 для оценивания параметров отрицательного биномиального распределения используются несмещенные оценки, а не ОМП [7]. Из сказанного следует априорно предпочитать ОМП другим видам оценок можно – если можно – лишь на этапе изучения асимптотического поведения оценок.

В отдельных случаях ОМП находятся явно, в виде конкретных формул, пригодных для вычисления.

*Пример 3.* Найдем ОМП для выборки из нормального распределения, каждый элемент которой имеет плотность

$$f(x; m, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}.$$

Таким образом, надо оценить двумерный параметр ($m$, $y^2$).

Произведение плотностей вероятностей для элементов выборки, т.е. функция правдоподобия, имеет вид

$$H(m; \sigma^2) = \sigma^{-n} (2\pi)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2\right\}. \quad (8)$$

Требуется решить задачу оптимизации

$$H(m; \sigma^2) \to \max.$$

Как и во многих иных случаях, задача оптимизации проще решается, если прологарифмировать функцию правдоподобия, т.е. перейти к функции

$$h(m; \sigma^2) = \ln H(m; \sigma^2),$$

называемой логарифмической функцией правдоподобия. Для выборки из нормального распределения

$$h(m; \sigma^2) = (-n) \ln \sigma + \left(-\frac{n}{2}\right) \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2. \quad (9)$$

Необходимым условием максимума является равенство 0 частных производных от логарифмической функции правдоподобия по параметрам, т.е.

$$\frac{\partial h(m, \sigma^2)}{\partial m} = 0, \quad \frac{\partial h(m, \sigma^2)}{\partial (\sigma^2)} = 0. \quad (10)$$

Система (10) называется системой уравнений максимального правдоподобия. В общем случае число уравнений равно числу неизвестных параметров, а каждое из уравнений выписывается

путем приравнивания 0 частной производной логарифмической функции правдоподобия по тому или иному параметру.

При дифференцировании по $m$ первые два слагаемых в правой части формулы (9) обращаются в 0, а последнее слагаемое дает уравнение

$$\frac{\partial}{\partial m}\sum_{i=1}^{n}(x_i-m)^2=\sum_{i=1}^{n}2(x_i-m)(-1)=0,\ \ \sum_{i=1}^{n}x_i=nm\ .$$

Следовательно, оценкой $m^*$ максимального правдоподобия параметра $m$ является выборочное среднее арифметическое,

$$m^*=\bar{x}\ .$$

Для нахождения оценки дисперсии необходимо решить уравнение

$$\frac{\partial}{\partial(\sigma^2)}h(m;\sigma^2)=\frac{\partial}{\partial(\sigma^2)}(-n)\ln\sqrt{\sigma^2}-\frac{\partial}{\partial(\sigma^2)}\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2=0.$$

Легко видеть, что

$$\frac{\partial}{\partial(\sigma^2)}(-n)\ln\sqrt{\sigma^2}=\frac{(-n)}{2\sigma^2},\ \ -\frac{\partial}{\partial(\sigma^2)}\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2=\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-m)^2.$$

Следовательно, оценкой $(\text{у}^2)^*$ максимального правдоподобия для дисперсии $\text{у}^2$ с учетом найденной ранее оценки для параметра $m$ является выборочная дисперсия,

$$(\sigma^2)^*=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Итак, система уравнений максимального правдоподобия решена аналитически, ОМП для математического ожидания и дисперсии нормального распределения – это выборочное среднее арифметическое и выборочная дисперсия. Отметим, что последняя оценка является смещенной.

Отметим, что в условиях примера 3 оценки метода максимального правдоподобия совпадают с оценками метода моментов. Причем вид оценок метода моментов очевиден и не требует проведения каких-либо рассуждений.

В большинстве случаев аналитических решений не существует, для нахождения ОМП необходимо применять численные методы. Так обстоит дело, например, с выборками из гамма-распределения или распределения Вейбулла-Гнеденко. Во многих работах каким-либо итерационным методом решают систему уравнений максимального правдоподобия ([8] и др.) или впрямую максимизируют функцию правдоподобия типа (8) (см. [9] и др.).

Однако применение численных методов порождает многочисленные проблемы. Сходимость итерационных методов требует обоснования. В ряде примеров функция правдоподобия имеет много локальных максимумов, а потому естественные итерационные процедуры не сходятся [10]. Для данных ВНИИ железнодорожного транспорта по усталостным испытаниям стали уравнение максимального правдоподобия имеет 11 корней [11]. Какой из одиннадцати использовать в качестве оценки параметра?

Как следствие осознания указанных трудностей, стали появляться работы по доказательству сходимости алгоритмов нахождения оценок максимального правдоподобия для конкретных вероятностных моделей и конкретных алгоритмов. Примером является статья [12].

Однако теоретическое доказательство сходимости итерационного алгоритма – это еще не всё. Возникает вопрос об обоснованном выборе момента прекращения вычислений в связи с достижением требуемой точности. В большинстве случаев он не решен.

Но и это не все. Точность вычислений необходимо увязывать с объемом выборки – чем он больше, тем точнее надо находить оценки параметров, в противном случае нельзя говорить о состоятельности метода оценивания. Более того, при увеличении объема выборки необходимо увеличивать и количество используемых в компьютере разрядов, переходить от одинарной точности расчетов к двойной и далее – опять-таки ради достижения состоятельности оценок.

Таким образом, при отсутствии явных формул для оценок максимального правдоподобия нахождение ОМП натыкается на ряд проблем вычислительного характера. Специалисты по математической статистике позволяют себе игнорировать все эти проблемы, рассуждая об ОМП в теоретическом плане. Однако прикладная статистика не может их

игнорировать. Отмеченные проблемы ставят под вопрос целесообразность практического использования ОМП.

Нет необходимости абсолютизировать ОМП. Кроме них, существуют другие виды оценок, обладающих хорошими статистическими свойствами. Примером являются одношаговые оценки (ОШ-оценки).

В прикладной статистике разработано много видов оценок. Упомянем квантильные оценки. Они основаны на идее, аналогичной методу моментов, но только вместо выборочных и теоретических моментов приравниваются выборочные и теоретические квантили. Другая группа оценок базируется на идее минимизации расстояния (показателя различия) между эмпирическими данными и элементом параметрического семейства. В простейшем случае минимизируется евклидово расстояние между эмпирическими и теоретическими гистограммами, а точнее, векторами, составленными из высот столбиков гистограмм.

### 2.2.2. Одношаговые оценки

Одношаговые оценки имеют столь же хорошие асимптотические свойства, что и оценки максимального правдоподобия, при тех же условиях регулярности, что и ОМП. Грубо говоря, они представляют собой результат первой итерации при решении системы уравнений максимального правдоподобия по методу Ньютона-Ватсона. Одношаговые оценки выписываются в виде явных формул, а потому требуют существенно меньше машинного времени, а также могут применяться при ручном счете (на калькуляторах). Снимаются вопросы о сходимости алгоритмов, о выборе момента прекращения вычислений, о влиянии округлений при вычислениях на окончательный результат. ОШ оценки были использованы нами при разработке ГОСТ 11.011-83 вместо ОМП.

Как и раньше, рассмотрим выборку $x_1, x_2, \ldots, x_n$ из распределения с плотностью $f(x;и_0)$, где $f(x;и_0)$- элемент параметрического семейства плотностей распределения вероятностей $\{f(x;и), иєИ\}$. Здесь И – известное статистику $k$-мерное пространство параметров, являющееся подмножеством евклидова пространства $R^k$, а конкретное значение параметра $и_0$ неизвестно. Его и будем оценивать.

Обозначим и =($и^1$, $и^2$,…, $и^k$). Рассмотрим вектор-столбец частных производных логарифма плотности вероятности

$$s(x;\theta) = \left\| \frac{\partial}{\partial \theta^{\alpha}} \ln f(x;\theta), \quad \alpha = 1,2,...,k \right\|$$

и матрицу частных производных второго порядка для той же функции

$$b(x;\theta) = \left\| \frac{\partial^2}{\partial \theta^{\alpha} \partial \theta^{\beta}} \ln f(x;\theta), \quad \alpha, \beta = 1,2,...,k \right\|.$$

Положим

$$s_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} s(x_i,\theta), \quad b_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} b(x_i,\theta).$$

Пусть матрица информации Фишера $I(и_0) = M[-b_n(и_0)]$ положительно определена.

*Определение 1* [10, с.269]. Оценку и($n$) параметра $и_0$ называют наилучшей асимптотически нормальной оценкой (сокращенно НАН-оценкой), если распределение случайного вектора $\sqrt{n}(\theta(n) - \theta_0)$ сходится при n $\to \infty$ к нормальному распределению с нулевым математическим ожиданием и ковариационной матрицей, равной $I^{-1}(и_0)$.

Определение 1 корректно: $I^{-1}(и_0)$ является нижней асимптотической границей для ковариационной матрицы случайного вектора $\sqrt{n}(\theta^*(n) - \theta_0)$, где и*($n$) – произвольная оценка; кроме ОМП есть НАН-оценки (см. [10] и др.). Некоторые другие оценки также являются НАН-оценками, например, байесовские. Сказанное об ОМП и байесовских оценках справедливо при некоторых условиях регулярности (см., например, [13]). В ряде случаев несмещенные оценки являются НАН-оценками, более того, они лучше, чем ОМП (их дисперсия меньше), при конечных объемах выборки [6].

Для анализа реальных данных естественно рекомендовать какую-либо из НАН-оценок. (Это утверждение всегда верно на этапе асимптотики при изучении конкретной задачи прикладной статистики. Теоретически можно предположить, что при тщательном изучении для конкретных конечных объемов выборки наилучшей окажется какая-либо оценка, не являющаяся НАН-оценкой. Однако такие ситуации нам пока не известны.)

Пусть $и_1(n)$ и $I_n^{-1}$ - некоторые оценки $и_0$ и $I^{-1}(и_0)$ соответственно.

*Определение 2.* Одношаговой оценкой (ОШ-оценкой, или ОШО) называется оценка

$$\theta_2(n) = \theta_1(n) + I_n^{-1} s_n(\theta_1(n)).$$

*Теорема 1* [14]. Пусть выполнены следующие условия.

(I) Распределение $\sqrt{n} s_n(\theta_0)$ сходится при $n \to \infty$ к нормальному распределению с математическим ожиданием 0 и ковариационной матрицей $I(и_0)$ и, кроме того, существует $Mb_n(\theta_0) b_n'(\theta_0)$.

(II) При некотором $e > 0$ и $n \to \infty$

$$\sup_{\theta: 0<|\theta-\theta_0|<\varepsilon} \frac{| s_n(\theta) - s_n(\theta_0) - b_n(\theta_0)(\theta-\theta_0) |}{| \theta - \theta_0 |^2} = O_p(1).$$

(III) Для любого $e > 0$

$$\lim_{n\to\infty} P\left\{ n^{1/4} (| \theta_1(n) - \theta_0 | + \| I_n^{-1} - I^{-1}(\theta_0) \|) > \varepsilon \right\} = 0.$$

Тогда ОШ-оценка является НАН-оценкой.

*Доказательство*. Рассмотрим тождество

$$\sqrt{n}(\theta_2(n) - \theta_0) = \sqrt{n}(\theta_1(n) - \theta_0) + \sqrt{n} I_n^{-1} s_n(\theta_1(n)). \quad (1)$$

В силу условия (II) теоремы

$$\sqrt{n} I_n^{-1} s_n(\theta_1(n)) = \sqrt{n} I_n^{-1} s_n(\theta_0) + \sqrt{n} I_n^{-1} b_n(\theta_0)(\theta_1(n) - \theta_0) + \sqrt{n} I_n^{-1} O_p(| \theta_1(n) - \theta_0 |^2). \quad (2)$$

Из условия (I) теоремы следует, что первое слагаемое в правой части формулы (2) сходится при $n \to \infty$ по распределению к нормальному закону с математическим ожиданием 0 и ковариационной матрицей $I^{-1}(и_0)$. Согласно условию (III)

$$\sqrt{n} \, | \theta_1(n) - \theta_0 |^2 \to 0$$

по вероятности. Кроме того, согласно тому же условию последовательность матриц $I_n^{-1}$ ограничена по вероятности. Поэтому третье слагаемое в правой части формулы (2) сходится к 0 по вероятности. Для завершения доказательства теоремы осталось показать, что

$$\sqrt{n}(\theta_1(n) - \theta_0) + \sqrt{n} I_n^{-1} b_n(\theta_0)(\theta_1(n) - \theta_0) \to 0 \quad (3)$$

по вероятности. Левая часть формулы (3) преобразуется к виду

$$(E + I_n^{-1} b_n(\theta_0)) \sqrt{n}(\theta_1(n) - \theta_0), \quad (4)$$

где $E$ – единичная матрица. Поскольку из условия (I) теоремы следует, что для $b_n(и_0)$ справедлива (многомерная) центральная предельная теорема, то

$$b_n(\theta_0) = -I(\theta_0) + O_p(n^{-1/2}).$$

С учетом условия (III) теоремы заключаем, что

$$E + I_n^{-1} b_n(\theta_0) = o_p(n^{-1/4}). \quad (5)$$

Из соотношений (4), (5) и условия (III) теоремы вытекает справедливость формулы (3). Теорема доказана.

Прокомментируем условия теоремы. Условия (I) и (II) обычно предполагаются справедливыми при рассмотрении оценок максимального правдоподобия [10]. Эти условия можно выразить в виде требований, наложенных непосредственно на плотность $f(x;и)$ из параметрического семейства, как это сделано, например, в [13]. Условие (III) теоремы, наложенное на исходные оценки, весьма слабое. Обычно используемые оценки $и_1(n)$ и $I_n^{-1}$ являются не $n^{-1/4}$-состоятельными, а $\sqrt{n}$-состоятельными, т.е. условие (III) заведомо выполняется.

Какие оценки годятся в качестве начальных? В качестве $и_1(n)$ можно использовать оценки метода моментов, как это сделано в ГОСТ 11.011-83 [1], или, например, квантильные. В качестве $I_n^{-1}$ в теоретической работе [10] предлагается использовать простейшую оценку

$$I_n^{-1} = -b_n^{-1}(\theta_1(n)). \qquad (6)$$

Для гамма-распределения с неизвестными параметрами формы, масштаба и сдвига ОШ-оценки применены в [1]. При этом оценка (6) оказалась непрактичной, поскольку с точностью до погрешностей измерений и вычислений $\det(b_n) = 0$ для реальных данных о наработке резцов до предельного состояния, приведенных выше в табл.2 (пункт 2.2.1). Поскольку $\det(b_n) = 0$, то обратная матрица не существует, вычисления по формуле (6) невозможны. Поэтому в [1] в качестве ОШ-оценки была применена непосредственно первая итерация метода Ньютона-Рафсона решения системы уравнений максимального правдоподобия, т.е. была использована оценка

$$I_n^{-1} = I^{-1}(\theta_1(n)). \qquad (7)$$

В формуле (7) непосредственно используется явный вид зависимости матрицы информации Фишера от неизвестных параметров распределения.

В других случаях выбор тех или иных начальных оценок, в частности, выбор между (6) и (7), может определяться, например, простотой вычислений. Можно использовать также устойчивые аналоги [5] перечисленных выше оценок.

Полезно отметить, что еще в 1925 г., т.е. непосредственно при разработке метода максимального правдоподобия, его создатель Р.Фишер считал, что первая итерация по методу Ньютона-Рафсона дает хорошую оценку вектору неизвестных параметров [10, с.298]. Он однако рассматривал эту оценку как аппроксимацию ОМП. А.А.Боровков воспринимает ОШ-оценки как способ «приближенного вычисления оценок максимального правдоподобия» [15, с.225] и показывает асимптотическую эквивалентность ОШ-оценок и ОМП (в более сильных предположениях, чем в теореме 1; другими словами, теорема 1 обобщает результаты А.А. Боровкова относительно ОШ-оценок). Мы же полагаем, что ОШ-оценки имеют самостоятельную ценность, причем не меньшую, а в ряде случаев большую, чем ОМП. По нашему мнению, ОМП целесообразно применять (на этапе асимптотики) только тогда, когда они находятся явно. Во всех остальных случаях следует использовать на этом этапе ОШ-оценки (или какие-либо иные, выбранные из дополнительных соображений).

С чем связана популярность оценок максимального правдоподобия? Из всех НАН-оценок они наиболее просто вводятся, ранее других предложены. Поэтому среди математиков сложилась устойчивая традиция рассматривать ОМП в курсах математической статистики. Однако при этом игнорируются вычислительные вопросы, а также отодвигаются в сторону многочисленные иные НАН оценки.

В прикладной статистике – иные приоритеты. На первом месте – ОШ-оценки, все остальные НАН-оценки, в том числе ОМП, рассматриваются в качестве дополнительных возможностей.

*Пример 1*. Найдем ОШ-оценки для гамма-распределения с плотностью

$$f(x;a,b,c) = \begin{cases} \dfrac{1}{\Gamma(a)}(x-c)^{a-1}b^{-a}\exp\left[-\dfrac{x-c}{b}\right], & x \ge c, \\ 0, & x < c. \end{cases} \qquad (8)$$

Плотность вероятности в формуле (8) определяется тремя параметрами *a, b, c*, где $a>0$, $b>0$. При этом *a* является параметром формы, *b* - параметром масштаба и *c* - параметром сдвига. Здесь *Г(a)* - одна из используемых в математике специальных функций, так называемая "гамма-функция", по которой названо и распределение, задаваемое формулой (8),

$$\Gamma(a) = \int_0^{+\infty} x^{a-1}e^{-x}dx.$$

Как следует из явного вида плотности (8), логарифмическая функция правдоподобия имеет вид [16, с. 98]:

$$L = \sum_{i=1}^{n} \ln f(x_i; a, b, c) = -n \ln \Gamma(a) - na \ln b + (a-1) \sum_{i=1}^{n} \ln(x_i - c) - \frac{1}{b} \sum_{i=1}^{n} x_i + \frac{nc}{b},$$

а уравнения правдоподобия таковы:

$$\frac{\partial L}{\partial a} = -n\Psi(a) + \sum_{i=1}^{n} \ln\left(\frac{x_i - c}{b}\right) = 0,$$

$$\frac{\partial L}{\partial b} = -\frac{na}{b} + \frac{1}{b^2} \sum_{i=1}^{n} (x_i - c) = 0,$$

$$\frac{\partial L}{\partial c} = -(a-1) \sum_{i=1}^{n} \frac{1}{x_i - c} + \frac{n}{b} = 0,$$

где

$$\Psi(a) = \frac{d}{da} \ln \Gamma(a).$$

Ясно, что выписанная система нелинейных уравнений не имеет аналитического решения, в отличие от аналогичной системы для семейства нормальных распределений. Построим ОШ-оценки для задачи оценивания трех неизвестных параметров [17].

В качестве начальных оценок $и_1(n)$ будем использовать оценки метода моментов (см. пункт 2.2.1):

$$a^* = 4 \frac{s^6}{m_3^2}, \quad b^* = \frac{1}{2} \frac{m_3}{s^2}, \quad c^* = \bar{x} - a^* b^*,$$

где $\bar{x}$ - выборочное среднее арифметическое, $s^2$ – выборочная дисперсия, $m_3$ – выборочный третий центральный момент.

Матрица информации Фишера согласно [16, с.98] при $a > 2$ имеет вид

$$I(\theta) = I(a, b, c) = \begin{Vmatrix} \frac{d\Psi(a)}{da} & \frac{1}{b} & \frac{1}{b(a-1)} \\ \frac{1}{b} & \frac{a}{b^2} & \frac{1}{b^2} \\ \frac{1}{b(a-1)} & \frac{1}{b^2} & \frac{1}{b^2(a-2)} \end{Vmatrix}. \qquad (9)$$

Вектор-столбец частных производных логарифма плотности вероятности

$$s(x; \theta) = s(x; a, b, c) = (s(1), s(2), s(3))'$$

имеет координаты

$$s(1) = -\Psi(a) + \ln\left(\frac{x - c}{b}\right),$$

$$s(2) = -\frac{a}{b} + \frac{x - c}{b^2},$$

$$s(3) = -\frac{a-1}{x-c} + \frac{1}{b}.$$

Таким образом, для получения $s_n(a^*, b^*, c^*)$ необходимо вычислить две суммы

$$\sum_{i=1}^{n} \ln\left(\frac{x_i - c}{b}\right), \quad \sum_{i=1}^{n} \frac{1}{x_i - c}$$

и произвести еще несколько арифметических действий, число которых не зависит от объема выборки.

Одношаговые оценки $a_n$, $b_n$, $c_n$ для параметров гамма-распределения вычисляют по формуле

$$(a_n, b_n, c_n) = (a^*, b^*, c^*) + I^{-1}(a^*, b^*, c^*) s_n(a^*, b^*, c^*),$$

где $I^{-1}$ – обратная матрица к матрице информации Фишера $I$, заданной формулой (9). Матрицу $I^{-1}$ нетрудно рассчитать аналитически. Формулы для нахождения одношаговых оценок

расписаны в [1]. Расчеты облегчает то обстоятельство, что для гамма-распределения вторая координата вектора $s_n(a^*, b^*, c^*)$ тождественно равна 0, т.е. $s_n^{(2)}(a^*, b^*, c^*) \equiv 0$.

При $n \to \infty$ распределение вектора оценок $(a_n, b_n, c_n)$ приближается трехмерным нормальным распределением с математическим ожиданием, равным вектору истинных значений параметров $(a, b, c)$, и ковариационной матрицей $\Gamma^1(a_n, b_n, c_n)$. На этом приближении основаны правила расчета доверительных границ для параметров гамма-распределения [1]. Дисперсии оценок неизвестны, но зато имеются известные статистику зависимости этих дисперсий от параметров гамма-распределения. Эти зависимости непрерывные. Они стоят на главной диагонали ковариационной матрицы $\Gamma^1(a_n, b_n, c_n)$. Поэтому можно вместо неизвестных параметров подставить в них оценки этих параметров и на основе принципа наследования сходимости (глава 1.4 выше) получить состоятельные оценки дисперсий. Затем на основе оценок дисперсий обычным образом строятся доверительные интервалы для параметров гамма-распределения.

В табл.1 приведены результаты реализации описанной выше схемы расчетов - точечные и интервальные (при односторонней доверительной вероятности 0,95) оценки параметров гамма-распределения для данных, содержащихся в табл.2 предыдущего пункта 2.2.1.

Таблица 1.
Одношаговые оценки и доверительные границы
для параметров гамма-распределения

| Параметр | Одношаговая оценка | Верхняя доверительная граница | Нижняя доверительная граница |
|---|---|---|---|
| Формы | 7,32 | 16,41 | -1,77 |
| Масштаба | 8,77 | 15,24 | 2,30 |
| Сдвига | - 11,46 | 23,28 | - 46,20 |

Приведенные в табл.1 данные получены на основе асимптотических формул. Из-за конечности объема выборки необходимо внести некоторые коррективы. Поскольку параметр формы всегда положителен, $a > 0$, то нижняя доверительная граница для этого параметра должна быть неотрицательна, следует положить $a_H = 0$. Поскольку плотность гамма-распределения положительна только правее параметра $c$, то, очевидно, $c \le x_{min} = 9{,}00$, верхняя доверительная граница для параметра сдвига должна быть заменена на $c_B = 9{,}00$.

Может ли параметр сдвига быть отрицательным в данной прикладной задаче? Отрицательность параметра сдвига означает, что с положительной вероятностью рассматриваемая случайная величина отрицательна. Т.е. наработка резца до предельного состояния отрицательна. Ясно, что такого быть не может, хотя для специалиста по математической статистике отрицательность параметра сдвига вполне приемлема. Однако специалист по прикладной статистике должен признать неотрицательность параметра $c$ при обработке данных, составляющих рассматриваемую выборку. Следовательно, нижнюю доверительную границу для параметра сдвига необходимо заменить на $c_н = 0$.

Как следует из проведенных выше рассуждений и выкладок (см. также [16, с.98-100]), отношение дисперсий оценок метода моментов и ОШ-оценок имеет вид

$$\frac{Da_n}{Da^*} = \frac{\left\{(a-1)^3 + \frac{1}{5}(a-1)\right\}}{a(a+1)(a+5)}$$

при больших $a$. Это отношение, как и должно быть из общих соображений, всегда меньше 1. Отношение дисперсий возрастает при приближении к 0 коэффициента асимметрии распределения. Если $a > 39{,}1$ (коэффициент асимметрии меньше 0,102), то эффективность оценки метода моментов превышает 80%. При $a = 20$ (коэффициент асимметрии 0,20) она равна 65%. Напомним, что при безграничном росте параметра формы $a$ гамма-распределение приближается к нормальному, для которого оценки метода моментов и ОМП совпадают, а потому имеют равные дисперсии. Поэтому вполне естественно, что отношение дисперсий в формуле (10) стремится к 1 при безграничном росте параметра формы $a$.

Хотя дисперсии оценок метода моментов, как правило, меньше, чем дисперсии НАН-оценок, таких, как ОШО и ОМП, метод моментов играет большую роль в прикладной статистике. Во-первых, обычно их расчет проще (в частности, требует меньшего числа компьютерных операций), чем оценок других типов. К тому же оценки находятся с помощью выборочных моментов, которые, как правило, вычисляются на этапе описания статистических данных. Во-вторых, они служат основой для вычисления оценок других типов, например, ОШО. Для запуска итерационных методов нахождения ОМП также нужны начальные значения, и ими обычно являются оценки метода моментов. В-третьих, при учете погрешностей результатов наблюдений оценки метода моментов могут оказаться точнее ОМП и асимптотически эквивалентных им ОШО (см. главу 3.5 настоящего учебника).

Методы оценивания параметров гамма-распределения и примеры расчетов для всех семи постановок, перечисленных в табл.1 пункта 2.2.1, приведены в [1]. Большинство из них основано на асимптотических (при $n \to \infty$) теоретических результатах прикладной статистики. Методом статистических испытаний (Монте-Карло) показано, что уже при $n \geq 10$ используемые приближения удовлетворительны. Другими словами, асимптотической нормальностью оценок и другими важными для проведенных выше рассуждений предельными результатами можно пользоваться уже при $n \geq 10$.

Алгоритмическое и программное обеспечение ОШ-оценок для распределения Вейбулла-Гнеденко и гамма-распределения рассмотрено в содержательной монографии [18]. История вопроса освещена в статье [14].

### 2.2.3. Асимптотика решений экстремальных статистических задач

Если проанализировать приведенные выше в подразделе 2.1.5 постановки и результаты, касающиеся эмпирических и теоретических средних и законов больших чисел, то становится очевидной возможность их обобщения. Так, доказательства теорем практически не меняются, если считать, что функция $f(x,y)$ определена на декартовом произведении бикомпактных пространств $X$ и $Y$, а не на $X^2$. Тогда можно считать, что элементы выборки лежат в $X$, а $Y$ - пространство параметров, подлежащих оценке.

**Обобщения законов больших чисел**. Пусть, например, выборка $x_1 = x_1(\text{щ})$, $x_2 = x_2(\text{щ})$, … , $x_n = x_n(\text{щ})$ взята из распределения с плотностью $p(x,y)$, где $y$ – неизвестный параметр. Если положить

$$f(x,y) = - \ln p(x,y),$$

то задача нахождения эмпирического среднего

$$f_n(\omega, y) = \frac{1}{n}\sum_{k=1}^{n} f(x_k(\omega), y) \to \min$$

переходит в задачу оценивания неизвестного параметра $y$ методом максимального правдоподобия

$$\sum_{k=1}^{n} \ln p(x_k(\omega), y) \to \max .$$

Соответственно законы больших чисел переходят в утверждения о состоятельности этих оценок в случае пространств $X$ и $Y$ общего вида. При такой интерпретации функция $f(x,y)$ уже не является расстоянием или показателем различия. Однако для доказательства сходимости оценок к соответствующим значениям параметров это и не требуется. Достаточно непрерывности этой функции на декартовом произведении бикомпактных пространств $X$ и $Y$.

В случае функции $f(x,y)$ общего вида можно говорить об определении в пространствах произвольной природы оценок минимального контраста и их состоятельности. При этом при каждом конкретном значении параметра y справедливо предельное соотношение

$$f_n(\omega, y) = \frac{1}{n}\sum_{k=1}^{n} f(x_k(\omega), y) \to Mf(x_1(\omega), y) = g(y),$$

где $f$ – функция контраста. Тогда состоятельность оценок минимального контраста вытекает из справедливости предельного перехода

$$Arg\min\left\{\frac{1}{n}\sum_{k=1}^{n} f(x_k(\omega), y)\right\} \to Arg\min\{Mf(x_1(\omega), y)\}.$$

Частными случаями оценок минимального контраста являются, устойчивые (робастные) оценки Тьюки-Хубера (см. ниже), а также оценки параметров в задачах аппроксимации (параметрической регрессии) в пространствах произвольной природы.

Можно пойти и дальше в обобщении законов больших чисел. Пусть известно, что при каждом конкретном $y$ при безграничном росте $n$ имеет быть сходимость по вероятности

$$f_n(\omega, y) \to f(y),$$

где $f_n$(щ, $y$) – последовательность случайных функций на пространстве $Y$, а $f(y)$ – некоторая функция на $У$. В каких случаях и в каком смысле имеет место сходимость

$$\text{Argmin}\ \{f_n(\omega, y), y \in X\} \to \text{Argmin}\ \{f(y), y \in X\}?$$

Другими словами, когда из поточечной сходимости функций вытекает сходимость точек минимума?

Причем здесь можно под $n$ понимать натуральное число. А можно рассматривать сходимость по направленному множеству (подраздел 1.4.3), или же, что практически то же самое – «сходимость по фильтру» в смысле Картана и Бурбаки [19, с.118]. В частности, можно описывать ситуацию вектором, координаты которого - объемы нескольких выборок, и все они безгранично растут. В классической математической статистике такие постановки рассматривать не любят.

Поскольку, как уже отмечалось, основные задачи прикладной статистики можно представить в виде оптимизационных задач, то ответ на поставленный вопрос о сходимости точек минимума дает возможность единообразного подхода к изучению асимптотики решений разнообразных экстремальных статистических задач. Одна из возможных формулировок, основанная на бикомпактности пространств $X$ и $У$ и нацеленная на изучение оценок минимального контраста, дана и обоснована выше. Другой подход развит в работе [20]. Он основан на использовании понятий асимптотической равномерной разбиваемости и координатной асимптотической равномерной разбиваемости пространств. С помощью указанных подходов удается стандартным образом обосновывать состоятельность оценок характеристик и параметров в основных задачах прикладной статистики.

Рассматриваемую тематику можно развивать дальше, в частности, рассматривать аналоги законов больших чисел в случае пространств, не являющихся бикомпактными, а также изучать скорость сходимости $\text{Argmin}\{f_n(x(\omega), y), y \in X\}$ к $\text{Argmin}\{f(y), y \in X\}$.

Приведем примеры применения результатов о предельном поведении точек минимума.

**Задача аппроксимации зависимости (параметрической регрессии).** Пусть $X$ и $Y$ – некоторые пространства. Пусть имеются статистические данные - $n$ пар $(x_k, y_k)$, где $x_k \in X$, $y_k \in Y$, $k = 1, 2, \ldots, n$. Задано параметрическое пространство И произвольной природы и семейство функций $g(x,и)$: $X$ЧИ $\to Y$. Требуется подобрать параметр и $\in$ И так. чтобы $g(x_k, и)$ наилучшим образом приближали $y_k$, $k = 1, 2, \ldots, n$. Пусть $f_k$ – последовательность показателей различия в $У$. При сделанных предположениях параметр и естественно оценивать путем решения экстремальной задачи:

$$\theta_n = Arg\min_{\theta \in \Theta} \sum_{k=1}^{n} f_k(g(x_k, \theta), y_k). \qquad (1)$$

Часто, но не всегда, все $f_k$ совпадают. В классической постановке, когда $X = R^k$, $Y = R^1$, функции $f_k$ различны при неравноточных наблюдениях, например, когда число опытов меняется от одной точки $x$ проведения опытов к другой.

Если $f_k(y_1,y_2) = f(y_1,y_2) = (y_1 - y_2)^2$, то получаем общую постановку метода наименьших квадратов (см. подробности в главе 3.2):

$$\theta_n = Arg\min_{\theta \in \Theta} \sum_{k=1}^{n} (g(x_k, \theta) - y_k)^2.$$

В рамках детерминированного анализа данных остается единственный теоретический вопрос – о существовании $и_n$. Если все участвующие в формулировке задачи (1) функции непрерывны, а минимум берется по бикомпакту, то $и_n$ существует. Есть и иные условия существования $и_n$ [20-22].

При появлении нового наблюдения x в соответствии с методологией восстановления зависимости рекомендуется выбирать оценку соответствующего $y$ по правилу

$$y^* = g(x, и_n).$$

Обосновать такую рекомендацию в рамках детерминированного анализа данных невозможно. Это можно сделать только в вероятностной теории, равно как и изучить асимптотическое поведение $и_n$, доказать состоятельность этой оценки.

Кпк и в классическом случае, вероятностную теорию целесообразно строить для трех различных постановок.

1. Переменная $x$ – детерминированная (например, время), переменная $y$ – случайная, ее распределение зависит от $x$.

2. Совокупность $(x_k, y_k)$, $k = 1, 2, \ldots, n$, – выборка из распределения случайного элемента со значениями в $X$Ч$Y$.

3. Имеется детерминированный набор пар $(x_{k0}, y_{k0})$, $k = 1, 2, \ldots, n$, результат наблюдения $(x_k, y_k)$ является случайным элементом, распределение которого зависит от $(x_{k0}, y_{k0})$. Это – постановка конфлюэнтного анализа.

Во всех трех случаях

$$f_n(\omega,\theta) = \sum_{k=1}^{n} f_k(g(x_k,\theta), y_k),$$

однако случайность входит в правую часть по-разному в зависимости от постановки, от которой зависит и определение предельной функции $f(и)$.

Проще всего выглядит $f(и)$ в случае второй постановки при $f_k \equiv$ f:

$$f(и) = Mf(g(x_1,и),y).$$

В случае первой постановки

$$f(\theta) = \lim_{n\to\infty} \sum_{k=1}^{n} Mf_k(g(x_k,\theta), y_k(\omega))$$

в предположении существования указанного предела. Ситуация усложняется для третьей постановки:

$$f(\theta) = \lim_{n\to\infty} \sum_{k=1}^{n} Mf_k(g(x_k(\omega),\theta), y_k(\omega)).$$

Во всех трех случаях на основе общих результатов о поведении решений экстремальных статистических задач можно изучить [20-22] асимптотику оценок $и_n$. При выполнении соответствующих внутриматематических условий регулярности оценки оказываются состоятельными, т.е. удается восстановить зависимость.

**Аппроксимация и регрессия**. Соотношение (1) дает решение задачи аппроксимации. Поясним, как эта задача соотносится с нахождением регрессии. Согласно [23] для случайной величины (о, з) со значениями в $X$Ч$Y$ регрессией з на о относительно меры близости $f$ естественно назвать решение задачи

$$Mf(g(о), з) \to \min_g, \qquad (2)$$

где $f$: $Y$Ч$Y \to R^1$, $g$: $X \to Y$, минимум берется по множеству всех измеримых функций.

Можно исходить и из другого определения. Для каждого $x \in X$ рассмотрим случайную величину з($x$), распределение которой является условным распределением з при условии о = $x$. В соответствии с определением математического ожидания в пространстве общей природы назовем условным математическим ожиданием решение экстремальной задачи

$$M(\eta \mid \xi = x) = Arg\min\{Mf(y,\eta(x)), y \in Y\}.$$

Оказывается, при обычных предположениях измеримости решение задачи (2) совпадает с $M(\eta \mid \xi = x)$. (Внутриматематические уточнения типа «равенство имеет место почти всюду» здесь опущены.)

Если заранее известно, что условное математическое ожидание $M(\eta \mid \xi = x)$ принадлежит некоторому параметрическому семейству g(x,и), то задача нахождения регрессии сводится к оцениванию параметра и в соответствии с рассмотренной выше второй постановкой вероятностной теории параметрической регрессии. Если же нет оснований считать, что регрессия принадлежит параметрическому семейству, то можно использовать

непараметрические оценки регрессии. Они строятся с помощью непараметрических оценок плотности (см. главу 2.1).

Пусть $н_1$ – мера в $X$, $н_2$ – мера в $У$, а их прямое произведение $н = н_1 Ч н_2$ – мера в $X Ч У$. Пусть $g(x,y)$ – плотность случайного элемента (о,з) по мере н. Тогда условная плотность g(y|x) распределения з при условии о=$x$ имеет вид

$$g(y \mid x) = \frac{g(x,y)}{\int\limits_Y g(x,y)\nu_2(dy)} \qquad (3)$$

(в предположении, что интеграл в знаменателе отличен от 0). Следовательно,

$$Mf(y,\eta(x)) = \int\limits_Y f(y,a)g(a \mid x)\nu_2(da),$$

а потому

$$M(\eta \mid \xi = x) = Arg\min_{y \in Y} Mf(y,\eta(x)) = Arg\min_{y \in Y} \int\limits_Y f(y,a)g(a \mid x)\nu_2(da).$$

Заменяя $g(x,y)$ в (3) непараметрической оценкой плотности $g_n(x,y)$, получаем оценку условной плотности

$$g_n(y \mid x) = \frac{g_n(x,y)}{\int\limits_Y g_n(x,y)\nu_2(dy)}. \qquad (4)$$

Если $g_n(x,y)$ – состоятельная оценка $g(x,y)$, то числитель (4) сходится к числителю (3). Сходимость знаменателя (4) к знаменателю (3) обосновывается с помощью предельной теории статистик интегрального типа (см главу 2.3). В итоге получаем утверждение о состоятельности непараметрической оценки (4) условной плотности (3).

Непараметрическая оценка регрессии ищется как

$$M_n(\eta \mid \xi = x) = Arg\min_{y \in Y} \int\limits_Y f(y,a)g_n(a \mid x)\nu_2(da).$$

Состоятельность этой оценки следует из приведенных выше общих результатов об асимптотическом поведении решений экстремальных статистических задач.

**Применение к методу главных компонент.** Исходные данные – набор векторов $о_1$, $о_2$, … , $о_n$, лежащих в евклидовом пространстве $R^k$ размерности $k$. Цель состоит в снижении размерности, т.е. в уменьшении числа рассматриваемых показателей. Для этого берут всевозможные линейные ортогональные нормированные центрированные комбинации исходных показателей, получают $k$ новых показателей, из них берут первые $m$, где $m < k$ (подробности см. в главе 3.2). Матрицу преобразования $C$ выбирают так, чтобы максимизировать информационный функционал

$$I_n(C) = \frac{s^2(z(1)) + s^2(z(2)) + ... + s^2(z(m))}{s^2(x(1)) + s^2(x(2)) + ... + s^2(x(k))}, \qquad (5)$$

где $x(i)$, $i = 1, 2, \dots, k$, - исходные показатели; исходные данные имеют вид $о_j = (x_j(1), x_j(2), \dots, x_j(k))$, $j = 1, 2, \dots, n$; при этом $z(б)$, $б = 1, 2, \dots, m$, - комбинации исходных показателей, полученные с помощью матрицы $C$. Наконец, $s^2(z(б))$, $б = 1, 2, \dots, m$, $s^2(x(i))$, $i = 1, 2, \dots, k$, - выборочные дисперсии переменных, указанных в скобках.

Укажем подробнее, как новые показатели (главные компоненты) $z(б)$ строятся по исходным показателям $x(i)$ с помощью матрицы $C$:

$$z_j(\alpha) = \sum_{\beta=1}^{k} c_{\alpha\beta}(x_j(\beta) - \overline{x(\beta)}), \ \alpha = 1,2,...,m, \ j = 1,2,...,n,$$

где

$$\overline{x(\beta)} = \frac{1}{n}\sum_{j=1}^{n} x_j(\beta).$$

Матрица $C = ||c_{бв}||$ порядка $m Ч k$ такова, что

$$\sum_{\beta=1}^{k} c_{\alpha\beta}^2 = 1, \ \alpha = 1,2,...,m \qquad (6)$$

(нормированность),

$$\sum_{\beta=1}^{k} c_{\alpha\beta} c_{\gamma\beta} = 0, \ \alpha, \gamma = 1,2,...,m, \ \alpha \neq \gamma \qquad (7)$$

(ортогональность).

Решением основной задачи метода главных компонент является

$$C_n = Arg \min_C(-I_n(C)),$$

где минимизируемая функция определена формулой (5), а минимизация проводится по всем матрицам $C$, удовлетворяющим условиям (6) и (7).

Вычисление матрицы $C_n$ – задача детерминированного анализа данных. Однако, как и в иных случаях, например, для медианы Кемени, возникает вопрос об асимптотическом поведении $C_n$. Является ли решение основной задачи метода главных компонент устойчивым, т.е. существует ли предел $C_n$ при $n \to \infty$? Чему равен этот предел?

Ответ, как обычно, может быть дан только в вероятностной теории. Пусть $o_1$, $o_2$, … , $o_n$ - независимые одинаково распределенные случайные вектора. Положим

$$z_\infty(\alpha) = \sum_{\beta=1}^{k} c_{\alpha\beta}(x_1(\beta) - Mx_1(\beta)), \ \alpha = 1,2,...,m,$$

где матрица $C = ||c_{\text{бв}}||$ удовлетворяет условиям (6) и (7). Введем функцию от матрицы

$$I(C) = \frac{D(z_\infty(1)) + D(z_\infty(2)) + ... + D(z_\infty(m))}{D(x(1)) + D(x(2)) + ... + D(x(k))}.$$

Легко видеть, что при $n \to \infty$ и любом С

$$I_n(C) \to I(C).$$

Рассмотрим решение предельной экстремальной задачи

$$C_\infty = Arg \min_C(-I(C)).$$

Естественно ожидать, что

$$\lim_{n\to\infty} C_n = C_\infty.$$

Действительно, это соотношение вытекает из приведенных выше общих результатов об асимптотическом поведении решений экстремальных статистических задач.

Таким образом, теория, развитая для пространств произвольной природы, позволяет единообразным образом изучать конкретные процедуры прикладной статистики.

### 2.2.4. Робастность статистических процедур

Термин "робастность" (*robustness* - англ.) образован от *robust* - крепкий, грубый (англ.). Сравните с названием одного из сортов кофе - *robusta*. Имеется в виду, что робастные статистические процедуры должны "выдерживать" ошибки, которые теми или иными способами могут попадать в исходные данные или искажать предпосылки используемых вероятностно-статистических моделей.

Термин "робастный" стал популярным в нашей стране в 1970-е годы. Сначала он использовался фактически как сужение термина "устойчивый" на алгоритмы статистического анализа данных классического типа (не включая теорию измерений, статистику нечисловых и интервальных данных). Затем реальная сфера его применения сузилась.

Пусть исходные данные - это выборка, т.е. совокупность независимых одинаково распределенных случайных величин с одной и той же функцией распределения *F(x)*. Наиболее простая модель изучения устойчивости - это модель засорения

$$F(x) = (1-\varepsilon)F_0(x) + \varepsilon H(x). \qquad (1)$$

Эта модель имеются также моделью Тьюки - Хубера. (Джон Тьюки - американский исследователь, П. Хубер, или Хьюбер - швейцарский ученый.) Модель (1) показывает, что с близкой к 1 вероятностью, а именно, с вероятностью $(1-\varepsilon)$, наблюдения берутся из совокупности с функцией распределения $F_0(x)$, которая предполагается обладающей "хорошими" свойствами. Например, она имеет известный статистику вид (хотя бы с точностью до параметров), у нее существуют все моменты, и т.д. Но с малой вероятностью $\varepsilon$ появляются

наблюдения из совокупности с "плохим" распределением, например, взятые из распределения Коши, не имеющего математического ожидания, резко выделяющиеся аномальные наблюдения, выбросы.

Актуальность модели (1) не вызывает сомнений. Наличие засорений (выбросов) может сильно исказить результаты эконометрического анализа данных. Ясно, что если функция распределения элементов выборки имеет вид (1), где первое слагаемое соответствует случайной величине с конечным математическим ожиданием, а второе - такой, для которого математического ожидания не существует (например, если $H(x)$ - функция распределения Коши), то для итоговой функций распределения (1) также не существует математического ожидания. Исследователя обычно интересуют характеристики первого слагаемого, но найти их, т.е. освободиться от влияния засорения, не так-то просто. Например, среднее арифметическое результатов наблюдений не будет иметь никакого предела (это - строгое математическое утверждение, вытекающее из того, что математическое ожидание не существует [24]).

Существуют различные способы борьбы с засорением. Эмпирическое правило "борьбы с засорениями" при подведении итогов работы команды судей найдено в фигурном катании: наибольшая и наименьшая оценки отбрасываются, а по остальным рассчитывается средняя арифметическая. Ясно, что единичное "засорение" окажется среди отброшенных оценок.

Оценивать характеристики и параметры, проверять статистические гипотезы, вообще осуществлять статистический анализ данных все чаще рекомендуют на основе эмпирических квантилей (другими словами, порядковых статистик, членов вариационного ряда), отделенных от концов вариационного ряда. Речь идет об использовании статистик вида

$$ax(0{,}1n) + bx(0{,}3n) + cx(0{,}5n) + dx(0{,}7n) + ex(0{,}9n),$$

где $a, b, c, d, e$ – заданные числа, $x(0{,}1n)$, x$(0{,}3n)$, $x(0{,}5n)$, $x(0{,}7n)$, $x(0{,}9n)$ – члены вариационного ряда с номерами, наиболее близкими к числам, указанным в скобках. Так ценой небольшой потери в эффективности избавляемся от засоренности типа описанной в модели (1).

Вариантом этого подхода является переход к сгруппированным данным. Отрезок прямой, содержащий основную часть наблюдений, разбивается на интервалы, и вместо количественных значений статистик подсчитывает лишь, сколько наблюдений попало в те или иные интервалы. Особое значение приобретают крайние интервалы - к ним относят все наблюдения, которые больше некоторого верхнего порога и меньше некоторого нижнего порога. Любым методам анализа сгруппированных данных резко выделяющиеся наблюдения не страшны.

Можно поставить под сомнение и саму опасность засорения. Дело в том, что практически все реальные величины ограничены. Все они лежат на каком-то интервале - от и до. Это совершенно ясно, если речь идет о физическом измерении - все результаты измерений укладывается в шкалу прибора. По-видимому, и для иных статистических измерений наибольшие сложности создают не сверхбольшие помехи, а те засорения, что находятся "на грани" между "интуитивно возможным" и "интуитивно невозможным".

Что же это означает для практики статистического анализа данных? Если элементы выборки по абсолютной величине не превосходят числа $A$, то все засорение может сдвинуть среднее арифметическое на величину $\varepsilon A$. Если засорение невелико, то и сдвиг мал.

Построена достаточно обширная и развитая теория, посвященная разработке и изучению методов анализа данных в модели (1). С ней можно познакомиться по монографиям [25-27]. К сожалению, в теории обычно предполагается известной степень засорения $\varepsilon$, а на практике эта величина неизвестна. Кроме того, теория обычно направлена на защиту от воздействий, якобы угрожающих из бесконечности (например, отсутствием математического ожидания), а на самом деле реальные данные финитны (сосредоточены на конечных отрезках). Все это объясняет, почему теория робастности, исходящая из модели (1), популярна среди теоретиков, но мало интересна тем, кто анализирует реальные технические, экономические, медицинские и иные статистические данные.

Рассмотрим несколько более сложную модель. Пусть наблюдаются реализации $x_1, x_2, ..., x_n$ независимых случайных величин с функциями распределения $F_1(x), F_2(x), ...., F_n(x)$ соответственно. Эта модель соответствует гипотезе о том, что в процессе наблюдения (измерения) условия несколько менялись. Естественной представляется

модель малых отклонений функций распределений наблюдаемых случайных величин от некоторой "базовой" функции распределения $F_0(x)$. Множество возможных значений функций распределений наблюдаемых случайных величин (т.е. совокупность допустимых отклонений согласно общей схеме устойчивости, рассмотренной в главе 1.4) описывается следующим образом:

$$E((F_1, F_2, ...., F_n); \varepsilon) = \{(F_1, F_2, ...., F_n) : \sup_x | F_i(x) - F_0(x) | < \varepsilon, i = 1,2,..., n].$$

Следующий тип моделей - это введение малой (т.е. слабой) зависимости между рассматриваемыми случайными величинами (см., например, монографию [28]). Ограничения на взаимную зависимость можно задать разными способами. Пусть $F(x_1, x_2, ..., x_n)$ - совместная функция распределения *n*-мерного случайного вектора, $F_1(x_1)$, $F_2(x_2)$, … , $F_n(x_n)$ – функции распределения его координат. Если все координаты независимы, то $F(x_1, x_2, ..., x_n) = F_1(x_1)F_2(x_2)\ldots F_n(x_n)$. Пусть $\rho(i, j)$ – коэффициент корреляции между *i*-ой и *j*-ой случайными величинами – координатами вектора. Множество возможных совместных функций распределения (т.е. совокупность допустимых отклонений согласно общей схеме устойчивости, рассмотренной в главе 1.4) описывается следующим образом:

$$E\big(F(x_1, x_2, ..., x_n); \varepsilon\big) = \{F(x_1, x_2, ..., x_n) : P(x_i(\omega) < x) = F_i(x),\ | \rho(i, j) | \le \varepsilon,\ 1 \le i < j \le n\}.$$

Таким образом, фиксируются функции распределения координат, а коэффициенты корреляции предполагаются малыми (по абсолютной величине).

Есть еще целый ряд постановок задач робастности. Если накладывать погрешности непосредственно на результаты наблюдений (измерений) и предполагать лишь, что эти погрешности не превосходят (по абсолютной величине) заданных величин, то получаем постановки задач статистики интервальных данных (см. главу 3.5). При этом каждый результат наблюдения превращается в интервал - исходное значение плюс-минус максимально возможная погрешность.

Разработано много вариантов робастных методов анализа статистических данных (см. монографии [5, 25-28]). Иногда говорят, что робастные методы позволяют использовать информацию о том, что реальные наблюдения лежат "около" тех или иных параметрических семейств, например, нормальных. В этом, дескать, их преимущество по сравнению с непараметрическими методами, которые предназначены для анализа данных, распределенных согласно произвольной непрерывной функции распределения. Однако количественных подтверждений этих уверений любителей робастных методов обычно не удается найти. В основном потому, что термин «около» трудно формализовать.

На примере различных подходов к изучению робастности статистических процедур оценивания и проверки гипотез видны сложности, связанные с изучением устойчивости. Дело в том, что для каждой конкретной статистической задачи можно самыми разными способами задать совокупность допустимых отклонений. Так, выше кратко рассмотрены четыре такие совокупности, соответствующие модели засорения Тьюки - Хубера, модели малых отклонений функций распределения, модели слабых связей и модели интервальных данных.

В каждой из этих моделей общая схема устойчивости (глава 1.4) предлагает для решения целый спектр задач устойчивости. Кроме изучения свойств робастности известных статистических процедур можно в каждой из постановок находить оптимальные процедуры. Однако практическая ценность этих оптимальных процедур, как правило, невелика, поскольку в других постановках оптимальными будут уже другие процедуры.

## Литература

1. ГОСТ 11.011-83. Прикладная статистика. Правила определения оценок и доверительных границ для параметров гамма-распределения. - М.: Изд-во стандартов, 1984. - 53 с. - Переиздание: М.: Изд-во стандартов, 1985. - 50 с.
2. Рао С.Р. Линейные статистические методы и их применения. – М.: Наука, 1968. - 548 с.
3. Вентцель Е.С. Теория вероятностей. – М.: Наука, 1964. -576 с.
4. Крамер Г. Математические методы статистики. – М.: Мир, 1975. – 648 с.

5. Орлов А.И. Устойчивость в социально-экономических моделях. – М.: Наука, 1979. – 396 с.
6. Лумельский Я.П. К вопросу сравнения несмещенных и других оценок // Прикладная статистика. – М.: Наука, 1983, С.316-319.
7. ГОСТ 11.010-81. Прикладная статистика. Правила определения оценок параметров и доверительных границ для биномиального и отрицательного биномиального распределений. – М.: Изд-во стандартов, 1982. – 32 с.
8. Сатаров Г.А., Шмерлинг Д.С. Новая статистическая модель парных сравнений // Экспертные оценки в задачах управления. – М.: Изд-во Института проблем управления АН СССР, 1982. – С.67-79.
9. Лапига А.Г. Многокритериальные задачи управления качеством: построение прогноза качества в балльной шкале // Заводская лаборатория. 1983. Т.49. № 7. С.55-59.
10. Закс Ш. Теория статистических выводов. – М.: Мир, 1975. – 776 с.
11. Бахмутов В.О., Косарев Л.Н. Использование метода максимального правдоподобия для оценки однородности результатов усталостных испытаний // Заводская лаборатория. 1986. Т.52. № 5. С.52-57.
12. Резникова А.Я., Шмерлинг Д.С. Оценивание параметров вероятностных моделей парных и множественных сравнений // Статистические методы оценивания и проверки гипотез/ Межвузовский сборник научных трудов. – Пермь: Изд-во Пермского госуниверситета, 1984. – С.110-120.
13. Ибрагимов И.А., Хасьминский Р.З. Асимптотическая теория оценивания. – М.: Наука, 1979. – 528 с.
14. Орлов А.И. О нецелесообразности использования итеративных процедур нахождения оценок максимального правдоподобия // Заводская лаборатория. 1986. Т.52. No.5. С.67-69.
15. Боровков А.А. Математическая статистика / Учебное пособие для вузов. – М.: Наука, 1984. – 472 с.
16. Кендалл М.Дж., Стьюарт А. Статистические выводы и связи. – М.: Наука, 1973. – 900 с.
17. Орлов А.И., Миронова Н.Г. Одношаговые оценки для параметров гамма-распределения // Надежность и контроль качества. 1988. No.9. С.18-22.
18. Петрович М.Л., Давидович М.И. Статистическое оценивание и проверка гипотез на ЭВМ. – М.: Финансы и статистика, 1989. – 191 с.
19. Келли Дж. Общая топология. - М.: Наука, 1968. - 384 с.
20. Орлов А.И. Асимптотика решений экстремальных статистических задач. – В сб.: Анализ нечисловых данных в системных исследованиях. Сборник трудов. Вып.10. - М.: Всесоюзный научно-исследовательский институт системных исследований, 1982. - С. 4-12.
21. Орлов А.И. Общий взгляд на статистику объектов нечисловой природы. - В сб.: Анализ нечисловой информации в социологических исследованиях. - М.: Наука, 1985. С.58-92.
22. Орлов А.И. Некоторые неклассические постановки в регрессионном анализе и теории классификации. - В сб.: Программно-алгоритмическое обеспечение анализа данных в медико-биологических исследованиях. - М.: Наука, 1987. с.27-40.
23. Орлов А.И. Статистика объектов нечисловой природы и экспертные оценки. – В сб.: Экспертные оценки / Вопросы кибернетики. Вып.58. - М.: Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1979. С.17-33.
24. Гнеденко Б.В. Курс теории вероятностей: Учебник. 7-е изд., исправл. - М.: Эдиториал УРСС, 2001.- 320 с.
25. Смоляк С.А., Титаренко Б.П. Устойчивые методы оценивания: Статистическая обработка неоднородных совокупностей. - М;: Статистика, 1980. - 208 с.
26. Хьюбер П. Робастность в статистике. - М.: Мир, 1984. - 304 с.
27. Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В. Робастность в статистике. Подход на основе функций влияния. - М.: Мир, 1989. - 512 с.
28. Эльясберг П.Е. Измерительная информация. Сколько ее нужно, как ее обрабатывать? - М.: Наука, 1983. - 208 с.

## Контрольные вопросы и задачи

1. Чем задачи оценивания параметров распределения отличаются от задач оценивания характеристик распределения?
2. С помощью метода линеаризации обоснуйте вид асимптотических дисперсий оценок метода моментов для параметров гамма-распределения (табл.3 подраздела 2.2.1).
3. Почему одношаговые оценки предпочтительнее оценок максимального правдоподобия?
4. Как связаны законы больших чисел в пространствах произвольной природы и утверждения об асимптотическом поведении решений экстремальных статистических задач?
5. Как соотносятся параметрическая регрессия и непараметрическая регрессия?
6. Сопоставьте различные постановки задач изучения робастности статистических процедур.

**Темы докладов, рефератов, исследовательских работ**

1. Квантильные оценки.
2. Минимизация расстояния как способ построения оценок параметров.
3. Одношаговые оценки параметров распределения Вейбулла-Гнеденко.
4. Оптимизационные постановки основных задач прикладной статистики.
5. Роль функции влияния при изучении робастности в модели засорения Тьюки – Хубера.
6. На основе четырех указанных в настоящем учебнике моделей сформулируйте новые постановки задач устойчивости статистических процедур.

## 2.3. Проверка гипотез

### 2.3.1. Метод моментов проверки гипотез

К методу моментов относят все статистические процедуры, основанные на использовании выборочных моментов и функций от них. Метод моментов оценивания параметров распределения рассмотрен в главе 2.2. В непараметрической статистике на основе выборочных моментов проводится точечное и интервальное оценивание характеристик распределения, таких, как математическое ожидание, дисперсия, среднее квадратическое отклонение, коэффициент вариации (глава 3.1). Для проверки гипотез в непараметрической статистике также используется метод моментов. Примером является критерий Крамера-Уэлча, предназначенный для проверки равенства математических ожиданий по двум независимым выборкам (глава 3.1).

В практике применения статистических методов (согласно классическим схемам) довольно часто возникает необходимость проверки гипотезы о том, что функция распределения результатов наблюдений $X_1$, $X_2$, … , $X_n$ принадлежит параметрическому семейству распределений $\{F(x, и), и \in И\}$, где $И \subseteq R^k$. Как проверять эту гипотезу?

Давно разработан универсальный метод – критерий минимума хи-квадрат [1]. Однако у него имеется существенный недостаток – необходимость группирования наблюдений, что приводит к потере информации. Как хорошо известно [2], это приводит к существенному снижению мощности критерия минимума хи-квадрат по сравнению с критериями типа Колмогорова и типа омега-квадрат. Кроме того, нахождение минимума статистики хи-квадрат – достаточно сложная вычислительная процедура. Поэтому иногда вместо оценок, получаемых при указанной оптимизации, подставляют оценки максимального правдоподобия или какие-либо еще. Такая замена приводит к тому, что распределение рассматриваемой статистики существенно отличается от классического, причем различие не исчезает при росте объема выборки. Предложенная член-корр. АН СССР Л.Н. Большевым и проф. М.С. Никулиным [3] модификация критерия минимума хи-квадрат не снимает недостатков, связанных с группированием и необходимостью существенной вычислительной работы.

Общий подход, основанный на дистанционном методе, предложен Дж. Вольфовицем (США) в 1950-х годах. Согласно этому методу следует основываться на том или ином расстоянии между эмпирической функцией распределения и параметрическим семейством распределений (как многообразием в пространстве всех функций распределения). Конкретная реализация этого подхода приводит к критериям типа Колмогорова и типа омега-квадрат. Однако для каждого конкретного параметрического семейства приходится разрабатывать самостоятельную теорию и рассчитывать только ему соответствующие предельные и точные распределения [4, 5]. Предельные распределения найдены лишь для нескольких семейств, а точных почти ничего не известно. До сих пор часто делают ошибку, применяя для произвольных семейств предельные распределения, найденные для проверки согласия с фиксированным распределением (см. подробности в главе 1.2).

Отметим, что критерии минимума хи-квадрат и аналогичные им не являются состоятельными, поскольку вероятности попадания в области группирования не задают однозначно функцию распределения. С этим недостатком можно бороться, увеличивая число интервалов группирования вместе с ростом объема выборки, однако на этом пути еще не выработаны рекомендации, пригодные для широкого практического использования. Критерии типа Колмогорова и типа омега-квадрат – состоятельные, т.е. любую альтернативную функцию распределения, не входящую в рассматриваемое параметрическое семейство, они отвергают с вероятностью, стремящейся к 1 при росте объема выборки.

Для конкретности обсудим проверку согласие результатов наблюдений с трехпараметрическим семейством гамма-распределений с плотностями

$$f(x;a,b,c)=\begin{cases}\dfrac{1}{\Gamma(a)}(x-c)^{a-1}b^{-a}\exp\left[-\dfrac{x-c}{b}\right], & x\ge c,\\ 0, & x<c.\end{cases} \qquad (1)$$

Здесь $a>2$ - параметр формы, $b>0$ - параметр масштаба и $c$ - параметр сдвига, $\Gamma(a)$ - одна из используемых в математике специальных функций, так называемая "гамма-функция". Критерий минимума хи-квадрат имеет указанные выше недостатки. Критерии типа Колмогорова и типа омега-квадрат для этого случая не разработаны.

В подобных ситуациях целесообразно строить критерии согласия на основе функций от выборочных моментов, т.е. пользоваться методом моментов. Для оценивания параметров метод моментов хорошо известен и обычно рассматривается в учебной литературе по теории вероятностей и математической статистике. Реализацией метода моментов для проверки нормальности являются известные критерии асимметрии и эксцесса [6].

*Пример 1.* Если случайная величина $X$ имеет нормальное распределение с математическим ожиданием $a$ и дисперсией $\sigma^2$, то, как известно [6],

$$\delta=\frac{M\,|X-a|}{\sigma}=\sqrt{\frac{2}{\pi}}=0{,}79788,\ \gamma_1=\frac{M(X-a)^3}{\sigma^3}=0,\ \beta_1=\frac{M(X-a)^4}{\sigma^4}=3,$$

где д – нормированное среднее абсолютное отклонение, $г_1$ – коэффициент асимметрии и ($в_1$ – 3) – коэффициент эксцесса. Таким образом, если выборочные оценки указанных моментных отношений существенно отличаются от соответствующих теоретических значений, то следует признать, что распределение результатов наблюдений отлично от нормального. Так как указанные выше значения моментных отношений могут приниматься и для распределений, отличных от нормальных, то близость выборочных значений к только что выписанным не обязательно свидетельствует о нормальности распределения результатов наблюдений. Критерии, полученные методом моментов, служат не столько для проверки нормальности, сколько для выявления отклонений распределения от нормального, или, точнее, для проверки гипотез $д \ne \sqrt{2/\pi}$, $г_1 \ne 0$, $в_1 \ne 3$. Рассматриваемые критерии построены на основе выборочных моментных отношений:

$$d=\frac{1}{ns}\sum_{k=1}^{n}|X_k-\overline{X}|,\ g_1=\frac{1}{ns^3}\sum_{k=1}^{n}(X_k-\overline{X})^3,\ b_1=\frac{1}{ns^4}\sum_{k=1}^{n}(X_k-\overline{X})^4.$$

Здесь, как обычно, $\overline{X}$ - выборочное среднее арифметическое и $s^2$ – выборочная дисперсия, соответственно, $s$ – выборочное среднее квадратическое отклонение. Как вытекает из результатов главы 1.4, все три статистики являются асимптотически нормальными. Выражения для параметров их асимптотических распределений приведены в [6]. Процентные точки распределений рассматриваемых выборочных моментных отношений при конечных объемах выборки найдены в предположении нормальности результатов наблюдений [6].

Как и критерии минимума хи-квадрат, критерии метода моментов никогда не являются состоятельными. Однако они, как и в случае критериев асимметрии и эксцесса, позволяют в ряде случаев отвергнуть гипотезу согласия. Использование несостоятельных критериев часто встречается в прикладной статистике. Отметим, например, что применение критерия Вилкоксона для проверки гипотезы однородности двух выборок широко распространено, хотя против общей альтернативы он является несостоятельным (см. главу 3.1).

Критерии метода моментов основаны на использовании функций от выборочных моментов, имеющих асимптотически нормальные распределения, параметры которых легко могут быть вычислены по методике, описанной в главе 1.4. Метод моментов по сравнению с другими методами проверки согласия требует существенно меньше вычислений (число операций пропорционально объему выборки). Поэтому он может быть рекомендован для использования при проверке согласия с семействами распределений, для которых не разработаны более совершенные методы, а также в качестве быстрого (экспрессного) метода. Что же касается хорошо изученных семейств, например, нормального, то основанные на использовании моментов критерии

асимметрии и эксцесса применять для проверки нормальности нецелесообразно. Судя по специальным исследованиям, следует рекомендовать критерий *W* Шапиро - Уилка.

Продемонстрируем применение метода моментов на примере проверки гипотезы согласия с двухпараметрическим семейством гамма-распределений без сдвига, т.е. выделяемого из семейства (1) условием *c*=0. Поскольку для трехпараметрического семейства гамма-распределений (1)

$$M(X) = ab + c,\ D(X) = ab^2,\ м_3 = M(X - M(X))^3 = 2ab^3,$$

то при справедливости гипотезы $H_0$: $c = 0$ выполнено соотношение

$$\frac{M(X)\mu_3}{2\sigma^4} - 1 = 0. \qquad (2)$$

Для специалистов по техническим наукам большое значение имеет альтернативная гипотеза

$$H_1: c > 0.$$

В частности, она связана с дискуссией о выборе нормируемых показателей надежности технических устройств. Альтернативная гипотеза соответствует предположению, что в течение некоторого времени (до момента $c > 0$) отказы невозможны, а нулевая – с отрицанием этого предположения и признанием того, что отказы возможны в любой момент.

При справедливости альтернативной гипотезы

$$\frac{M(X)\mu_3}{2\sigma^4} - 1 = \frac{c}{ab} > 0,$$

поэтому для проверки гипотезы согласия в рассматриваемой постановке целесообразно использовать критерий со статистикой

$$Z = \frac{\overline{X} m_3}{2s^4} - 1.$$

С помощью описанной в главе 1.4 методики вычисления предельного распределения функции от выборочных моментов можно установить, что при $n \to \infty$ распределение статистики $\sqrt{n}Z$ сходится к нормальному, причем при справедливости нулевой гипотезы, т.е. соотношения (2), асимптотическое распределение имеет нулевое математическое ожидание и дисперсию

$$\frac{1}{2a}(3a^2 + 13a + 10). \qquad (3)$$

Поскольку параметр формы *a* неизвестен статистику, необходимо в выражении (3) заменить а на его состоятельную оценку, например, на оценку метода моментов (см. главу 2.2)

$$a^* = \frac{\left(\overline{X}\right)^2}{s^2}.$$

Рассмотрим критерий с критической областью вида

$$\left\{ Z:\ Z > u(1-\alpha)\frac{3(a^*)^2 + 13a^* + 10}{2a^*\sqrt{n}} \right\}, \qquad (4)$$

где $u(1 - б)$ – квантиль порядка 1 - б стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. При $n \to \infty$ уровень значимости этого критерия стремится к б.

Если альтернативная гипотеза является двусторонней, т.е. $H_1^{'}: c \neq 0$, то аналогично строится двусторонняя критическая область.

Критерий (4) состоятелен против альтернативы $H_1$: $c > 0$, а также против непараметрической альтернативы

$$\frac{M(X)\mu_3}{2\sigma^4} > 1,$$

в которой не предполагается, что функция распределения элементов выборки имеет гамма-распределение (1) с какими-либо конкретными значениями параметров, но не является состоятельным против общей альтернативы.

*Пример 2.* Применим критерий (4) для проверки согласия с гамма-распределением при $c = 0$, т.е. с двухпараметрическим семейством, данных о наработке $n = 50$ резцов до предельного состояния (в часах), приведенных в табл.2 подраздела 2.2.1.

Для рассматриваемых данных $\overline{X} = 57{,}88$, $s^2 = 663{,}00$, выборочный третий центральный момент $m_3 = 14927{,}91$, откуда $Z = - 0{,}01719$. При этом $a^* = 5{,}05$, и потому

$$\frac{3(a^*)^2 + 13a^* + 10}{2a^*\sqrt{n}} = 0{,}4488 .$$

Следовательно, гипотеза согласия рассматриваемых данных с двухпараметрическим гамма-распределением не отвергается на любом из обычно используемых уровней значимости, как для односторонней критической области, так и для двухсторонней.

### 2.3.2. Неустойчивость параметрических методов отбраковки выбросов

При обработке реальных технических, экономических, медицинских и иных данных, полученных в процессе наблюдений, измерений, расчетов, иногда один или несколько результатов наблюдений резко выделяются, т.е. далеко отстоят от основной массы данных. Такие резко выделяющиеся результаты наблюдений часто считают содержащими грубые погрешности, соответственно называют промахами или выбросами. В рассматриваемых случаях возникает естественная мысль о том, что подобные наблюдения не относятся к изучаемой совокупности, поскольку содержат грубую погрешность, а получены они в результате ошибки, промаха. В справочнике по метрологии об этом явлении говорится так: "Грубые погрешности и промахи возникают из-за ошибок или неправильных действий оператора (его психофизиологического состояния, неверного отсчета, ошибок в записях или вычислениях, неправильного включения приборов и т.п.). А также при резких кратковременных изменениях условий проведения измерений (в результате вибрации, поступления холодного воздуха, толчка прибора оператором и т.п.). Если грубые погрешности и промахи обнаруживают в процессе измерений, то результаты, содержащие их, отбрасывают. Однако чаще всего их выявляют только при окончательной обработке результатов измерений с помощью специальных критериев оценки грубых погрешностей" [7, с.46-47].

Есть два подхода к обработке данных, которые могут быть искажены грубыми погрешностями и промахами:

1) отбраковка резко выделяющихся результатов наблюдений, т.е. обнаружение наблюдений, искаженных грубыми погрешностями и промахами, и исключение их из дальнейшей статистической обработки;

2) применение устойчивых (робастных) методов обработки данных, на результаты работы которых мало влияет наличие небольшого числа грубо искаженных наблюдений (см. подраздел 2.2.4).

Обсудим методы отбраковки. Наиболее изучена ситуация, когда результаты наблюдений - числа $x_1, x_2, \ldots, x_n$, среди них резко выделяется один результат наблюдения, для определенности, максимальный $x_{\max}$.

Простейшая вероятностно-статистическая модель такова [6]. При нулевой гипотезе $H_0$ результаты наблюдения $x_1, x_2, \ldots, x_n$ рассматриваются как реализация независимых одинаково распределенных случайных величин числа $X_1, X_2, \ldots, X_n$ с функцией распределения $F(x)$. При альтернативной гипотезе $H_1$ случайные величины $X_1, X_2, \ldots, X_n$ также независимы, $X_1, X_2, \ldots, X_{n-1}$ имеют распределение $F(x)$, а $X_n$ - распределение $G(x)$, оно "существенно сдвинуто вправо" относительно $F(x)$, например, $G(x)=F(x - A)$, где $A$ достаточно велико. Если альтернативная гипотеза справедлива, то при $A \to \infty$ вероятность равенства

$$X_n = \max(X_1, X_2, \ldots, X_n)$$

стремится к 1, поэтому естественно применять решающее правило следующего вида:

если $x_{max.} > d$, то принять $H_1$,

если $x_{max.} \le d$, то принять $H_0$, (1)

где $d$ - параметр решающего правила, который следует определять из вероятностно-статистических соображений.

При справедливости нулевой гипотезы

$$P\{\max_{1\le i\le n} X_i \le d\} = \{F(d)\}^n.$$

Статистический критерий проверки гипотезы $H_0$, основанный на решающем правиле вида (1), имеет уровень значимости $\alpha$, если

$$P\{\max_{1\le i\le n} X_i > d\} = 1 - \{F(d)\}^n = \alpha,$$

т.е.

$$F(d) = \sqrt[n]{1-\alpha}. \qquad (2)$$

Из соотношения (2) определяют граничное значение $d = d(\alpha, n)$ в решающем правиле (1).

При больших $n$ и малых $\alpha$ согласно известным результатам математического анализа

$$F(d) = \sqrt[n]{1-\alpha} = 1 - \frac{\alpha}{n} + O\left(\frac{\alpha^2}{n^2}\right), \qquad (3)$$

поэтому в качестве хорошего приближения к $d(\alpha, n)$ рассматривают $(1-\alpha/n)$ - квантиль распределения $F(x)$.

Пусть правило отбраковки задано в соответствии с соотношениями (1) и (2) с некоторой функцией распределения $F$, однако выборка берется из функции распределения $G$, мало отличающейся от $F$ в смысле расстояния Колмогорова:

$$\rho(F, G) = \sup_x |F(x) - G(x)| \le \delta. \qquad (4)$$

С помощью соотношения (3) получаем, что величина $\gamma = G(d)$ для $d$ из уравнения (2) находится между $\gamma_1 = \max(0,\ 1 - \frac{\alpha}{n} - \delta)$ и $\gamma_2 = \min(1 - \frac{\alpha}{n} + \delta,\ 1)$. Таким образом, уровень значимости критерия, построенного для $F$, при применении к наблюдениям из $G$ есть $1-\gamma^n$ и может принимать любые значения в отрезке $[1-\gamma_2;\ 1-\gamma_1]$.

В частности, при $\delta = 0{,}01$, $\alpha = 0{,}05$, $n = 5$ возможные значения уровня значимости заполняют отрезок [0; 0,1], т.е. уровень значимости может быть в 2 раза выше номинального. А если $n$ возрастает до 30, то максимальный уровень значимости есть 0,297, т.е. почти в 6 раз выше номинального. При дальнейшем росте $n$ верхняя граница для уровня значимости, как нетрудно видеть, приближается к 1.

Рассмотрим и другой вопрос - насколько правило отбраковки с уровнем значимости $\alpha$ для $G$ может отличаться от такового для $F$ при справедливости неравенства (4). С использованием соотношения (3) заключаем, что из

$$G(d) = 1 - \frac{\alpha}{n} \qquad (5)$$

следует, что $\gamma_1 \le F(d) \le \gamma_2$, где $\gamma_1$ и $\gamma_2$ выписаны выше. Решение уравнения (5) может принимать любое значение в отрезке $[F^{-1}(\gamma_1);\ F^{-1}(\gamma_2)]$. В частности, при $\alpha = 0{,}05$ и $n = 5$ для стандартного нормального распределения $F$ имеем $d(\alpha, n) = 2{,}319$, при $\delta = 0{,}01$ решение уравнения (5) может принимать любое значение в отрезке $[2{,}054; +\infty]$, при $\delta = 0{,}005$ - любое значение в [2,170; 2,576].

При использовании любого другого расстояния между функциями распределения выводы о неустойчивости правил отбраковки также справедливы. Отметим, что проведенные рассмотрения выполнены в рамках "общей схемы устойчивости" (см. главу 1.4).

Рассмотренные примеры показывают, что при конкретном значении $\delta = 0{,}01$ в неравенстве (4) весьма неустойчивы как уровни значимости при фиксированном правиле отбраковки, так и параметр $d$ правила отбраковки при фиксированном уровне значимости. Обсудим, насколько реалистично определение функции распределения с точностью $\delta \le 0{,}01$.

Есть два подхода к определению функции распределения результатов наблюдений: эвристический подбор с последующей проверкой с помощью критериев согласия и вывод из некоторой вероятностной модели.

Пусть с помощью критерия согласия Колмогорова проверяется гипотеза о том, что выборка взята из распределения $F$. Пусть функции распределения $F$ и $G$ удовлетворяют соотношению (4). Пусть на самом деле выборка взята из распределения $G$, а не $F$. При каких $\delta$ не удастся различить $F$ и $G$? Для определенности, при каких $\delta$ гипотеза согласия с $F$ будет приниматься не менее чем в 50% случаев?

Критерий согласия Колмогорова основан на статистике

$$\lambda_n = \sqrt{n}\,\rho(F_n, H), \qquad (6)$$

где расстояние $\rho$ между функциями распределения определено выше в формуле (4); $H$ - та функция распределения, согласие с которой проверяется, а $F_n$ - эмпирическая функция распределения (т.е. $F_n(x)$ равно доле наблюдений, меньших $x$, в выборке объема $n$). Как показал А.Н. Колмогоров в 1933 г., функция распределения случайной величины $\lambda_n$ при росте объема выборки $n$ сходится к некоторой функции распределения $K(x)$, которую ныне называют функцией Колмогорова. При этом $K(1{,}36) = 0{,}95$ и $K(0{,}83) = 0{,}50$.

Поскольку выборка взята из распределения $G$, то с вероятностью 0,50

$$\rho(F_n, G) < 0{,}83 / \sqrt{n} \qquad (7)$$

(при больших $n$). Тогда для рассматриваемой выборки с учетом неравенства (4) и неравенства треугольника для расстояния Колмогорова и симметричности этого расстояния имеем

$$\rho(F_n, F) \le \rho(F_n, G) + \rho(G\ , F) = \rho(F_n, G) + \rho(F\ , G) < 0{,}83 / \sqrt{n} + \delta.$$

Если

$$0{,}83 / \sqrt{n} + \delta \le 1{,}36 / \sqrt{n},$$

т.е.

$$\delta\sqrt{n} \le 0{,}53, \qquad (8)$$

то, согласно формуле (6), гипотеза согласия принимается (на уровне значимости 0,95) по крайней мере с той же вероятностью, с которой выполнено неравенство (7), т.е. с вероятностью не менее 0,50. Для $\delta = 0{,}01$ это условие выполняется при $n \le 2809$. Таким образом, для определения функции распределения с точностью $\delta \le 0{,}01$ с помощью критерия согласия Колмогорова необходимо несколько тысяч наблюдений, что для большинства задач прикладной статистики нереально.

При втором из названных выше подходов к определению функции распределения ее конкретный вид выводится из некоторой системы аксиом, в частности, из некоторой модели порождения соответствующей случайной величины. Например, из модели суммирования вытекает нормальное распределение. А из мультипликативной модели (т.е. модели перемножения) - логарифмически нормальное распределение. Как правило, при выводе используется предельный переход. Так, из Центральной Предельной Теоремы теории вероятностей вытекает, что сумма независимых случайных величин может быть приближена нормальным распределением. Однако более детальный анализ, в частности, с помощью неравенства Берри-Эссеена (см. подраздел 2.1.1) показывает, что для гарантированного достижения точности $\delta \le 0{,}01$ необходимо более полутора тысяч слагаемых. Такого количества слагаемых реально, конечно, указать почти никогда нельзя. Это означает, что при решении практических статистических задач теория дает возможность лишь сформулировать гипотезу о виде функции распределения, а проверять ее надо с помощью анализа

реальной выборки объема, как показано выше, не менее нескольких тысяч.

Таким образом, в большинстве реальных ситуаций определить функцию распределения с точностью $\delta \le 0{,}01$ невозможно.

Итак, показано, что правила отбраковки, основанные на использовании конкретной функции распределения, являются крайне неустойчивыми к отклонениям от нее распределения элементов выборки, а гарантировать отсутствие подобных отклонений почти всегда невозможно. Поэтому отбраковка по классическим правилам математической статистики [6] не является научно обоснованной, особенно при больших объемах выборок. Указанные правила целесообразно применять лишь для выявления "подозрительных" наблюдений, вопрос об отбраковке которых должен решаться из соображений соответствующей предметной области, а не из формально-математических соображений.

Выше для простоты изложения рассмотрен лишь случай полностью известного распределения $F$, для которого изучено правило отбраковки, заданное формулами (1) и (2). Аналогичные выводы о крайней неустойчивости правил отбраковки справедливы, если "истинное распределение" принадлежит какому-либо параметрическому семейству, например, нормальному, Вейбулла-Гнеденко, гамма.

Параметрическим методам отбраковки, основанным на моделях тех или иных параметрических семейств распределений, посвящены тысячи книг и статей. Приходится признать, что они имеют в основном внутриматематический интерес. При обработке реальных данных следует применять устойчивые методы (см. подразделы 1.4.7 и 2.2.4). Прежде всего можно рекомендовать непараметрические методы, а среди них – ранговые (т.е. инвариантные в порядковой шкале).

### 2.3.3. Предельная теория непараметрических критериев

В прикладной статистике широко используются статистики типа омега-квадрат и типа Колмогорова-Смирнова. Они применяются для проверки согласия с фиксированным распределением или семейством распределений, для проверки однородности двух выборок, симметрии распределения относительно 0, при оценивании условной плотности и регрессии в пространствах произвольной природы и т.д.

**Статистики интегрального типа и их асимптотика.** Рассмотрим статистики интегрального типа

$$\xi_\alpha = \xi(f_\alpha, F_\alpha) = \int_X f_\alpha(x,\omega)\, dF_\alpha(x,\omega), \qquad (1)$$

где $X$ – некоторое пространство, по которому происходит интегрирование (например, $X = [0; 1]$, $X = R^1$ или $X = R^k$). Здесь $\{б\}$ – направленное множество, переход к пределу по которому обозначен как $б \to \infty$ (см. главу 1.4). Случайные функции $f_б$: $X$ЧЩ $\to Y$ обычно принимают значения, являющиеся числами. Но иногда рассматривают и постановки, в которых $У = R^k$ или $У$ – банахово пространство (т.е. полное нормированное пространство [8]). Наконец, $F_б(x,щ)$ – случайная функция распределения или случайная вероятностная мера; в последнем случае используют также обозначение $dF_б(x,щ) = F_б(dx,щ)$.

Предполагаются выполненными необходимые для корректности внутриматематические предположения измеримости, например, сформулированные в [9, 10].

*Пример 1.* Рассмотрим критерий Лемана – Розенблатта, т.е. критерий типа омега-квадрат для проверки однородности двух независимых выборок (см. главу 3.1). Его статистика имеет вид:

$$A = \frac{mn}{m+n} \int_{-\infty}^{\infty} (F_m(x) - G_n(x))^2 dH_{m+n}(x),$$

где $F_m(x)$ – эмпирическая функция распределения, построенная по первой выборке объема $m$, $G_n(x))$ - эмпирическая функция распределения, построенная по второй выборке объема $n$, а $H_{m+n}(x)$ –

эмпирическая функция распределения, построенная по объединенной выборке объема $m+n$. Легко видеть, что

$$H_{m+n}(x) = \frac{m}{m+n} F_m(x) + \frac{n}{m+n} G_n(x).$$

Ясно, что статистика $A$ имеет вид (1). При этом $x$ – действительное число, $X = Y = R^1$, в роли б выступает пара $(m, n)$, и б→∞ означает, что $\min(m, n) \to \infty$. Далее,

$$f_б(x,щ) = \frac{mn}{m+n}\left(F_m(x) - G_n(x)\right)^2.$$

Наконец, $F_б(x,щ) = H_{m+n}(x)$.

Теперь обсудим асимптотическое поведение функций $f_б(x,щ)$ и $F_б(x,щ)$, с помощью которых определяется статистика $A$. Ограничимся случаем, когда справедлива гипотеза однородности, функции распределения, соответствующие генеральным совокупностям, из которых взяты выборки, совпадают. Их общую функцию распределения обозначим $F(x)$. Она предполагается непрерывной. Введем в рассмотрение выборочные процессы

$$\xi_m(x) = \sqrt{m}(F_m(x) - F(x)),\ \ \eta_n(x) = \sqrt{n}(G_n(x) - F(x)).$$

Нетрудно проверить, что

$$f_\alpha(x,\omega) = \left(\sqrt{\frac{n}{m+n}}\xi_m(x) - \sqrt{\frac{m}{m+n}}\eta_n(x)\right)^2.$$

Сделаем замену переменной $t = F(x)$. Тогда выборочные процессы переходят в соответствующие эмпирические (см. главу 1.4):

$$f_\alpha(F^{-1}(t),\omega) = \left(\sqrt{\frac{n}{m+n}}\xi_m(t) - \sqrt{\frac{m}{m+n}}\eta_n(t)\right)^2,\ \ 0 \le t \le 1.$$

Конечномерные распределения этого процесса, т.е. распределения случайных векторов

$$(f_\alpha(F^{-1}(t_1),\omega), f_\alpha(F^{-1}(t_2),\omega), ..., f_\alpha(F^{-1}(t_k),\omega))$$

для всех возможных наборов $(t_1, t_2, \ldots, t_k)$, сходятся к конечномерным распределениям квадрата броуновского моста $o^2(t)$. В соответствии с подразделом 1.4.5 рассматриваемая сходимость по распределению обозначается так:

$$f_\alpha(F^{-1}(t_1),\omega) \Rightarrow \xi^2(t),\ \ 0 \le t \le 1. \qquad (2)$$

Нетрудно видеть, что

$$F_б(x,щ) = H_{m+n}(x) \to F(x)$$

при б→∞. С помощью замены переменной $t = F(x)$ получаем, что

$$F_б(F^{-1}(t),щ) = H_{m+n}(F^{-1}(t)) \to t \qquad (3)$$

при б→∞. Из соотношений (2) и (3) хотелось бы сделать вывод, что в случаен статистики Лемана - Розенблатта типа омега-квадрат

$$\xi_\alpha = \int_X f_\alpha(x,\omega) dF_\alpha(x,\omega) = A \Rightarrow \int_0^1 \xi^2(t) dt,$$

т.е. предельным распределением этой статистики является классическое распределение [6], найденное как предельное для одновыборочной статистики критерия согласия омега-квадрат Крамера-Мизеса-Смирнова.

Действительно, сформулированное утверждение справедливо. Однако доказательство нетривиально.

Так, может показаться очевидным следующее утверждение.

*Утверждение 1*. Пусть $f$: [0; 1] → $R^1$ – ограниченная функция, $G_n(x)$ и $G(x)$ – функции распределения, $G_n(0) = G(0) = 0$, $G_n(1) = G(1) = 1$, причем $G_n(x) \to G(x)$ при всех $x$. Тогда

$$\lim_{n\to\infty}\int_0^1 f(x)d(G_n(x)-G(x))=0. \qquad (4)$$

Это утверждение неверно (ср. [5, с.42]). Действительно, пусть $f(x) = 1$, если $x$ рационально, и $f(x) = 0$, если $x$ иррационально, $G(x) = x$, а $G_n(x)$ имеет скачки величиной $2^{-n}$ в точках $m/2^n$, $m = 1, 2, \dots, 2^n$ при всех $n = 1, 2, \dots$ Тогда $G_n(x) \to G(x)$ при всех $x$, однако

$$\int_0^1 f(x)dG_n(x)=1,\ \int_0^1 f(x)dG(x)=0$$

при всех $n = 1, 2, \dots$ Следовательно, вопреки сформулированному выше утверждению 1,

$$\int_0^1 f(x)d(G_n(x)-G(x))=1,$$

т.е. соотношение (4) неверно.

Итак, сформулируем проблему. Пусть известно, что последовательность случайных функций $f_б(x, щ)$ сходится по распределению при б→∞ к случайной функции $f(x,щ)$. Пусть последовательность случайных мер $F_б(A,щ)$ сходится по распределению к вероятностной мере $F(A)$ при б→∞. Если речь идет о конечномерном пространстве и меры задаются функциями распределения, то сходимость $F_б(x,щ)$ к $F(x)$ должна иметь место во всех точках непрерывности $F(x)$. В каких случаях можно утверждать, что при б→∞ справедлив предельный переход

$$\xi_\alpha=\xi(f_\alpha,F_\alpha)=\int_X f_\alpha(x,\omega)dF_\alpha(x,\omega)\Rightarrow\xi=\xi(f,F)=\int_X f(x,\omega)dF(x)\ ?$$

Выше показано, что, например, ограниченности $f_б(x, щ)$ для этого недостаточно.

**Метод аппроксимации ступенчатыми функциями.** Пусть $T = \{C_1, C_2, \dots, C_k\}$ – разбиение пространства $X$ на непересекающиеся подмножества. Пусть в каждом элементе $C_j$ разбиения $T$ выделена точка $x_j$, $j = 1, 2, \dots, k$. На множестве функций $f$: $X \to Y$ введем оператор $A_T$: если $x \in C_j$, то

$$A_T f(x) = f(x_j),\ j = 1, 2, \dots . k. \qquad (5)$$

Тогда $A_T f$ – аппроксимация функции $f$ ступенчатыми (кусочно-постоянными) функциями.

Пусть $f_б(x,щ)$ – последовательность случайных функций на $X$, а $K(\cdot)$ – функционал на множестве всех возможных их траекторий как функций от $x$. Для изучения распределения $K(f_б)$ методом аппроксимации ступенчатыми функциями используют разложение

$$K(f_б) = K(A_T f_б) + \{K(f_б) - K(A_T f_б)\}. \qquad (6)$$

Согласно (5) распределение первого слагаемого в (6) определяется конечномерным распределением случайного элемента, а именно, распределением вектора

$$(f_б(x_1,щ), f_б(x_2,щ), \dots\ f_б(x_k,щ)). \qquad (7)$$

В обычных постановках предельной теории непараметрических критериев распределение вектора (7) сходится при б→∞ к соответствующему конечномерному распределению предельной случайной функции $f(x,щ)$, т.е. к распределению случайного вектора

$$(f(x_1,щ), f(x_2,щ), \dots\ f(x_k,щ)). \qquad (8)$$

В соответствии с теорией наследования сходимости (глава 1.4) при слабых условиях на функционал $K(\cdot)$ из сходимости по распределению вектора (7) к вектору (8) следует сходимость по распределению $K(A_T f_б)$ к $K(A_T f)$.

Используя аналогичное (6) разложение

$$K(f) = K(A_T f) + \{K(f) - K(A_T f)\}, \qquad (9)$$

можно устанавливать сходимость по распределению $K(f_б)$ к $K(f)$ при б→∞ в два этапа: сначала выбрать разбиение $T$ так, чтобы вторые слагаемые в правых частях соотношений (6) и (9) были малы, а затем при фиксированном операторе $A_T$ воспользоваться сходимостью по распределению $K(A_T f_б)$ к $K(A_T f)$.

Рассмотрим простой пример применения метода аппроксимации ступенчатыми функциями.

**Обобщение теоремы Хелли.** Пусть $f$: [0; 1] → $R^1$ – измеримая функция, $F_n(x)$ – функции распределений, сосредоточенных на отрезке [0; 1]. Пусть $F_n(x)$ сходятся в основном к функции распределения $F(x)$, т.е.

$$\lim_{n\to\infty} F_n(x) = F(x) \qquad (10)$$

для всех $x$, являющихся точками непрерывности $F(x)$.

*Утверждение 2.* Если $f(x)$ – непрерывная функция, то

$$\lim_{n\to\infty} \int_0^1 f(x)dF_n(x) = \int_0^1 f(x)dF(x) \qquad (11)$$

(рассматриваются интегралы Лебега-Стилтьеса).

Утверждение 2 известно в литературе как первая теорема Хелли [8, с.344-346], вторая теорема Хелли [11, с.174-175], лемма Хелли-Брея [12, с.193-194].

Естественно поставить вопрос: при каких $f$ из (10) следует (11)? Необходимо ввести условия и на $F_n$: если $F_n \equiv F$, то соотношение (11) верно для любой измеримой функции $f$, для которой интеграл в (11) существует. Поэтому рассмотрим следующую постановку.

*Постановка 1*. Пусть функция $f$ такова, что для *любой* последовательности $F_n$, удовлетворяющей (10), справедливо (11). Что можно сказать о функции $f$?

В работах [9, 10] найдены следующие необходимые и достаточные условия на функцию $f$.

*Теорема 1*. Пусть ограниченная на [0; 1] функция $f$ интегрируема по Риману-Стилтьесу по функции распределения $F(x)$. Тогда для *любой* последовательности функций распределения $F_n$, сходящейся в основном к $F$, имеет место предельный переход (11).

*Теорема 2*. Пусть функция $f$ не интегрируема по Риману-Стилтьесу по функции распределения $F(x)$. Тогда *существует* последовательность функций распределения $F_n$, сходящаяся в основном к $F$, для которой соотношение (11) не выполнено.

Теоремы 1 и 2 в совокупности дают необходимые и достаточные условия для $f$ в постановке 1. А именно, необходимо и достаточно, чтобы ограниченная на [0; 1] функция $f$ была интегрируема по Риману-Стилтьесу по $F$.

Напомним определение интегрируемости функции $f$ по Риману-Стилтьесу по функции распределения $F$ [8, с.341]. Рассмотрим разбиение $T = \{C_1, C_2, \dots, C_k\}$, где

$$C_i = [y_{i-1}, y_i),\ i = 1, 2, \dots, m-1,\ C_m = [y_{m-1}, y_m], \qquad (12)$$
$$0 = y_0 < y_1 < y_2 < \dots < y_m = 1.$$

Выберем в $C_i$ произвольную точку $x_i$, $i = 1, 2, \dots, m$, и составим сумму

$$S(T) = \sum_{i=1}^{m} f(x_i)[F(y_i) - F(y_{i-1})].$$

Если при $\max(y_i - y_{i-1}) \to 0$ эти суммы стремятся к некоторому пределу (не зависящему ни от способа дробления отрезка [0; 1], ни от выбора точек $x_i$ в каждом из элементов разбиения), то этот предел называется интегралом Римана-Стилтьеса от функции $f$ по функции $F$ по отрезку [0; 1] и обозначается символом, приведенным в правой части равенства (11).

Рассмотрим суммы Дарбу-Стилтьеса

$$S_H(T) = \sum_{i=1}^{m} m_i[F(y_i) - F(y_{i-1})], \quad S_B(T) = \sum_{i=1}^{m} M_i[F(y_i) - F(y_{i-1})],$$

где

$$m_i = \inf\{f(x), x \in X_i\}, \quad M_i = \sup f\{f(x), x \in X_i\}.$$

Ясно, что

$$S_H(T) \le S(T) \le S_B(T).$$

Необходимым и достаточным условием интегрируемости по Риману-Стилтьесу является следующее: для любой последовательности разбиений $T_k$, $k = 1, 2, 3, \dots$ вида (12) такой, что $\max(y_i - y_{i-1}) \to 0$ при $k \to \infty$, имеем

$$\lim_{k\to\infty}[S_B(T_k)-S_H(T_k)]=0. \qquad (13)$$

Напомним, что согласно подразделу 1.4.3 колебанием д($f$, $B$) функции $f$ на множестве $B$ называется д($f$, $B$) = $\sup\{|f(x)-f(y)|, x\in B, y\in B\}$. Поскольку

$$д(f, C_i) = M_i - m_i,$$

то условие (13) можно записать в виде

$$\lim_{k\to\infty}\sum_{C\in T_k}\delta(f,C)F(C)=0. \qquad (14)$$

Условие (14), допускающее обобщение с $X$ = [0; 1] и $f$: [0; 1] → $R^1$ на $X$ и $f$ более общего вида, и будем использовать при доказательстве теорем 1 и 2.

*Доказательство теоремы 1.* Согласно методу аппроксимации ступенчатыми функциями рассмотрим оператор $A_T$. Как легко проверить, имеет место разложение

$$\beta_n=\int_0^1 f(x)dF_n(x)-\int_0^1 f(x)dF(x)=\int_0^1\{f(x)-A_Tf(x)\}dF_n(x)+$$
$$+\int_0^1\{A_Tf(x)-f(x)\}dF(x)+\left\{\int_0^1 A_Tf(x)dF_n(x)-\int_0^1 A_Tf(x)dF(x)\right\}. \qquad (15)$$

Поскольку

$$|f(x) - A_Tf(x)| \le д(f, X_i),\ x\in C_i,$$

то первое слагаемое в правой части (15) не превосходит

$$\sum_{C\in T}\delta(f,C)F_n(C), \qquad (16)$$

а второе не превосходит

$$\sum_{C\in T}\delta(f,C)F(C).$$

Согласно определению оператора $A_T$ третье слагаемое в (15) имеет вид

$$\sum_{i=1}^{m} f(x_i)(F_n(C_i)-F(C_i)).$$

Очевидно, оно не превосходит по модулю

$$\sup_{x\in X}|f(x)|\sum_{C\in T}|F_n(C)-F(C)|$$

(здесь используется ограниченность $f$ на $X$).

Согласно (16) первое слагаемое в правой части (15) не превосходит

$$\sum_{C\in T}\delta(f,C)F(C)+\sum_{C\in T}\delta(f,C)|F_n(C)-F(C)|.$$

Поскольку

$$\delta(f,C)\le 2\sup_{x\in X}|f(x)|,$$

то первое слагаемое в правой части (15) не превосходит

$$\sum_{C\in T}\delta(f,C)F(C)+2\sup_{x\in X}|f(x)|\sum_{C\in T}|F_n(C)-F(C)|.$$

Из оценок, относящихся к трем слагаемым в разложении (15), следует, что

$$|\beta_n|\le 2\sum_{C\in T}\delta(f,C)F(C)+3\sup_{x\in X}|f(x)|\sum_{C\in T}|F_n(C)-F(C)|. \qquad (17)$$

Используя оценку (17), докажем, что в$_n$ → 0 при $n$ → ∞. Пусть дано е > 0. Согласно условию интегрируемости функции $f$ по Риману-Стилтьесу, т.е. условию (14), можно указать разбиение $T$ = $T$(е) такое, что

$$\sum_{C\in T(\varepsilon)}\delta(f,C)F(C)<\frac{\varepsilon}{4}, \qquad (18)$$

и в точках $y_i$, $i$ = 1, 2, …, $m$ - 1 (см. (12)), функция $F$ непрерывна.

Поскольку

$$F_n(X_i) = F_n(y_i) - F_n(y_{i-1}),$$

то из (10) следует, что существует число $n = n(\varepsilon)$ такое, что при $n > n(\varepsilon)$ справедливо неравенство

$$\sum_{C\in T(\varepsilon)} |F_n(C) - F(C)| < \frac{\varepsilon}{6}\left(\sup_{x\in X}|f(x)|\right)^{-1}. \qquad (19)$$

Из (17), (18) и (19) следует, что при $n > n(\varepsilon)$ справедливо неравенство

$$\left|\int_0^1 f(x)dF_n(x) - \int_0^1 f(x)dF(x)\right| < \varepsilon,$$

что и требовалось доказать.

Обсудим условие ограниченности $f$. Если оно не выполнено, то из (10) не всегда следует (11).

*Пример 2.* Пусть $f(x) = 1/x$ при $x > 0$ и $f(0)=0$. Пусть $F(0,5) = 0$, т.е. предельное распределение сосредоточено на [1/2; 1]. Пусть распределение $F_n$ на [0; S) имеет единственный атом в точке $x = 1/n$ величиной $n^{-1/2}$, а на [1/2; 1] справедливо (10). Тогда по причинам, изложенным при доказательстве теоремы 1,

$$\lim_{n\to\infty}\int_{1/2}^1 f(x)dF_n(x) = \int_{1/2}^1 f(x)dF(x),$$

однако

$$\int_0^{1/2} f(x)dF_n(x) = \sqrt{n}, \quad \int_0^{1/2} f(x)dF(x) = 0,$$

т.е. соотношение (11) не выполнено.

Условие ограниченности подынтегральной функции $f$ можно заменить, как это сделано, например, в [9], на условие строгого возрастания функции распределения $F$.

*Лемма.* Пусть функции распределения $F$ всюду строго возрастает, т.е. из $x_1 < x_2$ вытекает $F(x_1) < F(x_2)$. Пусть функция $f$ интегрируема по Риману-Стилтьесу по $F$, т.е. выполнено (14). Тогда функция $f$ ограничена.

*Доказательство.* Рассмотрим точки $0 = y_0 < y_1 < y_2 < \ldots < y_{2m} = 1$ и два разбиения

$$T_1 = \{[0;y_1),[y_1;y_3),[y_3;y_5),...,[y_{2m-1};1]\},\ T_2 = \{[0;y_2),[y_2;y_4),[y_4;y_6),...,[y_{2m-2};1]\}.$$

Тогда для любых двух точек $x$ и $x'$ можно указать конечную последовательность точек $x_1 = x, x_2, x_3, \ldots, x_s, x_{s+1} = x'$ такую, что любые две соседние точки $x_i, x_{i+1}$, $i = 1, 2, \ldots, s$, одновременно принадлежат некоторому элементу $C_i$ разбиения $T_1$ или разбиения $T_2$, причем $C_i \neq C_j$ при $i \neq j$. Действительно, пусть $x \in [y_p; y_{p+1}), x' \in [y_q; y_{q+1})$. Пусть для определенности $q > p$. Тогда можно положить $x_2 = y_{p+1}, x_3 = y_{p+2}, \ldots, x_s = y_q$. Поскольку среди элементов разбиений $T_1$ и $T_2$ есть $C_1 = [y_p; y_{p+2})$, то $x \in x_1 \in C_1, x_2 = y_{p+1} \in C_1$. Далее, $x_2 \in [y_{p+1}, y_{p+3}) = C_2, x_3 \in C_2$, и т.д.

Из указанных выше свойств последовательности $x_1 = x, x_2, x_3, \ldots, x_s, x_{s+1} = x'$ следует, что

$$|f(x) - f(x')| \le \sum_{i=1}^{s} |f(x_{i+1}) - f(x_i)| \le \sum_{C\in T_1}\delta(f,C) + \sum_{C\in T_2}\delta(f,C).$$

Пусть теперь число $\max(y_i - y_{i-2})$ настолько мало, что согласно (14)

$$\sum_{C\in T_1}\delta(f,C)F(C) < 1,\ \sum_{C\in T_2}\delta(f,C)F(C) < 1.$$

Тогда согласно двум последним соотношениям

$$|f(x) - f(x')| \le 2[\min\{F(C): C \in T_1 \cup T_2\}]^{-1},$$

что и доказывает лемму.

*Доказательство теоремы 2.* Пусть условие (14) не выполнено, т.е. существуют число г > 0 и последовательность разбиений $T_n$, $n = 1, 2, \ldots$, такие, что $\max(y_i - y_{i-1}) \to 0$ при $n\to\infty$ и при всех $n$

$$\sum_{C\in T_n}\delta(f,C)F(C)\ge\gamma\,. \qquad (20)$$

Для доказательства теоремы построим две последовательности функций распределения $F_{1n}$ и $F_{2n}$, $n$ = 1, 2, …, для которых выполнено (10), но последовательность

$$\delta_n=\int_0^1 f(x)dF_{1n}(x)-\int_0^1 f(x)dF_{2n}(x)$$

не стремится к 0 при $n\to\infty$. Тогда (11) не выполнено хотя бы для одной из последовательностей $F_{1n}$ и $F_{2n}$.

Для любого $C$ – элемента некоторого разбиения $T$ – можно указать, как вытекает из определения д($f$, $C$), точки $x_1(C)$ и $x_2(C)$ такие, что

$$f(x_1(C))-f(x_2(C))>\text{S д}(f,C). \qquad (21)$$

Построим $F_{1n}$ и $F_{2n}$ следующим образом. Пусть $F_{1n}(C)=F_{2n}(C)=F(C)$ для любого $C$ из $T_n$. При этом $F_{1n}$ имеет в $C$ один атом в точке $x_1(C)$ величиной $F(C)$, а $F_{2n}$ имеет в $C$ также один атом в точке $x_2(C)$ той же величины $F(C)$. Другими словами, распределение $F_{1n}$ в $C$ сосредоточено в одной точке, а именно, в $x_1(C)$, а распределение $F_{2n}$ сосредоточено в $x_2(C)$. Тогда

$$\delta_n=\sum_{C\in T_n}(f(x_1(C))-f(x_2(C)))F(C)\,. \qquad (22)$$

Из (20), (21) и (22) следует, что

$$\delta_n\ge\frac{1}{2}\sum_{C\in T_n}\delta(f,C)F(C)\ge\frac{\gamma}{2}\,.$$

Остается показать, что для последовательностей функций распределения $F_{1n}$ и $F_{2n}$ выполнено (10). Пусть $x$ – точка непрерывности $F$. Пусть

$$y_1(x,\,T)=\max\{y_{kn}\colon y_{kn}<x\},\ y_2(x,\,T)=\min\{y_{kn}\colon y_{kn}>x\},$$

где $y_{kn}$ – точки, определяющие разбиения $T_n$ согласно (12). В соответствии с определением $F_{in}$

$$F_{in}(y_j(x,\,T_n))=F(y_j(x,\,T_n)),\ i=1,2,\ j=1,2,$$

а потому

$$|F_{in}(\text{x})-F(x)|\le F(y_2(x,\,T_n))-F(y_1(x,\,T_n)),\ i=1,2.$$

В силу условия $\max(y_{kn}-y_{(k-1)n})\to 0$ и непрерывности $F$ в точке $x$ правая часть последнего соотношения стремится к 0 при $n\to\infty$, что и заканчивает доказательство теоремы 2.

Теоремы 1 и 2 демонстрируют основные идеи предельной теории статистик интегрального типа и непараметрических критериев в целом. Как показывают эти теоремы, основную роль в рассматриваемой теории играет предельное соотношение (14). Отметим, что если д($f$, $T_n$) $\to 0$ при $n\to\infty$, то (14) справедливо, но, вообще говоря, не наоборот. Естественно возникает еще ряд постановок. Пусть (14) выполнено для $f_1$ и $f_2$. При каких функциях $h$ это соотношение выполнено для $h(x, f_1(x), f_2(x))$? В прикладной статистике вместо $f(x)$ рассматривают $f_б(x,$ щ) и $f(x,$ щ), а вместо интегрирования по функциям распределения $F_n(x)$ – интегрирование по случайным мерам $F_б$(щ). Как меняются формулировки в связи с такой заменой? В связи со слабой сходимостью (т.е. сходимостью по распределению) $A_T f_б$ к $A_T$ и переходом от $f_б(x,$ щ) к $h_б(x, f_{1б}(x,$ щ), $f_{2б}(x,$ щ)) возникает следующая постановка. Пусть $к_б$ слабо сходится к к при б$\to\infty$. Когда распределения $g_б(к_б)$ сближаются с распределениями $g_б$(к)? Полным ответом на последний вопрос являются необходимые и достаточные условия наследования сходимости. Они приведены в главе 1.4.

**Основные результаты.** Наиболее общая теорема типа теоремы 1 выглядит так [10].

*Теорема 3*. Пусть существует последовательность разбиений $T_n$, $n$ = 1, 2, …, такая, что при $n\to\infty$ и б$\to\infty$

$$\Delta(f_\alpha,T_n)=\sum_{C\in T_n}\delta(f_\alpha,C)F(C)\to 0\,. \qquad (23)$$

Пусть для любого $C$, входящего хотя бы в одно из разбиений $T_n$,

$$F_б(C,\text{ щ})\to F(C) \qquad (24)$$

при б→∞ (сходимость по вероятности). Пусть $f_б$ асимптотически ограничены по вероятности при б→∞. Тогда

$$\xi(f_\alpha, F_\alpha) - \xi(f_\alpha, F) \to 0 \qquad (25)$$

при б→∞ (сходимость по вероятности).

Как известно, полное сепарабельное метрическое пространство называется польским. Это понятие понадобится для формулировки аналога теоремы 2.

*Теорема 4.* Пусть $X$ – польское пространство, У конечномерно, существует измельчающаяся последовательность $T_n$ разбиений, для которой соотношение (23) не выполнено. Тогда существует удовлетворяющая (24) последовательность $F_б$, для которой соотношение (25) неверно, хотя $F_б$ слабо сходится к $F$ при б→∞.

Условие (23) естественно назвать условием римановости, поскольку в случае, рассмотренном в теореме 1, оно является условием интегрируемости по Риману-Стилтьесу. Рассмотрим *наследуемость римановости* при переходе от $f_{1б}(x, щ)$ со значениями в $У_1$ и $f_{2б}(x, щ)$ со значениями в $У_2$, удовлетворяющих (23), к $h_б(x, f_{1б}(x, щ), f_{2б}(x, щ))$ со значениями в $У_3$.

Положим

$$Y_k(a,\varepsilon) = \{(y,y') : y \in Y_k, y' \in Y_k, \| y \|_k < a, \| y' \| < a, \| y - y' \| < \varepsilon\}, k = 1,2,$$

где $\|\cdot\|_k$ – норма (т.е. длина вектора) в пространстве $Y_k$, $k = 1, 2$. Рассмотрим также множества

$$A(C,a,\varepsilon) = \{(x,x',y_1,y_1^*,y_2,y_2^*) : x,x' \in C, (y_k,y_k^*) \in Y_k(a,\varepsilon), k = 1,2\}$$

и функции

$$q_\alpha(x,x',y_1,y_1^*,y_2,y_2^*) = h_\alpha(x,y_1,y_2) - h_\alpha(x',y_1^*,y_2^*).$$

Наконец, понадобится измеритель колеблемости

$$c(h_\alpha, T, a, \varepsilon) = \sum_{C \in T} \sup_{A(C,a,\varepsilon)} \| q_\alpha \|_3 \, F(C)$$

и множество

$$Z(a) = X \times \{y_1 : \| y_1 \| < a\} \times \{y_2 : \| y_2 \| < a\}.$$

*Теорема 5.* Пусть $h_б$ асимптотически (при б→∞) ограничены на $Z(a)$ при любом положительном $a$, функции $f_{1б}$ и $f_{2б}$ асимптотически ограничены по вероятности и удовлетворяют условию (23). Пусть для участвующей в (23) последовательности $T_n$

$$c(h_б, T_n, a, е) \to 0 \qquad (26)$$

при б→∞, $n$→∞, е→ 0 и любом положительном $a$. Тогда $f_{3б}(x, щ) = h_б(x, f_{1б}(x, щ), f_{2б}(x, щ))$ удовлетворяют условию (23) и асимптотически ограничены по вероятности.

*Теорема 6.* Пусть условие (26) не выполнено для $h_б$. Тогда существуют детерминированные ограниченные функции $f_{1б}$ и $f_{2б}$ такие, что соотношение (23) выполнено для $f_{1б}$ и $f_{2б}$ и не выполнено для $f_{3б}$.

*Пример 3.* Пусть $X = [0; 1]^k$, пространства $Y_1$ и $Y_2$ конечномерны, функция $h_б \equiv h(x, y_1, y_2)$ непрерывна. Тогда условие (26) выполнено.

С помощью теорем 3 и 5 и результатов о наследовании сходимости можно изучить асимптотическое поведение статистик интегрального типа

$$\xi_\alpha = \int_X h_\alpha(x, f_{1\alpha}(x,\omega), f_{2\alpha}(x,\omega)) F_\alpha(dx,\omega)$$

со значениями в банаховом пространстве $У$.

*Теорема 7.* Пусть для некоторой последовательности $T_n$ разбиений $X$ справедливы соотношения (23) для $f_{1б}$ и $f_{2б}$ и (24) для $F_б$. Пусть последовательность функций $h_б$ удовлетворяет условию в теореме 5, конечномерные распределения ($f_{1б}(x, щ)$, $f_{2б}(x, щ)$) слабо сходятся к конечномерным распределениям ($f_1(x, щ)$, $f_2(x, щ)$), причем для $f_1$ и $f_2$ справедливо соотношение (23). Тогда

$$\lim_{\alpha \to \infty} L(\xi_\alpha, \eta_\alpha) = 0,$$

где L – расстояние Прохорова (см. подраздел 1.4.3),

$$\eta_\alpha = \int_X h_\alpha(x, f_1(x,\omega), f_2(x,\omega))F(dx)\,.$$

Теорема 7 дает общий метод получения асимптотических распределений статистик интегрального типа. Важно, что соотношение (23) выполнено для эмпирического процесса и для процессов, связанных с оцениванием параметров при проверке согласия [9].

Один из выводов общей теории состоит в том, что в качестве $F_6$ можно использовать практически любую состоятельную оценку истинной функции распределения. Этот вывод использовался при построении критерия типа омега-квадрат для проверки симметрии распределения относительно 0 и обнаружения различий в связанных выборках (глава 3.1).

Асимптотическое поведение критериев типа Колмогорова может быть получено с помощью описанного выше метода аппроксимации ступенчатыми функциями. Этот метод не требует обращения к теории сходимости вероятностных мер в функциональных пространствах. Для критериев Колмогорова и Смирнова достаточно использовать лишь свойства эмпирического процесса и броуновского моста. В случае проверки согласия добавляется необходимость изучения еще одного случайного процесса. Он является разностью между двумя функциями распределения. Одна - функция распределения элементов выборки. Вторая - член параметрического семейства распределений, полученный путем подстановки оценок параметров вместо их истинных значений.

### 2.3.4. Метод проверки гипотез по совокупности малых выборок

Одна из областей применения прикладной статистики – статистические методы управления качеством продукции [13, гл.13]. К ним относится статистический приемочный контроль, в котором по результатам испытаний элементов выборки делается вывод о качестве партии продукции. В простейшем варианте проводится контроль по альтернативному признаку, при котором возможны лишь два результата контроля конкретной единицы продукции – «соответствует требованиям» или «не соответствует требованиям», короче – «да» или «нет».

Рассмотрим статистический приемочный контроль по двум альтернативным признакам одновременно. На основе теории люсианов обсудим проблему проверки независимости двух альтернативных признаков. Ее приходится проводить по совокупности малых выборок, т.е. в так называемой асимптотике А.Н.Колмогорова, когда число неизвестных параметров распределения не является постоянным, а растет пропорционально объему данных.

**Испытания по двум альтернативным признакам.** При статистическом контроле качества продукции, в частности, при сертификации, чаще всего используют контроль по альтернативным признакам. При этом устанавливается, соответствует ли контролируемый параметр единицы продукции (изделия, детали) заданным в нормативно-технической документации требованиям или не соответствует. Если соответствует - единица продукции признается годной. Примем для определенности, что в этом случае результат контроля кодируется символом 0. Если же не соответствует - единица продукции признается дефектной, а результат контроля кодируется символом 1.

Таким образом, в рассматриваемой нами математической модели контроля альтернативный признак - это функция $X = X(w)$, определенная на множестве единиц продукции $W = \{w\}$ и принимающая два значения 0 и 1. Причем $X(w) = 0$ означает, что единица продукции $w$ является годной, а $X(w) = 1$ - что она является дефектной.

Методы статистического контроля, в частности, включенные в государственные стандарты и иную нормативно-техническую документацию (НТД), как правило, используют контроль по одному признаку. В НТД указывают правила выбора планов контроля и расчета различных их характеристик, приводят графики оперативных характеристик и т.п.

Однако на производстве контроль нередко проводится по нескольким альтернативным признакам. Возникает проблема выбора плана контроля и расчета его характеристик.

Рассмотрим сначала контроль по двум альтернативным признакам $X(w)$ и $Y(w)$. В вероятностной модели $X(w)$ и $Y(w)$ - случайные величины, принимающие два значения - 0 и 1. Пусть, пользуясь стандартной (для статистических методов управления качеством) терминологией,

$$p_1 = P(X(w) = 1)$$

- входной уровень дефектности для первого признака, а

$$p_2 = P(Y(w) = 1)$$

- для второго. Вероятности результатов контроля по двум признакам одновременно описываются четырьмя числами:

$$P(X(w) = 0,\ Y(w) = 0) = p_{00},\ P(X(w) = 1,\ Y(w) = 0) = p_{10},$$
$$P(X(w) = 0,\ Y(w) = 1) = p_{01},\ P(X(w) = 1,\ Y(w) = 1) = p_{11}.$$

При этом справедливы соотношения:

$$p_{00} + p_{10} + p_{01} + p_{11} = 1,\ \ p_{10} + p_{11} = p_1,\ p_{01} + p_{11} = p_2.$$

С прикладной точки зрения наиболее интересна вероятность $p_{00}$ того, что единица продукции является годной (по всем параметрам), и вероятность ее дефектности ($1-p_{00}$ ), т.е. входной уровень дефектности для изделия в целом.

В табл.1 сведены вместе введенные выше вероятности.

Таблица 1.
Вероятности результаты испытаний при контроле по двум альтернативным признакам

| | $X$=0 | $X$=1 | Всего |
|---|---|---|---|
| $Y$=0 | $p_{00}$ | $p_{10}$ | $1 - p_2$ |
| $Y$=1 | $p_{01}$ | $p_{11}$ | $p_2$ |
| Всего | $1 - p_1$ | $p_1$ | 1 |

Есть три важных частных случая - поглощения, несовместности и независимости дефектов. Другими словами, поглощения, несовместности и независимости событий $\{w: X(w) = 1\}$ и $\{w: Y(w) = 1\}$. В случае поглощения одно из этих событий содержит другое, а потому

$$p_{00} = 1 - \max(p_1, p_2).$$

В случае несовместности

$$p_{00} = 1 - p_1 - p_2.$$

В случае независимости

$$p_{00} = (1 - p_1)(1 - p_2) = 1 - p_1 - p_2 + p_1p_2.$$

Очевидно, что вероятность годности изделия всегда заключена между значениями, соответствующими случаям поглощения и несовместности. Кроме того, известно, что при большом числе признаков и малой вероятности дефектности по каждому из них случаи поглощения и независимости дают (в асимптотике) крайние значения для вероятности годности изделия, т.е. формулы, соответствующие независимости и несовместности, асимптотически совпадают. Причина этого явления состоит в том, что при малости $p_1$ и $p_2$ их произведение $p_1p_2$является бесконечно малой более высокого порядка по сравнения с $p_1$ и $p_2$.

Рассмотрим несколько примеров. Пусть некоторая продукция, скажем, гвозди, контролируются по двум альтернативным признакам, для определенности, по весу и длине. Пусть результаты контроля 1000 единиц продукции представлены в табл.2

Таблица 2.
Результаты 1000 испытаний по двум альтернативным признакам (случай поглощения)

| | $X$=0 | $X$=1 | Всего |
|---|---|---|---|

| | | | |
|---|---|---|---|
| *У*=0 | 952 | 0 | 952 |
| *У*=1 | 0 | 48 | 48 |
| Всего | 952 | 48 | 1000 |

Судя по данным табл.2, дефекты всегда встречаются парами - если есть один, то есть и другой. Входной уровень дефектности как по каждому показателю, так и по обоим вместе - один и тот же, а именно, 0,048. Получив по результатам статистического наблюдения данные типа приведенных в табл.2, целесообразно перейти к контролю только одного показателя, а не двух. Каково именно? Видимо, того, контроль которого дешевле. Однако совсем иная ситуация в случае несовместности дефектов (табл.3).

Таблица 3.
Результаты 1000 испытаний по двум альтернативным признакам (случай несовместности )

| | *Х*=0 | *Х*=1 | Всего |
|---|---|---|---|
| *У*=0 | 904 | 48 | 952 |
| *У*=1 | 48 | 0 | 48 |
| Всего | 952 | 48 | 1000 |

Судя по данным табл.3, дефекты всегда встречаются поодиночке - если есть один, то другого нет. В результате входной уровень дефектности по каждому признаку по-прежнему равен 0,048, в то время как доля дефектных изделий (т.е. имеющих хотя бы один дефект) вдвое выше, т.е. входной уровень дефектности для изделия в целом равен 0,096.

Случай независимости результатов контроля по двум независимым признакам (табл.4) лежит между крайними случаями поглощения и несовместности. Независимость альтернативных признаков обосновывается путем статистической проверки с помощью описанного ниже критерия $n^{1/2}V$.

Таблица 4.
Результаты 1000 испытаний по двум альтернативным признакам (случай независимости)

| | *Х*=0 | *Х*=1 | Всего |
|---|---|---|---|
| *У*=0 | 909 | 43 | 952 |
| *У*=1 | 43 | 5 | 48 |
| Всего | 952 | 48 | 1000 |

Согласно данным табл.4, входной уровень дефектности для каждого из двух альтернативных признаков по-прежнему равен 0,048, в то время как для изделий в целом он равен 0,091, т.е. на 5,2% меньше, чем в случае несовместности, и на 89,6% больше, чем в случае поглощения.

Проблема состоит в том, что таблицы и стандарты по статистическому приемочному контролю относятся обычно к случаю одного контролируемого параметра. А как быть, если контролируемых параметров несколько? Приведенные выше примеры показывают, что входной уровень дефектности изделия в целом не определяется однозначно по входным уровням дефектности отдельных его параметров.

**Гипотеза независимости.** Как должны соотноситься характеристики планов контроля по отдельным признакам с характеристиками плана контроля по двум (или многим) признакам одновременно? Рассмотрим распространенную рекомендацию - складывать уровни дефектности, т.е. считать, что уровень дефектности изделия в целом равен сумме уровней дефектности по отдельным его параметрам. Она, очевидно, опирается на гипотезу несовместности дефектов, а

потому во многих случаях преувеличивает дефектность, следовательно, ведет к использованию излишне жестких планов контроля, что экономически невыгодно.

Зная специфику применяемых технологических процессов, в ряде конкретных случаев можно предположить, что дефекты по различным признакам возникают независимо друг от друга. Это предположение необходимо обосновывать по статистическим данным. Если же оно обосновано, следует рассчитывать входной уровень дефектности по формуле

$$1 - p_{00} = p_1 + p_2 - p_1 p_2,$$

соответствующей независимости признаков.

Итак, необходимо уметь проверять по статистическим данным гипотезу независимости двух альтернативных признаков. Речь идет о статистической проверке нулевой гипотезы

$$H_0: \ p_{11} = p_1 p_2 \qquad (1)$$

(что эквивалентно проверке равенства $p_{00} = (1 - p_1)(1 - p_2)$). Нетрудно проверить, что гипотеза о справедливости равенства (1) эквивалентна гипотезе

$$H_0 : p_{00} p_{11} - p_{10} p_{01} = 0. \qquad (2)$$

В простейшем случае предполагается, что проведено $n$ независимых испытаний $(Xi\ ,\ Y_i)$, $i = 1, 2, ..., n,$ в каждом из которых проконтролированы два альтернативных признака, а вероятности результатов контроля не меняются от испытания к испытанию. Общий вид статистических данных приведен в табл.5.

Таблица 5.
Общий вид результатов контроля
по двум альтернативным признакам.

| | $X$=0 | $X$=1 | Всего |
|---|---|---|---|
| $У$=0 | $a$ | $b$ | $a+b$ |
| $У$=1 | $c$ | $d$ | $c+d$ |
| Всего | $a+c$ | $b+d$ | $n$ |

В табл.5 величина $a$ - число испытаний, в которых $(Xi\ ,\ Y_i) = (0,0)$, величина $b$ - число испытаний, в которых $(Xi\ ,\ Y_i) = (1,0)$, и т.д.

Случайный вектор $(a, b, c, d)$ имеет мультиномиальное распределение с числом испытаний $n$ и вектором вероятностей исходов $(p_{00}, p_{10}, p_{01}, p_{11})$. Состоятельными оценками этих вероятностей являются дроби $a/n, b/n, c/n, d/n$ соответственно. Следовательно, критерий проверки гипотезы (2) может быть основан на статистике

$$Z = ad - bc. \qquad (3)$$

Как вытекает из известной формулы для ковариаций мультиномиального вектора (см., например, формулу (6.3.5) в учебнике С.Уилкса [14] на с. 153),

$$M(Z) = n\,(p_{10} p_{01} - p_{00} p_{11}), \qquad (4)$$

что равно 0 при справедливости гипотезы независимости (2).

Связь между переменными $X$ и $Y$ обычно измеряется коэффициентом, отличающимся от $Z$ нормирующим множителем:

$$V = (ad - bc)\{(a + b)(a + c)(b + d)(c + d)\}^{-1/2}$$

(см. классическую монографию М. Дж. Кендалла и А. Стьюарта [15, с.723]). При справедливости гипотезы $H_0$ и больших $n$ случайная величина $nV^2$ имеет хи-квадрат распределение с одной степенью свободы, а $n^{1/2}V$ имеет стандартное нормальное распределение с математическим ожиданием 0 и дисперсией 1 (см. [15, с.736]). Значение $n^{1/2}V$ для данных табл.4 равно 1,866, т.е на уровне значимости 0,05 гипотезу независимости следует принять.

Рассмотрим еще один пример. Пусть проведено 100 испытаний, результаты которых описаны в табл.6. Тогда

$$V = (50 \cdot 20 - 10 \cdot 20)\,(60 \cdot 70 \cdot 30 \cdot 40)^{-1/2} =$$

$$= (1000 - 200) \cdot 5940000^{-1/2} = 800 / 2245 = 0{,}35635,$$
$$n^{1/2}V = 3{,}5635 \ .$$

Таблица 6.

Результаты 100 испытаний по двум альтернативным признакам.

| | $X$=0 | $X$=1 | Всего |
|---|---|---|---|
| $Y$=0 | 50 | 10 | 60 |
| $Y$=1 | 20 | 20 | 40 |
| Всего | 70 | 30 | 100 |

Поскольку полученное значение $n^{1/2}V$ превышает критическое значение при любом применяемом в статистике уровне значимости, то гипотезу о независимости признаков необходимо отклонить.

**Проверка гипотез по совокупности малых выборок.** К сожалению, приведенный простой метод годится не всегда. При статистическом анализе реальных данных возникают проблемы, связанные с отсутствием достаточно больших однородных выборок, т.е. выборок, в которых постоянны параметры вероятностных распределений. Реально единицы продукции представляются на контроль партиями, из каждой партии контролируются лишь несколько изделий, т.е. малая выборка. При этом от партии к партии меняются параметры $p_{00}$, $p_{10}$, $p_{01}$, $p_{11}$, описывающие уровень дефектности. Поэтому необходимы статистические методы, позволяющие проверять гипотезу независимости признаков по совокупности малых выборок. Построим один из возможных методов.

Рассмотрим вероятностную модель совокупности $k$ малых выборок объемов $n_1$ , $n_2$ ,..., $n_k$ соответственно. Пусть $j$ -я выборка ($X_{jt}$ , $Y_{jt}$), $t$ = 1, 2,..., $n_j$ , имеет распределение, задаваемое вектором параметров ($p_{00}j$, $p_{10}j$, $p_{01}j$, $p_{11}j$) в соответствии с ранее введенными обозначениями, $j$ = 1,2,...,$k$ . Будем проверять гипотезу

$$H_0: \ p_{11}j = (p_{10}j + p_{11}j)\,(p_{01}j + p_{11}j), j = 1,2,...,k, \qquad (5)$$

или в эквивалентной формулировке

$$H_0: \ p_{11}j\,p_{00}j - p_{10}j\,p_{01}j \ , j = 1,2,...,k \ . \qquad (6)$$

Основная идея состоит в нахождении асимптотического распределения статистики типа $n^{1/2}V$ при росте числа $k$ малых выборок, а именно, статистики

$$S = g_1 Z_1 + g_2 Z_2 + ... + g_k Z_k \ , \qquad (7)$$

где $Z_1$, $Z_2$, ..., $Z_k$ - статистики, рассчитанные по формуле (3) для каждой из $k$ выборок, т.е. $Z_j = a_j d_j - b_j c_j$ , $j$ = 1,2,...,$k$, а $g_1$ , $g_2$ , ... , $g_k$ - некоторые весовые коэффициенты, которые, в частности, могут совпадать. Поскольку

$$M(S) = g_1 M(Z_1) + g_2 M(Z_2) + ... + g_k M(Z_k),$$

то при справедливости гипотезы независимости (5) - (6) имеем $M(S)$ = 0 согласно соотношению (4). Поскольку слагаемые в сумме (7) независимы, то при росте $k$ случайная величина $S$ в силу Центральной Предельной Теоремы является асимптотически нормальной. Дисперсия этой величины равна сумме дисперсий слагаемых:

$$D(S) = g_1^2 D(Z_1) + g_2^2 D(Z_2) + ... + g_k^2 D(Z_k) \ . \qquad (8)$$

Для оценивания дисперсии $S$ необходимо использовать **несмещенные** оценки дисперсий в каждой из $k$ выборок (и в этом одна из основных "изюминок" разбираемого метода). Предположим, что построены статистики $T_j$ такие, что

$$M(T_j) = D(Z_j) \ , j = 1, 2, ..., k. \qquad (9)$$

Тогда при некоторых математических "условиях регулярности", на которых нет необходимости здесь останавливаться, несмещенная оценка дисперсии статистики $S$, имеющая согласно формулам (8) и (9) вид

$$L = g_1^2 T_1 + g_2^2 T_2 + ... + g_k^2 T_k \ ,$$

в силу закона больших чисел такова, что дробь $D(S)/L$ приближается к 1 при росте числа выборок (сходимость по вероятности). Отсюда следует, что распределение случайной величины $Q = SL^{-1|2}$ приближается при росте числа выборок к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1. Следовательно, критерий проверки гипотезы (5) - (6) независимости признаков, состоящий в том, что при (-1,96) < $Q$ < 1,96 гипотеза принимается, а при $Q$ , выходящих за пределы интервала (-1,96; 1,96), гипотеза отклоняется, имеет уровень значимости, приближающийся к 0,05 при росте числа выборок. Мощность этого критерия зависит от величины $M(S)D(S)^{-1|2}$ при альтернативе.

Для реализации намеченного плана осталось научиться несмещенно оценивать $D(Z_j)$. К сожалению, в литературе по несмещенному оцениванию не рассматривают случай мультиномиального распределения, поэтому кратко опишем процедуру построения несмещенной оценки $D(Z_j)$. Поскольку согласно формулам (3) и (4)

$$D(Z_j) = M(Z_j^2) - (M(Z_j))^2 = M(a_j^2 d_j^2) - 2M(a_j b_j c_j d_j) + \\ + M(b_j^2 c_j^2) + n_j^2 (p_{00j} p_{11j} - p_{01j} p_{10j})^2, \qquad (10)$$

то для вычисления $D(Z_j)$ достаточно найти входящие в правую часть формулы (10) начальные смешанные моменты мультиномиального распределения (четвертого порядка). Теоретически это просто - известен вид характеристической функции мультиномиального распределения (см., например, формулу (6.3.4) в монографии [14, с.152]), а начальные смешанные моменты равны значениям ее соответствующих производных в 0, деленным на нужную степень мнимой единицы (формула (5.2.3) в монографии [14, с.131]). Например, с помощью описанной процедуры после некоторых вычислений получаем, что (для упрощения записи здесь и далее опустим индекс $j$)

$$M(a^2 d^2) = n(n-1)(n-2)(n-3) p_{11}^2 p_{00}^2 + \\ + n(n-1)(n-2)(p_{11}^2 p_{00} + p_{11} p_{00}^2) + n(n-1) p_{11} p_{00}. \qquad (11)$$

Формула (11) показывает, что начальные смешанные моменты мультиномиального распределения являются многочленами от параметров $p_{11}$, $p_{00}$, $p_{10}$, $p_{01}$ этого распределения, однако конкретный вид этих многочленов достаточно громоздок, поэтому не будем их здесь выписывать, ограничившись формулой (11) в качестве образца.

Как вытекает из формул (10) и (11), для построения несмещенной оценки $D(Z_j)$ достаточно научиться несмещенно оценивать произведения типа $p_{11}^r p_{00}^m$, где целые неотрицательные числа $r$, $m$ не превосходят 2. Эта задача решается, начиная с меньших степеней. Известно, что для ковариации мультиномиального вектора

$$M(ad) = - n p_{00} p_{11} \qquad (12)$$

(см., например, формулу (6.3.5) в монографии [14, с.153]), а потому несмещенной оценкой для $p_{00}p_{11}$ является ($-ad/n$). Далее, поскольку справедлива аналогичная (11) формула

$$M(a^2 d) = n(n-1)(n-2) p_{11} p_{00}^2 + n(n-1) p_{11} p_{00}, \qquad (13)$$

то с помощью формулы (12) преобразуем формулу (13) к виду

$$M(a^2 d + (n-1)ad) = n(n-1)(n-2) p_{11} p_{00}^2, \qquad (14)$$

т.е. несмещенной оценкой $p_{11} p_{00}^2$ является $ad(a+n-1)\{n(n-1)(n-2)\}^{-1}$.

Следующий шаг - аналогичным образом с помощью формул (12) и (14) получаем несмещенную оценку для $p_{11}^2 p_{00}^2$, а затем и для $D(Z_j)$. Промежуточные формулы опущены из-за громоздкости. Окончательный результат таков:

$$T_j = (b_j + d_j)(c_j + d_j)(a_j + c_j)(a_j + b_j)(n-1)^{-1}.$$

Как легко видеть,

$$\frac{Z_j}{\sqrt{T_j}} = V_j \sqrt{n_j - 1},$$

т.е. в случае одной выборки предлагаемый метод совпадает с классическим.

Общая идея рассматриваемого метода проверки гипотез по совокупности малых выборок состоит в том, что подбирается статистика, математическое ожидание которой для каждой малой выборки равно 0 при справедливости проверяемой гипотезы. Затем для каждой выборки строится несмещенная оценка дисперсии этой статистики. Итоговая статистика критерия для проверки гипотезы - это сумма рассматриваемых статистик для всех малых выборок, деленная на квадратный корень из суммы всех несмещенных оценок дисперсий рассматриваемых статистик. При справедливости нулевой гипотезы эта итоговая статистика имеет в асимптотике стандартное нормальное распределение (при выполнении некоторых математических "условий регулярности", которые обычно выполняются при анализе реальных статистических данных).

Впервые такой способ проверки гипотез по совокупности малых выборок был предложен в монографии [16, раздел 4.5]. Нестандартность постановки состоит в том, что число неизвестных параметров растет пропорционально объему данных, т.е. имеет место т.н. "асимптотика Колмогорова", или асимптотика растущей размерности. Дальнейшее развитие применительно к данных типа "да" - "нет" (или "годен" - "дефектен") шло в рамках теории люсианов как части статистики объектов нечисловой природы (см. главу 3.4).

### 2.3.5. Проблема множественных проверок статистических гипотез

Практика применения методов прикладной статистики часто выходит за границы классической математико-статистической теории. В качестве примера рассмотрим проверку статистических гипотез.

Базовая теоретическая модель касается проверки одной-единственной статистической гипотезы. На практике же при выполнении того или иного прикладного исследования гипотезы зачастую проверяют неоднократно. При этом, как правило, остается неясным, как влияют результаты предыдущих проверок на характеристики (уровень значимости, мощность) последующих проверок. Есть ли вообще влияние? Как его оценить? Как его учесть при формулировке окончательных выводов?

Изучены лишь некоторые схемы множественных проверок, например, схема последовательного анализа А. Вальда или схема оценивания степени полинома в регрессии путем последовательной проверки адекватности модели (см. главу 3.2). В таких исключительных постановках удается рассчитать характеристики статистических процедур, включающих множественные проверки статистических гипотез.

Однако в большинстве важных для практики случаев статистические свойства процедур анализа данных, основанных на множественных проверках, остаются пока неизвестными. Примерами являются процедуры нахождения информативных подмножеств признаков (коэффициенты для таких и только таких признаков отличны от 0) в регрессионном анализе или выявления отклонений параметров в автоматизированных системах управления.

В таких системах происходит слежение за большим числом параметров. Резкое изменение значения параметра свидетельствует об изменении режима работы системы, что, как правило, требует управляющего воздействия. Существует теория для определения границ допустимых колебаний одного или фиксированного числа параметров. Например, можно использовать контрольные карты Шухарта или кумулятивных сумм, а также их многомерные аналоги (см. главу 13 в [13]). В подавляющем большинстве постановок, согласно обычно используемым вероятностным моделям, для каждого параметра, находящемся в стабильном ("налаженном") состоянии, существует хотя и малая, но положительная вероятность того, что его значение выйдет за заданные границы. Тогда система зафиксирует резкое изменение значения параметра ("ложная разладка"). При достаточно большом числе параметров с вероятностью, близкой к 1, будет

обнаружено несколько "случайных сбоев", среди которых могут "затеряться" и реальные отказы подсистем. Можно доказать, что при большом числе параметров имеется два крайних случая - независимых (в совокупности) параметров и функционально связанных параметров, а для всех остальных систем вероятность обнаружения резкого отклонения хотя бы у одного параметра лежит между соответствующими вероятностями для этих двух крайних случаев.

Почему трудно изучать статистические процедуры, использующие множественные проверки гипотез? Причина состоит в том, что результаты последовательно проводящихся проверок, как правило, не являются независимыми (в смысле независимости случайных элементов). Более того, последовательность проверок зачастую задается исследователем произвольно.

Проблема множественных проверок статистических гипотез - часть более общей проблемы "стыковки" (сопряжения, последовательного выполнения) статистических процедур. Дело в том, что каждая процедура может применяться лишь при некоторых условиях, а в результате применения предыдущих процедур эти условия могут нарушаться. Например, часто рекомендуют перед восстановлением зависимости (регрессионным анализом) разбить данные на однородные группы с помощью какого-либо алгоритма классификации, а затем строить зависимости для каждой из выделенных групп отдельно. Здесь идет речь о "стыковке" алгоритмов классификации и регрессии. Как вытекает из рассмотрений статьи [17], попадающие в одну однородную группу результаты наблюдений зависимы и их распределение не является нормальным (гауссовым), поскольку они лежат в ограниченной по некоторым направлениям области, причем границы зависят от всей совокупности результатов наблюдений. При этом при росте объема выборки зависимость уменьшается, но ненормальность остается. Распределение результатов наблюдений, попавших в одну группу, приближается не к нормальному, а к усеченному нормальному. Следовательно, алгоритмами регрессионного анализа, основанными на "нормальной теории", пользоваться некорректно. Целесообразно применять непараметрическую или робастную регрессию.

Проблема "стыковки" статистических процедур обсуждается давно. По проблеме "стыковки" проведен ряд исследований, результаты некоторые из которых упомянуты выше, но сколько-нибудь окончательных результатов получено не было. По нашему мнению, на скорое решение проблемы "стыковки" рассчитывать нельзя. Возможно, она является столь же "вечной", как и проблема выбора между средним арифметическим и медианой как характеристиками "центра" выборки.

В качестве примера обсудим одно интересное исследование по проблеме повторных проверок статистических гипотез - работу С.Г.Корнилова [18].

Как уже отмечалось, теоретическое исследование является весьма сложным, сколько-нибудь интересные результаты удается получить лишь для отдельных постановок. Поэтому вполне естественно, что С.Г. Корнилов применил метод статистического моделирования на ЭВМ. Однако нельзя забывать о проблеме качества псевдослучайных чисел. Достоинства и недостатки различных алгоритмов получения псевдослучайных чисел много лет обсуждаются в различных изданиях (см. главу 11 в [13]).

В работе С.Г.Корнилова хорошо моделируется *мышление* статистика-прикладника. Видно, насколько мешает устаревшее представление о том, что для проверки гипотез необходимо задавать определенный уровень значимости. Особенно оно мешает, если в дальнейшем понадобятся дальнейшие проверки. Гораздо удобнее использовать "достигаемый уровень значимости", т.е. вероятность того, что статистика критерия покажет большее отклонение от нулевой гипотезы, чем то отклонение, что соответствует имеющимся экспериментальным данным. Если есть желание, можно сравнивать "достигаемый уровень значимости" с заданными значениями 0,05 или 0,01. Так, если "достигаемый уровень значимости" меньше 0,01, то нулевая гипотеза отвергается на уровне значимости 0,01, в противном случае - принимается. Следует рассчитывать "достигаемый уровень значимости" всегда, когда для этого есть вычислительные возможности.

Переход к "достигаемому уровню значимости" может избавить прикладника от еще одной трудности, связанной с использованием непараметрических критериев. Дело в том, что их распределения, как правило, дискретны, поскольку эти критерии используют только ранги наблюдений. Поэтому обычно невозможно построить критерий с заданным номинальным уровнем значимости - реальный уровень значимости может принимать лишь конечное число значений, среди которых, как правило, нет ни 0,05, ни 0,01, ни других популярных номинальных значений.

Невозможность построения критических областей критериев с заданными уровнями значимости затрудняет сравнение критериев по мощности, как это продемонстрировано в работе [19]. Есть формальный способ достичь заданного номинального уровня значимости - провести рандомизацию, т.е. при определенном (граничном) значении статистики критерия провести независимый случайный эксперимент, в котором одни исходы (с заданной суммарной вероятностью) приводят к принятию гипотезы, а остальные - к ее отклонению. Однако подобную процедуру рандомизации прикладнику трудно принять - как оправдать то, что одни и те же экспериментальные данные могут быть основанием как для принятия гипотезы, так и для ее отклонения? Вспоминается обложка журнала "Крокодил", на которой один хозяйственник говорит другому: "Бросим монетку. Упадет гербом - будем строить завод, а упадет решкой - нет". Описанная процедура рандомизации имеет практический смысл лишь при массовой рутинной проверке гипотез, например, при статистическом контроле больших выборок изделий или деталей.

При использовании все еще распространенных критерия Стьюдента и других параметрических статистических критериев - свои проблемы. Такие критерии построены исходя из предположения о том, что функции распределения результатов наблюдений входят в определенные параметрические семейства небольшой размерности. Наиболее распространена гипотеза нормальности распределения. Однако давно известно, что подавляющее большинство реальных распределений результатов измерений не являются нормальными. Об этом говорится, например, в классической для инженеров и организаторов производства монографии проф. В.В. Налимова [20]. Ряд недавно полученных конкретных экспериментальных фактов и теоретических соображений, подтверждающих точку зрения В.В. Налимова, рассмотрен в главе 2.1.

Как же быть? Проверять нормальность распределения своих данных? Но это дело непростое, можно допустить те или иные ошибки, в частности, применяя критерии типа Колмогорова или типа омега-квадрат. Как уже говорилось (в главе 1.2), одна из наиболее распространенных ошибок состоит в том, что в статистики вместо неизвестных параметров подставляют их оценки, но при этом пользуются критическими значениями, рассчитанными для случая, когда параметры полностью известны. Кроме того, для сколько-нибудь надежной проверки нормальности нужны тысячи наблюдений (см. подраздел 2.3.2). Поэтому в подавляющем большинстве реальных задач нет оснований принимать гипотезу нормальности. В лучшем случае можно говорить о том, что распределение результатов наблюдений мало отличается от нормального.

Как влияют отклонения от нормальности на свойства статистических процедур? Для разных процедур - разный ответ. Если речь идет об отбраковке выбросов - влияние отклонений от нормальности настолько велико, что делает процедуру отбраковки с практической точки зрения эвристической, а не научно обоснованной (см. главу 4). Если же речь идет о проверке однородности двух выборок с помощью критерия Стьюдента (при априорном предположении о равенстве дисперсий) или Крамера-Уэлча (при отсутствии такого предположения), то при росте объемов выборок влияние отклонений от нормальности убывает, как это подробно показано в главе 3.1. Это вытекает из Центральной Предельной Теоремы. Правда, при этом оказывается, что процентные точки распределения Стьюдента не приносят реальной пользы, достаточно использовать процентные точки предельного нормального распределения.

Весьма важна обсуждаемая, в частности, в работе [18] постоянно встающая перед исследователем проблема выбора того или иного статистического критерия для решения конкретной прикладной задачи. Например, как проверять однородность двух независимых выборок

числовых результатов наблюдений? Известны параметрические критерии: Стьюдента, Лорда; непараметрические: Крамера-Уэлча, Вилкоксона, Ван-дер-Вардена, Сэвиджа, Мартынова, Смирнова, типа омега-квадрат (Лемана - Розенблатта) и многие другие (см., например, главу 3.1 и справочник [6]). Какой из них выбрать для конкретных расчетов?

Некоторые авторы предлагают формировать технологию принятия статистического решения, согласно которой решающее правило формируется на основе комбинации нескольких критериев. Например, технология может предусматривать проведение "голосования": если из 5 критериев большинство "высказывается" за отклонение гипотезы, то итоговое решение - отвергнуть ее, в противном случае - принять. Однако в таком подходе нет ничего принципиально нового, просто к уже имеющимся критериям добавляются их комбинации - очередные варианты критериев, тем или иным образом выделяющие критические области в пространствах возможных значений результатов измерений, т.е. попросту увеличивается число рассматриваемых критериев.

Итак, имеется некоторая совокупность критериев. У каждого - свой набор значений уровней значимости и мощностей на возможных альтернативах. Математическая статистика демонстрирует в этой ситуации виртуозную математическую технику для анализа частных случаев и полную беспомощность при выдаче практических рекомендаций. Так, оказывается, что практически каждый из известных критериев является оптимальным в том или ином смысле для какого-то набора нулевых гипотез и альтернатив. Математики изучают асимптотическую эффективность в разных смыслах - по Питмену, по Бахадуру и т.д., но - для узкого класса альтернативных гипотез, обычно для альтернативы сдвига. При попытке переноса асимптотических результатов на конечные объемы выборок возникают новые нерешенные проблемы, связанные, в частности, с численным оцениванием скорости сходимости (см. подраздел 1.4.7). В целом эта область математической статистики может активно развиваться еще многие десятилетия, выдавая "на гора" превосходные теоремы (которые могут послужить основанием для защит кандидатских и докторских диссертаций, выборов в академики РАН и т.д.), но не давая ничего практике. Хорошо бы, чтобы этот пессимистический прогноз не вполне оправдался!

С точки зрения прикладной статистики необходимо изучать проблему выбора критерия проверки однородности двух независимых выборок. Такое изучение было проведено, в том числе методом статистических испытаний, и в результате был получен вывод о том, что наиболее целесообразно применять критерий Лемана - Розенблатта типа омега-квадрат (см. главу 3.1).

В литературе по прикладным статистическим методам, как справедливо замечает С.Г. Корнилов в работе [1], имеется масса ошибочных рекомендаций. Чего стоят хотя бы принципиально неверные государственные стандарты СССР по статистическим методам, а также соответствующие им стандарты СЭВ и ИСО, т.е. Международной организации по стандартизации (о них см. главу 13 учебника [13], а также статью [21]). Особо выделяются своим количеством ошибочные рекомендации по применению критерия типа Колмогорова для проверки нормальности. Ошибки есть и в научных статьях, и в нормативных документах (государственных стандартах), и в методических разработках, и даже в вузовских учебниках. К сожалению, нет способа оградить инженера и научного работника, экономиста и менеджера, нуждающихся в применении статистических методов, от литературных источников и нормативно-технических и инструктивно-методических документов с ошибками, неточностями и погрешностями. Единственный способ - либо постоянно поддерживать профессиональные контакты с квалифицированными специалистами по прикладной статистике, либо самому стать таким специалистом.

Как оценить достигаемый уровень значимости конкретного критерия, предусматривающего повторные проверки? Сразу ясно, что в большинстве случаев никакая современная теория математической статистики не поможет. Остается использовать современные компьютеры. Методика статистического моделирования может стать ежедневным рабочим инструментом специалиста, занимающегося применением методов анализа данных. Для этого она должна быть реализована в виде соответствующей диалоговой программной системы. Современные

персональные компьютеры позволяют проводить статистическое моделирование весьма быстро (за доли секунд). Можно использовать различные модификации бутстрепа - одного из вариантов применения статистического моделирования (см. [13]).

Проведенное обсуждение показывает, как много нерешенных проблем стоит перед специалистом, занимающимся, казалось бы, рутинным применением стандартных статистических процедур. Прикладная статистика - молодая наука, ее основные проблемы, по нашему мнению, еще не до конца решены. Много работы как в сравнительно новых областях, например, в анализе нечисловых и интервальных данных, так и в классических.

## Литература

1. Крамер Г. Математические методы статистики / Пер. с англ. / 2-е изд. - М.: Мир, 1975. – 648 с.
2. Орлов А.И. Метод моментов проверки согласия с параметрическим семейством распределений. – Журнал «Заводская лаборатория». 1989. Т.55. No.10. С.90-93.
3. Большев Л.Н. Избранные труды. Теория вероятностей и математическая статистика. – М.: Наука, 1987. – 286 с.
4. Тюрин Ю.Н. Исследования по непараметрической статистике: Непараметрические методы и линейная модель. Автореф. дисс. докт. физ.-мат. наук. – М.: МГУ, 1985. – 33 с.
5. Мартынов Г.В. Критерии омега-квадрат. – М.: Наука, 1978. – 80 с.
6. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983. - 416 с.
7. Артемьев Б.Г., Голубов С.М. Справочное пособие для работников метрологических служб.- М.: Изд-во стандартов, 1982. - 280 с.
8. Колмогоров А.Н., Фомин С.В. Элементы теории функций и функционального анализа. Учебник. – М.: Наука, 1972. – 496 с.
9. Орлов А.И. Асимптотическое поведение статистик интегрального типа. – Журнал «Доклады АН СССР». 1974. Т.219. No. 4. С. 808-811.
10. Орлов А.И. Асимптотическое поведение статистик интегрального типа. – В сб.: Вероятностные процессы и их приложения. Межвузовский сборник научных трудов. - М.: МИЭМ, 1989. С.118-123.
11. Гнеденко Б.В. Курс теории вероятностей: Учебник. 7-е изд., исправл. - М.: Эдиториал УРСС, 2001. 320 с.
12. Лоэв М. Теория вероятностей. – М.: ИЛ, 1962. – 720 с.
13. Орлов А.И. Эконометрика. Учебник для вузов. Изд. 2-е, исправленное и дополненное. - М.: Изд-во "Экзамен", 2003. – 576 с.
14. Уилкс С. Математическая статистика. - М.: Наука, 1967. - 632 с.
15. Кендалл М.Дж., Стьюарт А. Статистические выводы и связи. - М.: Наука, 1973. - 900 с.
16. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука, 1979. - 296 с.
17. Орлов А.И. Некоторые вероятностные вопросы теории классификации. – В сб.: Прикладная статистика. Ученые записки по статистике, т.45. - М.: Наука, 1983. С.166-179.
18. Корнилов С.Г. Накопление ошибки первого рода при повторной проверке статистических гипотез. Регламент повторных проверок. // Заводская лаборатория. 1996. Т.62. No.5. С. 45-51.
19. Камень Ю.Э., Камень Я.Э., Орлов А.И. Реальные и номинальные уровни значимости в задачах проверки статистических гипотез. // Заводская лаборатория. 1986. Т.52. No.12. С.55-57.
20. Налимов В.В. Применение математической статистики при анализе вещества. - М.: Физматгиз, 1960. - 430 с.
21. Орлов А.И. Сертификация и статистические методы (обобщающая статья). // Заводская лаборатория. - 1997. - Т.63. - No.3. - С.55-62.

## Контрольные вопросы

1. Сколько выборочных моментов необходимо использовать для проверки согласия с двухпараметрическим семейством функций распределения?
2. Почему методы отбраковки резко выделяющихся результатов наблюдений, основанные на предположении нормальности, нельзя считать научно обоснованными?
3. Какую роль играет условие интегрируемости по Риману-Стилтьесу в предельной теории статистик интегрального типа?
4. Как проверяют независимость альтернативных признаков с помощью таблиц 2Ч2?
5. Как влияет предварительное выделение однородных групп на проведение регрессионного анализа?
6. Как повлияет проверка однородности двух совокупностей (с помощью критерия Лемана – Розенблатта) на последующую оценку дисперсии по объединенной выборке (в случае подтверждения однородности)?

**Темы докладов, рефератов, исследовательских работ**

1. На основе метода моментов разработайте критерий согласия с семейством экспоненциальных распределений.
2. Методы отбраковки выбросов и их анализ с точки зрения теории устойчивости статистических процедур.
3. С помощью метода аппроксимации ступенчатыми функциями найдите асимптотическое распределение статистики Колмогорова.
4. Статистический приемочный контроль по альтернативным признакам.
5. Асимптотика Колмогорова в задачах прикладной статистики.
6. Проблема «стыковки» алгоритмов в технологиях обработки статистических данных.
7. Статистическая теория множественных проверок гипотез о разладке с помощью независимых датчиков.

# Часть 3. Методы прикладной статистики

## 3.1. Статистический анализ числовых величин

### 3.1.1. Оценивание основных характеристик распределения

Одна из основных задач прикладной статистики – оценивание по выборочным данным характеристик генеральной совокупности, таких, как математическое ожидание, медиана, дисперсия, среднее квадратическое отклонение, коэффициент вариации. Точечные оценки строятся очевидным образом – используют выборочные аналоги теоретических характеристик. Для получения интервальных оценок приходится использовать асимптотическую нормальность выборочных моментов и функций от них.

Пусть исходные данные – это выборка $x_1, x_2, \dots, x_n$, где $n$ – объем выборки. Выборочные значения $x_1, x_2, \dots, x_n$ рассматриваются как реализации независимых одинаково распределенных случайных величин $X_1, X_2, \dots, X_n$ с общей функцией распределения $F(x) = P(X_i < x), i = 1,2, \dots, n$. Поскольку функция распределения произвольна (с точностью до условий регулярности типа существования моментов), то рассматриваемые задачи доверительного оценивания характеристик распределения являются *непараметрическими*. Существование моментов является скорее математическим ограничением, чем реальным, поскольку практически все реальные статистические данные финитны (т.е. ограничены сверху и снизу, например, шкалой прибора).

В расчетах будут использоваться выборочное среднее арифметическое

$$\overline{X} = (X_1 + X_2 + \dots + X_n) / n,$$

выборочная дисперсия

$$s_0^2 = \{ (X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \dots + (X_n - \overline{X})^2 \} / (n-1)$$

и некоторые другие выборочные характеристики, которые мы введем позже.

**Точечное и интервальное оценивание математического ожидания.** Точечной оценкой для математического ожидания в силу закона больших чисел является выборочное среднее арифметическое $\overline{X}$. В некоторых случаях могут быть использованы и другие оценки. Например, если известно, что распределение симметрично относительно своего центра, то центр распределения является не только математическим ожиданием, но и медианой, а потому для его оценки можно использовать выборочную медиану.

Нижняя доверительная граница для математического ожидания имеет вид

$$\overline{X} - U(p)\, s_0 / n^{1/2},$$

где:

$\overline{X}$ – выборочное среднее арифметическое,

$p$ – доверительная вероятность (истинное значение математического ожидания находится между нижней доверительной границей и верхней доверительной границей с вероятностью, равной доверительной);

$U(p)$ – число, заданное равенством $\Phi(U(p)) = (1+ p)/2$, где $\Phi(x)$ – функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Например, при $p$ = 95% (т.е. при $p$ = 0,95) имеем $U(p)$ = 1,96. Функция $U(p)$ имеется в большинстве литературных источников по теории вероятностей и математической статистике (см., например, [1]);

$s_0$ – выборочное среднее квадратическое отклонение (квадратный корень из описанной выше выборочной дисперсии).

Верхняя доверительная граница для математического ожидания имеет вид

$$\overline{X} + U(p)\, s_0 / n^{1/2}.$$

Выражения для верхней и нижней доверительных границ получены с помощью Центральной Предельной Теоремы теории вероятностей, теоремы о наследовании сходимости и

других результатов главы 1.4. Они являются асимптотическими, т.е. становятся тем точнее, чем больше объем выборки. В частности, вероятность попадания истинного значения математического ожидания между нижней и верхней доверительными границами асимптотически приближается к доверительной вероятности, но, вообще говоря, может отличаться от нее. Это – недостатки непараметрического подхода. Достоинством же является то, что его можно применять всегда, когда случайная величина имеет математическое ожидание и дисперсию, что в силу финитности (ограниченности шкал) имеет быть практически всегда в реальных ситуациях.

Интересно сопоставить с параметрическим подходом. Обычно в таких случаях предполагают нормальность результатов наблюдений (которой, как уже было обосновано в главе 2.1, практически никогда нет). Тогда формулы для нижней и верхней доверительных границ для математического ожидания имеют похожий вид, только вместо *U(p)* стоят квантили распределения Стьюдента (а не нормального распределения, как в приведенных выше формулах), соответствующие объему выборки. Как известно, при росте объема выборки квантили распределения Стьюдента сходятся к соответствующим квантилям стандартного нормального распределения, так что при больших объемах выборок оба подхода дают близкие результаты. Отметим, что классические доверительные интервалы несколько длиннее, поскольку квантили распределения Стьюдента больше квантилей стандартного нормального распределения, хотя это различие, на наш взгляд, и невелико.

*Пример 1.* Рассмотрим данные о наработке резцов до отказа (раздел 2.2.1, табл.2). Для них выборочное среднее арифметическое $\overline{X}$ = 57,88 (это и есть точечная оценка для математического ожидания), выборочная дисперсия $s_0^2$ = 663,00, объем выборки $n$ = 50. Следовательно, выборочное среднее квадратическое отклонение $s_0 = \sqrt{663{,}00} = 25{,}75$ и согласно приведенным выше формулам при доверительной вероятности $p$ = 0,95 нижняя доверительная граница для математического ожидания такова:

$$57{,}88 - 1{,}96 \text{ Ч } 25{,}75 / \sqrt{50} = 57{,}88 - 7{,}14 = 50{,}74,$$

а верхняя доверительная граница есть 57,88 + 7,14 = 65,02.

Если заранее известно, что результаты наблюдения имеют нормальное распределение, то нижняя и верхняя доверительная границы для математического ожидания определяются по формулам

$$\overline{X} - t(p, n\text{-}1)\, s_0 / \sqrt{n}\,, \quad \overline{X} + t(p, n\text{-}1)\, s_0 / \sqrt{n}$$

соответственно. Эти формулы отличаются от предыдущих тем, что квантиль нормального распределения *U(p)* заменен на аналогичный квантиль распределения Стьюдента с ($n$ – 1) степенью свободы. Другими словами, *t(p, n-1)* – это число, заданное равенством $ST_{n-1}(p) = = (1+p)/2$, где $ST_{n-1}(x)$ – функция распределения Стьюдента с ($n$ – 1) степенью свободы.

Для доверительной вероятности $p$ = 0,95 при объеме выборки $n$ = 50 согласно [1] имеем *t(p, n-1)* = 2,0096. Следовательно, нижняя доверительная граница для математического ожидания такова:

$$57{,}88 - 2{,}0096 \text{ Ч } 25{,}75 / \sqrt{50} = 57{,}88 - 7{,}32 = 50{,}56,$$

а верхняя доверительная граница есть 57,88 + 7,32 = 65,20. Таким образом, длина доверительного интервала увеличилась с 14,28 до 14,64, т.е. на 2,5%.

Отметим, что рассматриваемые данные согласуются с гамма-распределением (см. раздел 2.3.1), а не с нормальным распределением, поэтому использование распределения Стьюдента для получения доверительных границ явно некорректно.

Иногда рекомендуют сначала проверить нормальность результатов наблюдений, а потом, в случае принятия гипотезы нормальности, рассчитывать доверительные границы с использованием квантилей распределения Стьюдента. Однако проверка нормальности - более сложная статистическая процедура, чем оценивание математического ожидания. Кроме того, применение одной статистической процедуры, как правило, нарушает предпосылки следующей процедуры, в частности, независимость результатов наблюдений (см. раздел 2.3.5). Поэтому цепочка

статистических процедур, следующих друг за другом, как правило, образует статистическую технологию, свойства которой неизвестны на современном уровне развития прикладной статистики.

Из сказанного вытекает, что только непараметрическую статистическую процедуру, основанную на асимптотических результатах главы 1.4, следует применять для анализа реальных данных. Как правило, встречающиеся на практике распределения не являются нормальными (см. раздел 2.1.1), а потому применение квантилей распределения Стьюдента неправомерно.

**Точечное и интервальное оценивание медианы.** Точечной оценкой для медианы является выборочная медиана.

*Пример 2.* Для данных о наработке резцов до отказа объем выборки – четное число, поэтому выборочной медианой является полусумма 25-го и 26-го членов вариационного ряда, т.е. (56 + 56,5)/2 = 56,25.

Чтобы построить доверительные границы для медианы, по доверительной вероятности $p$ находят $U(p)$, как разъяснено выше. Затем вычисляют натуральное число

$$C(p) = [n/2 - U(p)n^{1/2}/2],$$

где [.] – знак целой части числа. Нижняя доверительная граница для медианы имеет вид

$$X(C(p)),$$

где $X(i)$ – член вариационного ряда с номером $i$, построенного по исходной выборке (т.е. $i$-я порядковая статистика). Верхняя доверительная граница для медианы имеет вид

$$X(n + 1 - C(p)).$$

Теоретическое основание для приведенных доверительных границ содержится в литературе по порядковым статистикам (см., например, монографию [2, с.68]).

*Пример 3.* Для данных о наработке резцов до отказа $n = 50$. Рассмотрим как обычно, доверительную вероятность $p = 0{,}95$. Тогда

$$C(p) = [50/2 - 1{,}96\sqrt{50}/2] = [18{,}07] = 18.$$

Следовательно, нижней доверительной границей является $X(18) = 47{,}5$, а верхней доверительной границей $X(50 + 1 - 18) = X(33) = 61{,}5$.

Поскольку в случае нормального распределения медиана совпадает с математическим ожиданием, то каких-либо специальных способов ее оценивания в классическом случае нет.

**Точечное и интервальное оценивание дисперсии.** Точечной оценкой дисперсии является выборочная дисперсия $s_0^2$. Эта оценка является несмещенной и состоятельной. Доверительные границы находятся с помощью величины

$$d^2 = (m_4 - ((n-1)/n)^4 s_0^4)/n,$$

где $m_4$ - выборочный четвертый центральный момент, т.е.

$$m_4 = \{(X_1 - \overline{X})^4 + (X_2 - \overline{X})^4 + \ldots + (X_n - \overline{X})^4\}/n.$$

Нижняя доверительная граница для дисперсии случайной величины имеет вид

$$s_0^2 - U(p)d,$$

где:

$s_0^2$ – выборочная дисперсия,

$U(p)$ – квантиль нормального распределения порядка $(1+p)/2$ (как и раньше),

$d$ – положительный квадратный корень из величины $d^2$, введенной выше.

Верхняя доверительная граница для дисперсии случайной величины имеет вид

$$s_0^2 + U(p)d,$$

где все составляющие имеют тот же смысл, что и выше.

При выводе приведенных соотношений используется асимптотическая нормальность выборочной дисперсии, установленная, например, в учебнике по математической статистике [3, с.419]. Соответственно доверительный интервал является непараметрическим и асимптотическим. В классическом случае точечная оценка имеет тот же вид, а вот доверительные границы находят с

помощью квантилей распределения хи-квадрат с числом степеней свободы, на 1 меньшим объема выборки. Отметим, что в случае нормального распределения четвертый момент в 3 раза больше квадрата дисперсии, а потому можно оценить $d^2$ как $(2\ s_0^4)\ /\ n$ . Это дает быстрый способ для интервальной оценки дисперсии в нормальном случае.

*Пример 4.* Для данных о наработке резцов до отказа объем выборки $n$ = 50, выборочная дисперсия $s_0^2$ = 663,00, четвертый выборочный момент $m_4$ = 1702050,71. Поэтому

$$d^2 = (1702050{,}71 - ((50 - 1)/50)^4\, 663{,}00^2)/50 = 25932{,}13.$$

Тогда $d$ = 161,03. Для доверительной вероятности $p$ =0,95 нижняя доверительная граница для дисперсии случайной величины такова:

$$663{,}00 - 1{,}96 \text{Ч} 161{,}03 = 663{,}00 - 315{,}63 = 347{,}37,$$

а верхняя доверительная граница для дисперсии есть 663,00 + 315,63 = 978,63.

*Пример 5.* В случае нормального распределения с целью быстрого получения доверительного интервала величина $d^2$ оценивается как

$$(2\ s_0^4)\ /\ n = (2 \text{ Ч } 663{,}00^2)\ /\ 50 = 17582{,}76,$$

а потому $d$ = 132,6. Для доверительной вероятности $p$ =0,95 нижняя доверительная граница для дисперсии заменяется на

$$663{,}00 - 1{,}96 \text{Ч} 132{,}6 = 663{,}00 - 259{,}90 = 403{,}10,$$

а верхняя доверительная граница – на 663,00 + 259,90 = 922,9.

Сужение границ для дисперсии вполне естественно. Данные о наработке резцов до предельного состояния (т.е. до отказа) соответствуют гамма-распределению, а это распределение является асимметричным, с «тяжелым» правым «хвостом». Последнее означает, что плотность убывает заметно медленнее, чем для нормального распределения. Как следствие, четвертый момент заметно больше, чем для нормального распределения с теми же математическим ожиданием и дисперсией. А потому больше и параметр $d$. Из проведенных расчетов видно, что использование алгоритмов расчетов, соответствующих нормальному распределению, в ситуации, когда распределение результатов наблюдений отлично от нормального, может привести к заметно искаженным выводам.

*Пример 6.* В классическом случае нормального распределения исходят из того, что величина $(n - 1)\ s_0^2/y^2$ имеет распределение хи-квадрат с $(n - 1)$ степенью свободы. Для доверительной вероятности $p$ =0,95 следует рассмотреть неравенство

$$31{,}555 < (n - 1)\ s_0^2/y^2 < 70{,}222,$$

справедливое с вероятностью 0,95, поскольку

$$F(31{,}555) = 0{,}025,\ F(70{,}222) = 0{,}975,$$

где $F(x)$ – функция хи-квадрат распределения с 49 степенями свободы. Следовательно, нижняя доверительная граница для дисперсии нормально распределенной случайной величины такова:

$$(n - 1)\ s_0^2/70{,}222 = (49 \text{Ч} 663{,}00)/70{,}222 = 462{,}63,$$

а верхняя доверительная граница есть

$$(n - 1)\ s_0^2/31{,}555 = (49 \text{Ч} 663{,}00)/\ 31{,}555 = 1029{,}54.$$

Полученный доверительный интервал не является симметричным относительно точечной оценки. Нижняя доверительная граница больше, чем в примерах 4 и 5, но и верхняя доверительная граница тоже больше. Несимметричность доверительного интервала в примере 6 приводит к тому, что его трудно сопоставить с симметричными интервалами примеров 4 и 5. Что же касается практических рекомендаций, то они однозначны: поскольку обычно нет основания считать данные имеющими нормальное распределение, то при анализе реальных данных надо пользоваться непараметрическими методами, не предполагающими нормальность, т.е. методами, примененными в примере 4.

**Точечное и интервальное оценивание среднего квадратического отклонения.** Точечной оценкой является выборочное среднее квадратическое отклонение, т.е. неотрицательный

квадратный корень из выборочной дисперсии. Дисперсия рассматриваемой случайной величины - выборочного среднего квадратического отклонения $s_0$ – оценивается как дробь

$$d^2 / (4\ s_0^2\ ).$$

Нижняя доверительная граница для среднего квадратического отклонения исходной случайной величины имеет вид

$$s_0 - U(p)d / (2\ s_0)\ ,$$

где:

$s_0^2$ – выборочная дисперсия,

$U(p)$ – квантиль нормального распределения порядка $(1+p)/2$ (как и раньше),

$d$ – положительный квадратный корень из величины $d^2$, введенной выше при оценивании дисперсии.

Верхняя доверительная граница для среднего квадратического отклонения исходной случайной величины имеет вид

$$s_0\ + U(p)d / (2\ s_0)\ ,$$

где все составляющие имеют тот же смысл, что и выше.

*Пример 7.* Для данных о наработке резцов до отказа точечной оценкой для среднего квадратического отклонения является $s_0 = \sqrt{663{,}00} = 25{,}75$ . При доверительной вероятности $p$ = 0,95 нижняя доверительная граница такова:

$$25{,}75 – 1{,}96Ч161{,}03 / (2Ч25{,}75) = 25{,}75 – 6{,}13 = 19{,}62.$$

Соответственно верхняя доверительная граница симметрична нижней относительно точечной оценки и равна = 25,75 + 6,13 = 31,88.

Правила интервального оценивания для среднего квадратического отклонения получены из аналогичных правил для оценивания дисперсии с помощью метода линеаризации (см. главу 1.4 или, например, [4, п.2.4]). Как и раньше, доверительный интервал является симметричным, непараметрическим и асимптотическим.

Поскольку среднее квадратическое отклонение – это квадратный корень их дисперсии, то доверительные границы можно получить, извлекая квадратные корни из одноименных границ для дисперсии.

*Пример 8*. Для данных о наработке резцов до отказа при доверительной вероятности $p$ = 0,95 согласно примеру 4 доверительный интервал для дисперсии – это [347,37; 978,63]. Извлекая квадратные корни, получаем доверительный интервал [18,64; 31,28] для среднего квадратического отклонения, соответствующий тому же значению доверительной вероятности. Он не является симметричным относительно точечной оценки. Его длина 12,64 несколько больше длины симметричного доверительного интервала 12,26 в примере 7.

Классический подход, основанный на гипотезе нормальности распределения результатов наблюдения, связан с использованием распределения хи-квадрат и сводится к извлечению квадратных корней из доверительных границ для дисперсии.

*Пример 9*. Применяя формально классический подход к данным о наработке резцов до отказа, исходим из доверительного интервала для дисперсии [462,63; 1029,54], соответствующего доверительной вероятности $p$ = 0,95. Извлекая квадратные корни, находим доверительный интервал для среднего квадратического отклонения [21,51; 32,09]. Как и следовало ожидать, длина этого несимметричного интервала 10,58 меньше длины непараметрического доверительного интервала.

**Точечное и интервальное оценивание коэффициента вариации.** Коэффициент вариации $V$ = у / $M(X)$ широко используется при анализе конкретных технических, экономических, социологических, медицинских и иных данных (поскольку они, как правило, положительны), но не очень популярен среди теоретиков в области математической статистики. Точечной оценкой теоретического коэффициента вариации $V$ является выборочный коэффициент вариации

$$V_n = s_0 / \overline{X}$$

Дисперсия выборочный коэффициент вариации состоятельно оценивается с помощью вспомогательной величины

$$D^2 = (V_n^4 - V_n^2 / 4 + m_4 / (4 s_0^2 \overline{X}^2) - m_3 / \overline{X}^3) / n ,$$

где:

$\overline{X}$ – выборочное среднее арифметическое,

$s_0^2$ – выборочная дисперсия,

$m_3$ - выборочный третий центральный момент, т.е.

$$m_3 = \{ (X_1 - \overline{X})^3 + (X_2 - \overline{X})^3 + ... + (X_n - \overline{X})^3 \} / n ,$$

$m_4$ - выборочный четвертый центральный момент (см. выше),

$V_n$ – выборочный коэффициент вариации,

$n$ - объем выборки.

Нижняя доверительная граница для (теоретического) коэффициента вариации исходной случайной величины имеет вид

$$V_n - U(p) D,$$

где:

$V_n$ – выборочный коэффициент вариации,

$U(p)$ – квантиль нормального распределения порядка $(1+p)/2$ (как и ранее),

$D$ – положительный квадратный корень из величины $D^2$, введенной выше.

Верхняя доверительная граница для (теоретического) коэффициента вариации исходной случайной величины имеет вид

$$V_n + U(p) D,$$

где все составляющие имеют тот же смысл, что и выше.

Как и в предыдущих случаях, доверительный интервал является непараметрическим и асимптотическим. Он получен в результате применения специальной технологии вывода асимптотических соотношений прикладной статистики (см. главу 1.4). Напомним, что эта технология в качестве первого шага использует многомерную центральную предельную теорему, примененную к сумме векторов, координаты которых – степени исходных случайных величин. Второй шаг – преобразование предельного многомерного нормального вектора с целью получения интересующего исследователя вектора. При этом используются соображения линеаризации и отбрасываются бесконечно малые величины. Третий шаг – строгое обоснование полученных результатов на стандартном для асимптотических математико-статистических рассуждений уровне. При этом обычно приходится использовать необходимые и достаточные условия наследования сходимости, полученные в монографии [4, п.2.4]. Именно таким образом были получены приведенные выше результаты для выборочного коэффициента вариации. Формулы оказались существенно более сложными, чем в предыдущих случаях. Это объясняется тем, что выборочный коэффициент вариации - функция двух выборочных моментов, а ранее рассматривались либо выборочные моменты поодиночке, либо функция от одного выборочного момента - выборочной дисперсии.

*Пример 10*. Для данных о наработке резцов до отказа выборочное среднее арифметическое $\overline{X}$ = 57,88, выборочная дисперсия $s_0^2$ = 663,00, выборочное среднее квадратическое отклонение $s_0$ = 25,75, выборочный третий центральный момент $m_3$ = 14927,91, выборочный четвертый центральный момент $m_4$ = 1702050,71. Следовательно, выборочный коэффициент вариации таков:

$$V_n = 25,75 / 57,88 = 0,4449.$$

Рассчитаем значение вспомогательной величины

$$D^2 = ((0,4449)^4 - (0,4449)^2 / 4 + 1702050,71 / (4Ч663,00Ч(57,88)^2) -$$
$$- 14927,91 / (57,88)^3 ) / 50 = (0,0392 - 0,0495 + 0,1916 - 0,0770)\backslash 50 =$$
$$= 0,1043/50 = 0,002086.$$

Следовательно, $D$ = 0,04567. При доверительной вероятности p = 0,95 нижняя доверительная граница для теоретического коэффициента вариации имеет вид

$$0{,}4449 – 1{,}96Ч0{,}04567 = 0{,}4449 – 0{,}0895 = 0{,}3554,$$

а верхняя доверительная граница такова:

$$0{,}4449 + 0{,}0895 = 0{,}5344.$$

Среди классических результатов математической статистики, основанных на гипотезе нормальности результатов наблюдений, нет методов построения доверительных границ для коэффициента вариации, поскольку задача построения таких границ не выражается в терминах обычно используемых распределений, например, распределений Стьюдента и хи-квадрат.

Примеры применения доверительных границ для коэффициентов вариации при решении прикладных задач приведены, например, в работе [5], посвященной анализу технических характеристик и показателей качества.

### 3.1.2. Методы проверки однородности характеристик двух независимых выборок

В прикладных исследованиях часто возникает необходимость выяснить, различаются ли генеральные совокупности, из которых взяты две независимые выборки. Например, надо выяснить, влияет ли способ упаковки подшипников на их потребительские качества через год после хранения. Или: отличается ли потребительское поведение мужчин и женщин. Если отличается – рекламные ролики и плакаты надо делать отдельно для мужчин и отдельно для женщин. Если нет – рекламная кампания может быть единой.

В математико-статистических терминах постановка задачи такова: имеются две выборки $x_1, x_2,...,x_m$ и $y_1, y_2,...,y_n$ (т. е. наборы из $m$ и $n$ действительных чисел), требуется проверить их однородность. Термин «однородность» уточняется ниже.

Противоположным понятием является «различие». Можно переформулировать задачу: требуется проверить, есть ли различие между выборками. Если различия нет, то для дальнейшего изучения две рассматриваемые выборки часто объединяют в одну.

Например, в маркетинге важно выделить сегменты потребительского рынка. Если установлена однородность двух выборок, то возможно объединение сегментов, из которых они взяты, в один. В дальнейшем это позволит осуществлять по отношению к ним одинаковую маркетинговую политику (проводить одни и те же рекламные мероприятия и т.п.). Если же установлено различие, то поведение потребителей в двух сегментах различно, объединять эти сегменты нельзя, и могут понадобиться различные маркетинговые стратегии, своя для каждого из этих сегментов.

**Традиционный метод проверки однородности (критерий Стьюдента).** Для дальнейшего критического разбора опишем традиционный статистический метод проверки однородности. Вычисляют выборочные средние арифметические в каждой выборке

$$\bar{x} = \frac{1}{m}\sum_{1\le i\le m} x_i, \qquad y = \frac{1}{n}\sum_{1\le i\le n} y_i \ ,$$

затем выборочные дисперсии

$$s_x^2 = \frac{1}{m-1}\sum_{1\le i\le m}(x_i - \bar{x})^2 \ , \quad s_y^2 = \frac{1}{n-1}\sum_{1\le i\le n}(y_i - \bar{y})^2$$

и статистику Стьюдента $t$, на основе которой принимают решение,

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{(m-1)s_x^2 + (n-1)s_y^2}}\sqrt{\frac{mn(m+n-2)}{m+n}} \quad . \quad (1)$$

По заданному уровню значимости $\alpha$ и числу степеней свободы $(m+n-2)$ из таблиц распределения Стьюдента находят критическое значение $t_{кр}$. Если $|t|>t_{кр}$, то гипотезу однородности (отсутствия различия) отклоняют, если же $|t|\leq t_{кр}$, то принимают. (При односторонних альтернативных гипотезах вместо условия $|t|>t_{кр}$ проверяют, что $t>t_{кр}$; эту постановку рассматривать не будем, так как в ней нет принципиальных отличий от обсуждаемой здесь.)

Рассмотрим условия применимости традиционного метода проверки однородности, основанного на использовании статистики $t$ Стьюдента, а также укажем более современные методы.

**Вероятностная модель порождения данных.** Для обоснованного применения эконометрических методов необходимо прежде всего построить и обосновать вероятностную модель порождения данных. При проверке однородности двух выборок общепринята модель, в которой $x_1, x_2,...,x_m$ рассматриваются как результаты $m$ независимых наблюдений некоторой случайной величины $X$ с функцией распределения $F(x)$, неизвестной статистику, а $y_1, y_2,...,y_n$ - как результаты $n$ независимых наблюдений, вообще говоря, другой случайной величины $Y$ с функцией распределения $G(x)$, также неизвестной статистику. Предполагается также, что наблюдения в одной выборке не зависят от наблюдений в другой, поэтому выборки и называют независимыми.

Возможность применения модели в конкретной реальной ситуации требует обоснования. Независимость и одинаковая распределенность результатов наблюдений, входящих в выборку, могут быть установлены или исходя из методики проведения конкретных наблюдений, или путем проверки статистических гипотез независимости и одинаковой распределенности с помощью соответствующих критериев [1].

Если проведено $(m+n)$ измерений объемов продаж в $(m+n)$ торговых точках, то описанную выше модель, как правило, можно применять. Если же, например, $x_i$ и $y_i$ - объемы продаж одного и того же товара до и после определенного рекламного воздействия, то рассматриваемую модель применять нельзя. В последнем случае используют модель связанных выборок. В ней обычно строят новую выборку $z_i = x_i - y_i$ и используют статистические методы анализа одной выборки, а не двух. Методы проверка однородности для связанных выборок рассматривается в разделе 3.1.6.

При дальнейшем изложении принимаем описанную выше вероятностную модель двух выборок.

**Уточнения понятия однородности.** Понятие «однородность», т. е. «отсутствие различия», может быть формализовано в терминах вероятностной модели различными способами.

Наивысшая степень однородности достигается, если обе выборки взяты из одной и той же генеральной совокупности, т. е. справедлива нулевая гипотеза

$$H_0 : F(x)=G(x) \text{ при всех } x.$$

Отсутствие однородности означает, что верна альтернативная гипотеза, согласно которой

$$H_1 : F(x_0)\neq G(x_0)$$

хотя бы при одном значении аргумента $x_0$. Если гипотеза $H_0$ принята, то выборки можно объединить в одну, если нет - то нельзя.

В некоторых случаях целесообразно проверять не совпадение функций распределения, а совпадение некоторых характеристик случайных величин $X$ и $Y$ - математических ожиданий, медиан, дисперсий, коэффициентов вариации и др. Например, однородность математических ожиданий означает, что справедлива гипотеза

$$H'_0 : M(X)=M(Y),$$

где $M(X)$ и $M(Y)$ - математические ожидания случайных величин $X$ и $Y$, результаты наблюдений над которыми составляют первую и вторую выборки соответственно. Доказательство различия между выборками в рассматриваемом случае - это доказательство справедливости альтернативной гипотезы

$$H'_1 : M(X) \neq M(Y .$$

Если гипотеза $H_0$ верна, то и гипотеза $H'_0$ верна, но из справедливости $H'_0$, вообще говоря, не следует справедливость $H_0$. Математические ожидания могут совпадать для различающихся

между собой функций распределения. В частности, если в результате обработки выборочных данных принята гипотеза $H'_0$, то отсюда *не следует,* что две выборки можно объединить в одну. Однако в ряде ситуаций целесообразна проверка именно гипотезы $H'_0$. Например, пусть функция спроса на определенный товар или услугу оценивается путем опроса потребителей (первая выборка) или с помощью данных о продажах (вторая выборка). Тогда маркетологу важно проверить гипотезу об отсутствии систематических расхождений результатов этих двух методов, т.е. гипотезу о равенстве математических ожиданий. Другой пример – из производственного менеджмента. Пусть изучается эффективность управления бригадами рабочих на предприятии с помощью двух организационных схем, результаты наблюдения - объем производства на одного члена бригады, а показатель эффективности организационной схемы - средний (по предприятию) объем производства на одного рабочего. Тогда для сравнения эффективности препаратов достаточно проверить гипотезу $H'_0$.

**Классические условия применимости критерия Стьюдента.** Согласно математико-статистической теории должны быть выполнены два классических условия применимости критерия Стьюдента, основанного на использовании статистики $t$, заданной формулой (1):

а) результаты наблюдений имеют нормальные распределения:

$$F(x)=N(x;\ m_1,\ \sigma_1^2),\ G(x)=N(x;\ m_2,\ \sigma_2^2)$$

с математическими ожиданиями $m_1$ и $m_2$ и дисперсиями $\sigma_1^2$ и $\sigma_2^2$ в первой и во второй выборках соответственно;

б) дисперсии результатов наблюдений в первой и второй выборках совпадают:

$$D(X)=\sigma_1^2=D(Y)=\sigma_2^2.$$

Если условия а) и б) выполнены, то нормальные распределения $F(x)$ и $G(x)$ отличаются только математическими ожиданиями, а поэтому *обе* гипотезы $H_0$ и $H'_0$ сводятся к гипотезе

$$H''_0 : m_1=m_2,\ ,$$

а *обе* альтернативные гипотезы $H_1$ и $H'_1$ сводятся к гипотезе

$$H''_1 : m_1\neq m_2,\ .$$

Если условия а) и б) выполнены, то статистика $t$ при справедливости $H''_0$ имеет распределение Стьюдента с $(m + n - 2)$ степенями свободы. Только в этом случае описанный выше традиционный метод обоснован безупречно. Если хотя бы одно из условий а) и б) не выполнено, то нет никаких оснований считать, что статистика $t$ имеет распределение Стьюдента, поэтому применение традиционного метода, строго говоря, не обосновано. Обсудим возможность проверки этих условий и последствия их нарушений.

**Имеют ли результаты наблюдений нормальное распределение?** Как показано в главе 2.1, априори нет оснований предполагать нормальность распределения результатов экономических, технико-экономических, технических, медицинских и иных наблюдений. Следовательно, нормальность надо проверять. Разработано много статистических критериев для проверки нормальности распределения результатов наблюдений [1]. Однако проверка нормальности - более сложная и трудоемкая статистическая процедура, чем проверка однородности (как с помощью статистики $t$ Стьюдента, так и с использованием непараметрических критериев, рассматриваемых ниже).

Для достаточно надежного установления нормальности требуется весьма большое число наблюдений. В главе 2.1 показано, что для того, чтобы гарантировать, что функция распределения результатов наблюдений отличается от некоторой нормальной не более чем на 0,01 (при любом значении аргумента), требуется порядка 2500 наблюдений. В большинстве технических, экономических, медицинских и иных исследований число наблюдений существенно меньше.

Как уже отмечалось, есть и одна общая причина отклонений от нормальности: любой результат наблюдения записывается конечным (обычно 2-5) количеством цифр, а с математической точки зрения вероятность такого события равна 0. Следовательно, в прикладной статистике распределение результатов наблюдений практически всегда более или менее отличается от нормального распределения.

**Последствия нарушения условия нормальности.** Если условие а) не выполнено, то распределение статистики $t$ не является распределением Стьюдента. Однако при справедливости $H'_0$ и условии б) распределение статистики $t$ при росте объемов выборок приближается к стандартному нормальному распределению $\Phi(x)=N(x;\ 0,\ 1)$. К этому же распределению приближается распределение Стьюдента при возрастании числа степеней свободы. Другими словами, несмотря на нарушение условия нормальности традиционный метод (критерий Стьюдента) можно использовать для проверки гипотезы $H'_0$ при больших объемах выборок. При этом вместо таблиц распределения Стьюдента достаточно пользоваться таблицами стандартного нормального распределения $\Phi(x)$.

Сформулированное в предыдущем абзаце утверждение справедливо для любых функций распределения $F(x)$ и $G(x)$ таких, что $M(X)=M(Y)$, $D(X)=D(Y)$ и выполнены некоторые внутриматематические условия, обычно считающиеся справедливыми в реальных задачах. Если же $M(X)\neq M(Y)$, то нетрудно вычислить, что при больших объемах выборок

$$P(t\leq x)\approx\Phi(x-a_{mn}), \qquad (2)$$

где

$$a_{mn}=\frac{\sqrt{mn}[M(X)-M(Y)]}{\sqrt{mD(X)+nD(Y)}}. \qquad (3)$$

Формулы (2) - (3) позволяют приближенно вычислять мощность $t$-критерия (точность возрастает при увеличении объемов выборок $m$ и $n$).

**О проверке условия равенства дисперсий.** Иногда условие б) вытекает из методики получения результатов наблюдений, например, когда с помощью одного и того же прибора или методики $m$ раз измеряют характеристику первого объекта и $n$ раз - второго, а параметры распределения погрешностей измерения при этом не меняются. Однако ясно, что в постановках большинства исследовательских и практических задач нет основании априори предполагать равенство дисперсий.

Целесообразно ли проверять равенство дисперсий статистическими методами, например, как это иногда предлагают, с помощью $F$-критерия Фишера? Этот критерий основан на нормальности распределений результатов наблюдений, от которой неизбежны отклонения (см. выше). Причем хорошо известно, что в отличие от $t$-критерия распределение $F$-критерия Фишера сильно меняется при малых отклонениях от нормальности [3]. Кроме того, $F$-критерий отвергает гипотезу $D(X)=D(Y)$ лишь при большом различии выборочных дисперсий. Так, для данных [1] о двух группах результатов химических анализов отношение выборочных дисперсий равно 1,95, т.е. существенно отличается от 1. Тем не менее гипотеза о равенстве теоретических дисперсий принимается на 1%-м уровне значимости. Следовательно, при проверке однородности применение $F$-критерия для предварительной проверки равенства дисперсий нецелесообразно.

Итак, в большинстве технических, экономических, медицинских и иных задач условие б) нельзя считать выполненным, а проверять его нецелесообразно.

**Последствия нарушения условия равенства дисперсий.** Если объемы выборок $m$ и $n$ велики, то можно показать, что распределение статистики $t$ описывается с помощью только математических ожиданий $M(X)$ и $M(Y)$, дисперсий $D(X)$, $D(Y)$ и отношения объемов выборок, а именно:

$$P(t\leq x)\approx\Phi(b_{mn}x-a_{mn}), \qquad (4)$$

где $a_{mn}$ определено формулой (3),

$$b_{mn}^{\ 2}=\frac{\lambda D(X)+D(Y)}{D(X)+\lambda D(Y)},\quad \lambda=\frac{m}{n}. \qquad (5)$$

Если $b_{mn}\neq 1$, то распределение статистики $t$ отличается от распределения, заданного формулой (2), полученной в предположении равенства дисперсий. Когда $b_{mn}=1$? В двух случаях - при $m=n$ и при $D(X)=D(Y)$. Таким образом, при больших и равных объемах выборок требовать выполнения условия б) нет необходимости. Кроме того, ясно, что если объемы выборок мало различаются, то

$b_{mn}$ близко к 1. Так, для данных [1] о двух группах результатов химических анализов имеем $b^*_{mn}$= 0,987, где $b^*_{mn}$ - оценка $b_{mn}$ , полученная заменой в формуле (5) теоретических дисперсий на выборочные.

**Область применимости традиционного метода проверки однородности с помощью критерия Стьюдента.** Подведем итоги рассмотрения *t*-критерия. Он позволяет проверять гипотезу $H'_0$ о равенстве математических ожиданий, но не гипотезу $H_0$ о том, что обе выборки взяты из одной и той же генеральной совокупности. Классические условия применимости критерия Стьюдента в подавляющем большинстве технических, экономических, медицинских и иных задач не выполнены. Тем не менее при больших и примерно равных объемах выборок его можно применять. При конечных объемах выборок традиционный метод носит неустранимо приближенный характер.

**Критерий Крамера-Уэлча равенства математических ожиданий.** Вместо критерия Стьюдента целесообразно для проверки $H'_0$ использовать критерий Крамера-Уэлча [6], основанный на статистике

$$T = \frac{\sqrt{mn}(\bar{x} - \bar{y})}{\sqrt{ns_x^2 + ms_y^2}}. \qquad (6)$$

Критерий Крамера-Уэлча имеет прозрачный смысл – разность выборочных средних арифметических для двух выборок делится на естественную оценку среднего квадратического отклонения этой разности. Естественность указанной оценки состоит в том, что неизвестные статистику дисперсии заменены их выборочными оценками. Из многомерной центральной предельной теоремы и из теорем о наследовании сходимости [4] вытекает (см. главу 1.4), что при росте объемов выборок распределение статистики *T* Крамера-Уэлча сходится к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1. Итак, при справедливости $H'_0$ и больших объемах выборок распределение статистики *T* приближается с помощью стандартного нормального распределения $\Phi(x)$, из таблиц которого следует брать критические значения.

При $m=n$, как следует из формул (1) и (6), $t=T$. При $m \neq n$ этого равенства нет. В частности, при $s_x^2$ в (1) стоит множитель ($m$ - 1), а в (6)- множитель $n$.

Если $M(X) \neq M(Y)$, то при больших объемах выборок

$$P(T \leq x) \approx \Phi(x - c_{mn}), \qquad (7)$$

где

$$c_{mn} = \frac{\sqrt{mn}[M(X) - M(Y)]}{\sqrt{nD(X) + mD(Y)}}. \qquad (8)$$

При $m=n$ или $D(X)=D(Y)$, согласно формулам (3) и (8), $a_{mn}=c_{mn}$ , в остальных случаях равенства нет.

Из асимптотической нормальности статистики *T*, формул (7) и (8) следует, что правило принятия решения для критерия Крамера-Уэлча выглядит так:

- если $|T| \leq \Phi(1 - \frac{\alpha}{2})$, то гипотеза однородности (равенства) математических ожиданий принимается на уровне значимости $\alpha$,
- если же $|T| > \Phi(1 - \frac{\alpha}{2})$, то гипотеза однородности (равенства) математических ожиданий отклоняется на уровне значимости $\alpha$.

В прикладной статистике наиболее часто применяется уровень значимости $\alpha = 0{,}05$. Тогда значение модуля статистики *T* Крамера-Уэлча надо сравнивать с граничным значением $\Phi(1 - \frac{\alpha}{2}) = 1{,}96$.

Из сказанного выше следует, что применение критерия Крамера-Уэлча не менее обосновано, чем применение критерия Стьюдента. Дополнительное преимущество - не требуется равенства дисперсий *D(X)=D(Y).* Распределение статистики *T* не является распределением Стьюдента, однако и распределение статистики *t,* как показано выше, не является таковым в реальных ситуациях.

Распределение статистики *T* при объемах выборок *m=n=6, 8, 10,* 12 и различных функциях распределений выборок *F*(x) и *G*(x) изучено нами совместно с Ю.Э. Камнем и Я.Э. Камнем методом статистических испытаний (Монте-Карло). Рассмотрены различные варианты функций распределения *F(x)* и *G(x).* Результаты показывают, что даже при таких небольших объемах выборок точность аппроксимации предельным стандартным нормальным распределением вполне удовлетворительна. Поэтому представляется целесообразным во всех тех случаях, когда в настоящее время используется критерий Стьюдента, заменить его на критерий Крамера-Уэлча. Конечно, такая замена потребует переделки ряда нормативно-технических и методических документов, исправления учебников и учебных пособий для вузов.

*Пример 1.* Пусть объем первой выборки $m = 120, \bar{x} = 13{,}7, s_x = 5{,}3.$ Для второй выборки $n = 541, \bar{y} = 14{,}1, s_y = 8{,}4.$ Вычислим величину статистики Крамера-Уэлча

$$T = \frac{\sqrt{mn}(\bar{x} - \bar{y})}{\sqrt{ns_x^2 + ms_y^2}} = \frac{\sqrt{120 \times 541}(13{,}7 - 14{,}1)}{\sqrt{541 \times 5{,}3^2 + 120 \times 8{,}4^2}} = \frac{\sqrt{64920}(-0{,}4)}{\sqrt{541 \times 28{,}09 + 120 \times 141{,}12}} =$$

$$= \frac{254{,}79 \times (-0{,}4)}{\sqrt{15196{,}69 + 16934.4}} = \frac{-101{,}916}{\sqrt{32131{,}09}} = \frac{-101{,}916}{179{,}25} = -0{,}57.$$

Поскольку полученное значение по абсолютной величине меньше 1,96, то гипотеза однородности математических ожиданий принимается на уровне значимости 0,05.

**Непараметрические методы проверки однородности.** В большинстве технических, экономических, медицинских и иных задач представляет интерес не проверка равенства математических ожиданий или иных характеристик распределения, а обнаружение различия генеральных совокупностей, из которых извлечены выборки, т.е. проверка гипотезы $H_0$. Методы проверки гипотезы $H_0$ позволяют обнаружить не только изменение математического ожидания, но и любые иные изменения функции распределения результатов наблюдений при переходе от одной выборки к другой (увеличение разброса, появление асимметрии и т. д.). Как установлено выше, методы, основанные на использовании статистик *t* Стьюдента и *T* Крамера-Уэлча, не позволяют проверять гипотезу $H_0$ . Априорное предположение о принадлежности функций распределения *F(x)* и *G(x)* к какому-либо определенному параметрическому семейству (например, семействам нормальных, логарифмически нормальных, распределений Вейбулла-Гнеденко, гамма-распределений и др.), как также показано выше, обычно нельзя достаточно надежно обосновать. Поэтому для проверки $H_0$ следует использовать методы, пригодные при любом виде *F(x)* и *G(x),* т.е. непараметрические методы. (Напомним, что термин «непараметрический метод» означает, что при использовании этого метода нет необходимости предполагать, что функции распределения результатов наблюдений принадлежат какому-либо определенному параметрическому семейству.)

Для проверки гипотезы $H_0$ разработано много непараметрических методов - критерии Смирнова, типа омега-квадрат (Лемана - Розенблатта), Вилкоксона (Манна-Уитни), Ван-дер-Вардена, Сэвиджа, хи-квадрат и др. [1, 2, 7]. Распределения статистик всех этих критериев при справедливости $H_0$ не зависят от конкретного вида совпадающих функций распределения $F(x) \equiv G(x)$. Следовательно, таблицами точных и предельных (при больших объемах выборок) распределений статистик этих критериев и их процентных точек [1, 2] можно пользоваться при любых непрерывных функциях распределения результатов наблюдений.

**Каким из непараметрических критериев пользоваться?** Как известно [3], для выбора одного из нескольких критериев необходимо сравнить их мощности, определяемые видом альтернативных гипотез. Сравнению мощностей критериев посвящена обширная литература.

Хорошо изучены свойства критериев при альтернативной гипотезе сдвига

$$H_{1c} : G(x)=F(x-d),\ d\neq 0.$$

Критерии Вилкоксона, Ван-дер-Вардена и ряд других ориентированы для применения именно в этой ситуации. Если $m$ раз измеряют характеристику одного объекта и $n$ раз - другого, а функция распределения погрешностей измерения произвольна, но не меняется при переходе от объекта к объекту (это более жесткое требование, чем условие равенства дисперсий), то рассмотрение гипотезы $H_{1c}$ оправдано. Однако в большинстве технических, экономических, медицинских и иных исследований нет оснований считать, что функции распределения, соответствующие выборкам, различаются только сдвигом.

### 3.1.3. Двухвыборочный критерий Вилкоксона

Покажем (и это - основной результат настоящего пункта), что двухвыборочный критерий Вилкоксона (в литературе его называют также критерием Манна-Уитни) предназначен для проверки гипотезы

$$H_0 :\ P(X < Y) = 1/2,$$

где $X$ - случайная величина, распределенная как элементы первой выборки, а $Y$ - второй.

В описанной выше вероятностной модели двух независимых выборок без ограничения общности можно считать, что объем первой из них не превосходит объема второй, $m \le n$, в противном случае выборки можно поменять местами. Обычно предполагается, что функции $F(x)$ и $G(x)$ непрерывны и строго возрастают. Из непрерывности этих функций следует, что с вероятностью 1 все $m + n$ результатов наблюдений различны. В реальных эконометрических данных иногда встречаются совпадения, но сам факт их наличия - свидетельство нарушений предпосылок только что описанной базовой математической модели.

Статистика $S$ двухвыборочного критерия Вилкоксона определяется следующим образом. Все элементы объединенной выборки $X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n$ упорядочиваются в порядке возрастания. Элементы первой выборки $X_1, X_2, ..., X_m$ занимают в общем вариационном ряду места с номерами $R_1, R_2, ..., R_m$, другими словами, имеют ранги $R_1, R_2, ..., R_m$ . Тогда статистика Вилкоксона - это сумма рангов элементов первой выборки

$$S = R_1 + R_2 + ... + R_m .$$

Статистика U Манна-Уитни определяется как число пар $(X_i, Y_j)$ таких, что $X_i < Y_j$ , среди всех $mn$ пар, в которых первый элемент - из первой выборки, а второй - из второй. Как известно [7, с.160],

$$U = mn + m(m+1)/2 - S .$$

Поскольку $S$ и $U$ линейно связаны, то часто говорят не о двух критериях - Вилкоксона и Манна-Уитни, а об одном - критерии Вилкоксона (Манна-Уитни).

Критерий Вилкоксона - один из самых известных инструментов непараметрической статистики (наряду со статистиками типа Колмогорова-Смирнова и коэффициентами ранговой корреляции). Свойствам этого критерия и таблицам его критических значений уделяется место во многих монографиях по математической и прикладной статистике (см., например, [1, 2, 7]).

Однако в литературе имеются и неточные утверждения относительно возможностей критерия Вилкоксона. Так, одни полагают, что с его помощью можно обнаружить любое различие между функциями распределения $F(x)$ и $G(x)$. По мнению других, этот критерий нацелен на проверку равенства медиан распределений, соответствующих выборкам. И то, и другое, строго говоря, неверно. Это будет ясно из дальнейшего изложения.

Введем некоторые обозначения. Пусть $F^{-1}(t)$ - функция, обратная к функции распределения $F(x)$. Она определена на отрезке [0;1]. Положим $L(t) = G(F^{-1}(t))$. Поскольку $F(x)$ непрерывна и

строго возрастает, то $F^{-1}(t)$ и $L(t)$ обладают теми же свойствами. Важную роль в дальнейшем изложении будет играть величина $a = P(X< Y)$ . Как нетрудно показать,

$$a = P(X < Y) = \int_0^1 tdL(t).$$

Введем также параметры

$$b^2 = \int_0^1 L^2(t)dt - (1-a)^2, \quad g^2 = \int_0^1 t^2 dL(t) - a^2.$$

Тогда математические ожидания и дисперсии статистик Вилкоксона и Манна-Уитни согласно [7, с.160] выражаются через введенные величины:

$$M(U) = mna\ ,\ M(S) = mn + m(m+1)/2 - M(U) = mn(1- a) + m(m+1)/2,$$

$$D(S) = D(U) = mn\ [\ (n - 1)\ b^2\ + (m - 1)\ g^2\ + a(1 -a)\ ] \quad . (1)$$

Когда объемы обеих выборок безгранично растут, распределения статистик Вилкоксона и Манна-Уитни являются асимптотически нормальными (см., например, [7, гл. 5 и 6]) с параметрами, задаваемыми формулами (1) .

Если выборки полностью однородны, т.е. их функции распределения совпадают, справедлива гипотеза

$$H_0:\quad F(x) = G(x) \text{ при всех } x, \quad (2)$$

то $L(t) = t$ для $t$ из отрезка [0, 1], $L(t)= 0$ для всех отрицательных $t$ и $L(t)= 1$ для $t > 1$, соответственно $a= 1/2$. Подставляя в формулы (1), получаем, что

$$M(S) = m(m+n+1)/2,\quad D(S) = mn(m+n+1)/ 12 \quad (3) .$$

Следовательно, распределение нормированной и центрированной статистики Вилкоксона

$$T = (\ S - m(m+n+1)/2)\ (mn(m+n+1)/ 12\ )^{-1/2} \quad (4)$$

при росте объемов выборок приближается к стандартному нормальному распределению (с математическим ожиданием 0 и дисперсией 1).

Из асимптотической нормальности статистики $T$ следует, что правило принятия решения для критерия Вилкоксона выглядит так:

- если $|T| \le \Phi(1-\frac{\alpha}{2}),$ то гипотеза (2) однородности (тождества) функций распределений принимается на уровне значимости $\alpha$,

- если же $|T| > \Phi(1-\frac{\alpha}{2}),$ то гипотеза (2) однородности (тождества) функций распределений отклоняется на уровне значимости $\alpha$.

В прикладной статистике наиболее часто применяется уровень значимости $\alpha = 0{,}05$. Тогда значение модуля статистики $T$ Вилкоксона надо сравнивать с граничным значением $\Phi(1-\frac{\alpha}{2}) = 1{,}96.$

*Пример 1.* Пусть даны две выборки. Первая содержит $m = 12$ элементов 17; 22; 3; 5; 15; 2; 0; 7; 13; 97; 66; 14. Вторая содержит $n=14$ элементов 47; 30; 2; 15; 1; 21; 25; 7; 44; 29; 33; 11; 6; 15. Проведем проверку однородности функций распределения двух выборок с помощью критерия Вилкоксона.

Первым шагом является построение общего вариационного ряда для элементов двух выборок (табл.1).

Табл.1. Общий вариационный ряд для элементов двух выборок

| Ранги | 1 | 2 | 3,5 | 3,5 | 5 | 6 | 7 | 8,5 | 8,5 | 10 | 11 | 12 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Элементы выборок | 0 | 1 | 2 | 2 | 3 | 5 | 6 | 7 | 7 | 11 | 13 | 14 | 15 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Номера выборок | 1 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 1 |
| Ранги | 14 | 14 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
| Элементы выборок | 15 | 15 | 17 | 21 | 22 | 25 | 29 | 30 | 33 | 44 | 47 | 66 | 97 |
| Номера выборок | 2 | 2 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |

Хотя с точки зрения теории математической статистики вероятность совпадения двух элементов выборок равна 0, в реальных выборках экономических данных совпадения встречаются. Так, в рассматриваемых выборках, как видно из табл.1, два раза повторяется величина 2, два раза - величина 7 и три раза - величина 15. В таких случаях говорят о наличии "связанных рангов", а соответствующим совпадающим величинам приписывают среднее арифметическое тех рангов которые они занимают. Так, величины 2 и 2 занимают в объединенной выборке места 3 и 4, поэтому им приписывается ранг (3+4)/2=3,5. Величины 7 и 7 занимают в объединенной выборке места 8 и 9, поэтому им приписывается ранг (8+9)/2=8,5. Величины 15, 15 и 15 занимают в объединенной выборке места 13, 14 и 15, поэтому им приписывается ранг (13+14+15)/3=14.

Следующий шаг - подсчет значения статистики Вилкоксона, т.е. суммы рангов элементов первой выборки

$$S = R_1 + R_2 + ... + R_m = 1+3,5+5+6+8,5+11+12+14+16+18+25+26=146.$$

Подсчитаем также сумму рангов элементов второй выборки

$$S_1 = 2+3,5+7+8,5+10+14+14+17+19+20+21+22+23+24= 205.$$

Величина $S_1$ может быть использована для контроля вычислений. Дело в том, что суммы рангов элементов первой выборки $S$ и второй выборки $S_1$ вместе составляют сумму рангов объединенной выборки, т.е. сумму всех натуральных чисел от 1 до $m+n$. Следовательно,

$$S+ S_1 = (m+n)(m+n+1)/2= (12+14)(12+14+1)/2= 351.$$

В соответствии с ранее проведенными расчетами $S+S_1 = 146+205=351$. Необходимое условие правильности расчетов выполнено. Ясно, что справедливость этого условия не гарантирует правильности расчетов.

Перейдем к расчету статистики $T$. Согласно формуле (3)

$$M(S) = 12(12+14+1)/ 2 = 162, D(S) = 12\cdot 14(12+14+1)/ 12= 378 .$$

Следовательно,

$$T = ( S - 162) (378 )^{- 1/2} = (146-162) / 19,44 = - 0.82.$$

Поскольку $|T|\le 1,96$, то гипотеза однородности принимается на уровне значимости 0,05.

Что будет, если поменять выборки местами, вторую назвать первой? Тогда вместо $S$ надо рассматривать $S_1$ . Имеем

$$M(S_1 ) = 14(12+14+1)/ 2 = 189, D(S) = D(S_1 ) = 378 ,$$

$$T_1 = ( S_1 - 189) (378 )^{- 1/2} = (205-162) / 19,44 = 0.82.$$

Таким образом, значения статистики критерия отличаются только знаком (можно показать, что это утверждение верно всегда). Поскольку в правиле принятия решения используется только абсолютная величина статистики, то принимаемое решение не зависит от того, какую выборку считаем первой, а какую второй. Для уменьшения объема таблиц принято считать первой выборку меньшего объема.

Продолжим обсуждение критерия Вилкоксона. Правила принятия решений и таблица критических значений для критерия Вилкоксона строятся в предположении справедливости гипотезы полной однородности, описываемой формулой (2). А что будет, если эта гипотеза неверна? Другими словами, какова мощность критерия Вилкоксона?

Пусть объемы выборок достаточно велики, так что можно пользоваться асимптотической нормальностью статистики Вилкоксона. Тогда в соответствии с формулами (1) статистика $T$ будет асимптотически нормальна с параметрами

$$M(T) = (12mn)^{1/2} (1/2 - a) (m+n+1)^{-1/2},$$
$$D(T) = 12 [(n - 1) b^2 + (m - 1) g^2 + a(1 -a) ] (m+n+1)^{-1} . \quad (5)$$

Из формул (5) видно большое значение гипотезы

$$H_{01}: a = P(X < Y) = 1/2 . \quad (6)$$

Если эта гипотеза неверна, то, поскольку $m \le n$, справедлива оценка

$$|M(T)| \ge (12m n (2n+1)^{-1})^{1/2} |1/2 - a| ,$$

а потому $|M(T)|$ безгранично растет при росте объемов выборок. В то же время, поскольку

$$b^2 \le \int_0^1 L^2(t)dt \le 1, \quad g^2 \le \int_0^1 t^2 dL(t) \le 1, \quad \alpha(1-\alpha) \le 1/4,$$

то

$$D(T) \le 12 [(n - 1) + (m - 1) + 1/4] (m+n+1)^{-1} \le 12. \quad (7)$$

Следовательно, вероятность отклонения гипотезы $H_{01}$, когда она неверна, т.е. мощность критерия Вилкоксона как критерия проверки гипотезы (6), стремится к 1 при возрастании объемов выборок, т.е. критерий Вилкоксона является состоятельным для этой гипотезы при альтернативе

$$AH_{01}: a = P(X < Y) \ne 1/2 . \quad (8) .$$

Если же гипотеза (6) верна, то статистика T асимптотически нормальна с математическим ожиданием 0 и дисперсией, определяемой формулой

$$D(T) = 12 [(n - 1) b^2 + (m - 1) g^2 + 1/4 ] (m+n+1)^{-1} . \quad (9)$$

Гипотеза (6) является сложной, дисперсия (9), как показывают приводимые ниже примеры, в зависимости от значений $b^2$ и $g^2$ может быть как больше 1, так и меньше 1, но согласно неравенству (7) никогда не превосходит 12.

Приведем пример двух функций распределения $F(x)$ и $G(x)$ таких, что гипотеза (6) выполнена, а гипотеза (2) - нет. Поскольку

$$a = P(X < Y) = \int_{-\infty}^{+\infty} F(x)dG(x) , \quad 1 - a = P(Y < X) = \int_{-\infty}^{+\infty} G(x)dF(x) \quad (10)$$

и $a = 1/2$ в случае справедливости гипотезы (2), то для выполнения условия (6) необходимо и достаточно, чтобы

$$\int_{-\infty}^{+\infty} (F(x) - G(x))dF(x) = 0 \quad (11) ,$$

а потому естественно в качестве $F(x)$ рассмотреть функцию равномерного распределения на интервале (-1; 1). Тогда формула (11) переходит в условие

$$\int_{-\infty}^{+\infty} (F(x) - G(x))dF(x) = -\frac{1}{2}\int_{-1}^{+1}\left( G(x) - \frac{(x+1)}{2}\right)dx = 0. \quad (11) .$$

Это условие выполняется, если функция $(G(x) - (x + 1)/2)$ является нечетной.

*Пример 2.* Пусть функции распределения $F(x)$ и $G(x)$ сосредоточены на интервале (-1; 1), на котором

$$F(x) = (x + 1)/2 , \quad G(x) = ( x + 1 + 1/\pi \sin \pi x ) / 2 .$$

Тогда

$$x=F^{-1}(t)=2t -1, \; L(t)=G(F^{-1}(t))=(2t+1/\pi \sin \pi (2t-1))/2=t+1/2\pi \sin \pi (2t-1) .$$

Условие (11) выполнено, поскольку функция $(G(x) - (x + 1)/2)$ является нечетной. Следовательно, $a = 1/2$ . Начнем с вычисления

$$g^2 = \int_0^1 t^2 dL(t) - 1/4 = \int_0^1 t^2 d(t + \frac{1}{2\pi}\sin \pi(2t-1)) - \frac{1}{4}.$$

Поскольку

$$d(t+\frac{1}{2\pi}\sin\pi(2t-1))=(1+\cos\pi(2t-1))dt,$$

то

$$g^2=\int_0^1 t^2(1+\cos\pi(2t-1))dt-\frac{1}{4}=\frac{1}{12}+\int_0^1 t^2\cos\pi(2t-1)dt.$$

С помощью замены переменных $t = (x+1)/2$ получаем, что

$$\int_0^1 t^2\cos\pi(2t-1)dt=\frac{1}{8}\left(\int_{-1}^1 x^2\cos\pi x dx+2\int_{-1}^1 x\cos\pi x dx+\int_{-1}^1\cos\pi x dx\right).$$

В правой части последнего равенства стоят табличные интегралы (см., например, справочник [8, с.71]). Проведя соответствующие вычисления, получаем, что в правой части стоит $1/8 Ч(-4/\pi^2) = -1/(2\pi^2)$. Следовательно,

$$g^2 = 1/12 - 1/(2\pi^2) = 0{,}032672733...$$

Перейдем к вычислению $b^2$. Поскольку

$$b^2=\int_0^1 L^2(t)dt-\frac{1}{4}=\int_0^1\left(t+\frac{1}{2}\pi\sin\pi(2t-1)\right)^2dt-\frac{1}{4},$$

то

$$b^2=\frac{1}{12}+\frac{1}{\pi}\int_0^1(t\sin\pi(2t-1))dt+\left(\frac{\pi}{2}\right)^2\int_0^1\sin^2\pi(2t-1)dt.$$

С помощью замены переменных $t = (x+1)/2$ переходим к табличным интегралам (см., например, справочник [8, с.65]):

$$b^2=\frac{1}{12}+\frac{1}{4\pi}\int_{-1}^1 x\sin\pi x dx+\frac{1}{4\pi}\int_{-1}^1\sin\pi x dx+\frac{1}{8\pi^2}\int_{-1}^1\sin^2\pi x dx.$$

Проведя необходимые вычисления, получим, что

$$b^2=\frac{1}{12}+\frac{1}{4\pi}\left(-\frac{2}{\pi}\right)+0+\frac{1}{8\pi^2}=\frac{1}{12}-\frac{3}{8\pi^2}=0{,}045337893...$$

Следовательно, для рассматриваемых функций распределения нормированная и центрированная статистика Вилкоксона (см. формулу (4)) асимптотически нормальна с математическим ожиданием 0 и дисперсией (см. формулу (9))

$$D(T) = ( 0{,}544\, n + 0{,}392\, m + 2{,}064 )\, (m+n+1)^{-1} .$$

Как легко видеть, дисперсия всегда меньше 1. Это значит, что в рассматриваемом случае гипотеза полной однородности (2) при проверке с помощью критерия Вилкоксона будет приниматься чаще, чем если она на самом деле верна.

На наш взгляд, это означает, что критерий Вилкоксона нельзя считать критерием для проверки гипотезы (2) при альтернативе общего вида. Он не всегда позволяет проверить однородность - не при всех альтернативах. Точно так же критерии типа хи-квадрат нельзя считать критериями проверки гипотез согласия и однородности - они позволяют обнаружить не все различия, поскольку некоторые из них "скрадывает" группировка.

Обсудим теперь, действительно ли критерий Вилкоксона нацелен на проверку равенства медиан распределений, соответствующих выборкам.

*Пример 3.* Построим семейство пар функций распределения $F(x)$ и $G(x)$ таких, что их медианы различны, но для $F(x)$ и $G(x)$ выполнена гипотеза (6). Пусть распределения сосредоточены на интервале (0 ; 1), и на нем $G(x) = x$ , а $F(x)$ имеет кусочно-линейный график с вершинами в точках (0 ; 0), ( $\lambda$ , 1/2 ), ( $\delta$ , 3/4), (1 ; 1). Следовательно,

$F(x) = 0$ при $x < 0$ ;

$F(x) = x / (2\lambda)$ на $[0 ; \lambda)$ ;

$$F(x) = 1/2 + (x - \lambda) / (4\delta - 4\lambda) \text{ на } [\lambda ; \delta) ;$$
$$F(x) = 3/4 + (x - \delta) / (4 - 4\delta) \text{ на } [\delta ; 1] ;$$
$$F(x) = 1 \text{ при } x > 1.$$

Очевидно, что медиана $F(x)$ равна $\lambda$, а медиана $G(x)$ равна 1/2 .

Согласно соотношению (9) для выполнения гипотезы (6) достаточно определить $\delta$ как функцию $\lambda$, $\delta = \delta(\lambda)$, из условия

$$\int_0^1 F(x)dx = \frac{1}{2}.$$

Вычисления дают

$$\delta = \delta(\lambda) = 3(1 - \lambda)/2 .$$

Учитывая, что $\delta$ лежит между $\lambda$ и 1, не совпадая ни с тем, ни с другим, получаем ограничения на $\lambda$, а именно, $1/3 < \lambda < 3/5$ . Итак, построено искомое семейство пар функций распределения.

*Пример 4*. Пусть, как и в примере 3, распределения сосредоточены на интервале (0; 1), и на нем $F(x)=x$. А $G(x)$ - функция распределения, сосредоточенного в двух точках - $\beta$ и 1. Т.е. $G(x) = 0$ при $x$, не превосходящем $\beta$; $G(x) = h$ на $(\beta; 1]$; $G(x) = 1$ при $x > 1$. С такой функцией $G(x)$ легко проводить расчеты. Однако она не удовлетворяет принятым выше условиям непрерывности и строгого возрастания. Вместе с тем легко видеть, что она является предельной (сходимость в каждой точке отрезка [0; 1]) для последовательности функций распределения, удовлетворяющих этим условиям. А распределение статистики Вилкоксона для пары функций распределения примера 4 является предельным для последовательности соответствующих распределений статистики Вилкоксона, полученных в рассматриваемых условиях непрерывности и строгого возрастания.

Условие $P(X < Y) = 1/2$ выполнено, если $h = (1 - \beta)^{-1} / 2$ (при $\beta$ из отрезка [0; 1/2] ). Поскольку $h > 1/2$ при положительном $\beta$, то очевидно, что медиана $G(x)$ равна $\beta$, в то время как медиана $F(x)$ равна 1/2 . Значит, при $\beta = 1/2$ медианы совпадают, при всех иных положительных $\beta$ - различны. При $\beta = 0$ медианой $G(x)$ является любая точка из отрезка [0; 1].

Легко подсчитать, что в условиях примера 4 параметры предельного распределения имеют вид

$$b^2 = \beta(1- \beta)^{-1} / 4 , \; g^2 = (1- 2\beta) / 4 .$$

Следовательно, распределение нормированной и центрированной статистики Вилкоксона будет асимптотически нормальным с математическим ожиданием 0 и дисперсией

$$D(T) = 3 [(n-1) \beta(1- \beta)^{-1} + (m-1)(1-2\beta) + 1] (m+n+1)^{-1} .$$

Проанализируем величину $D(T)$ в зависимости от параметра $\beta$ и объемов выборок $m$ и $n$. При достаточно больших $m$ и $n$

$$D(T) = 3 w \beta (1 - \beta)^{-1} + 3 (1 - w)(1 - 2\beta) ,$$

с точностью до величин порядка $(m+n)^{-1}$ , где $w= n/(m+n)$. Значит, $D(T)$ - линейная функция от $w$, а потому достигает экстремальных значений на границах интервала изменения $w$, т.е. при $w = 0$ и $w = 1$. Легко видеть, что при $\beta(1- \beta)^{-1} <1-2\beta$ минимум равен $3\beta(1- \beta)^{-1}$ (при $w = 1$), а максимум равен $3(1 - 2\beta)$ (при $w = 0$). В случае $\beta(1- \beta)^{-1} >1-2\beta$ максимум равен $3\beta(1- \beta)^{-1}$ (при $w = 1$), а минимум равен $3(1 - 2\beta)$ (при $w = 0$). Если же $\beta(1- \beta)^{-1} =1-2\beta$ (это равенство справедливо при $\beta = \beta_0 = 1 - 2^{-1/2} = 0{,}293$), то $D(T)=3 (2^{1/2}-1)=1{,}2426...$ при всех $w$ из отрезка [0; 1].

Первый из описанных выше случаев имеет быть при $\beta < \beta_0$ , при этом минимум $D(T)$ возрастает от 0 (при $\beta=0$, $w=1$ - предельный случай) до $3(2^{1/2} - 1)$ (при $\beta = \beta_0$ , $w$ - любом), а максимум уменьшается от 3 (при $\beta=0$, $w=0$ - предельный случай) до $3 (2^{1/2} - 1)$ (при $\beta = \beta_0$ , $w$ -

любом). Второй случай относится к $\beta$ из интервала ( $\beta_0$ ; 1/2]. При этом минимум убывает от приведенного выше значения для $\beta = \beta_0$ до 0 (при $\beta$=1/2 , $w$=0 - предельный случай) , а максимум возрастает от того же значения при $\beta = \beta_0$ до 3 (при $\beta$=1/2 , $w$=0).

Таким образом, $D(T)$ может принимать все значения из интервала (0; 3) в зависимости от значений $\beta$ и $w$. Если $D(T) < 1$, то при применении критерия Вилкоксона к выборкам с рассматриваемыми функциями распределения гипотеза однородности (2) будет приниматься чаще (при соответствующих значениях $\beta$ и $w$ - с вероятностью, сколь угодно близкой к 1), чем если бы она самом деле была верна. Если $1<D(T)<3$, то гипотеза (2) также принимается достаточно часто. Так, если уровень значимости критерия Вилкоксона равен 0,05, то (асимптотическая) критическая область этого критерия, как показано выше, имеет вид $\{T: |T| \geq 1,96\}$. Если - самый плохой случай - $D(T)=3$, то гипотеза (2) принимается с вероятностью 0,7422.

**Гипотеза сдвига**. При проверке гипотезы однородности мы рассмотрели различные виды нулевых и альтернативных гипотез - гипотезу (2) и ее отрицание в качестве альтернативы, гипотезу (6) и ее отрицание, гипотезы о равенстве или различии медиан. В теоретических работах по математической статистике часто рассматривают гипотезу сдвига, в которой альтернативой гипотезе (2) является гипотеза

$$H_1: \quad F(x) = G(x + r) \quad (12)$$

при всех $x$ и некотором сдвиге $r$, отличным от 0. Если верна альтернативная гипотеза $H_1$, то вероятность $P(X < Y)$ отлична от 1/2, а потому при альтернативе (12) критерий Вилкоксона является состоятельным.

В некоторых прикладных постановках гипотеза (12) представляется естественной. Например, если одним и тем же прибором проводятся две серии измерений двух значений некоторой величины (физической, химической и т.п.). При этом функция распределения $G(x)$ описывает погрешности измерения одного значения, а $G(x+r)$ - другого. Вопреки распространенному заблуждению, хорошо известно, что распределение погрешностей измерений, как правило, не является нормальным (см. об этом главу 2.1). Однако при анализе конкретных статистических данных, как правило, нет никаких оснований считать, что отсутствие однородности всегда выражается столь однозначным образом, как следует из формулы (12). Поэтому эконометрику для проверки однородности необходимо использовать статистические критерии, состоятельные против любого отклонения от гипотезы однородности (2).

Почему же математики так любят гипотезу сдвига (12)? Да потому, что она дает возможность доказывать глубокие математические результаты, например, об асимптотической оптимальности критериев. К сожалению, с точки зрения эконометрики это напоминает поиск ключей под фонарем, где светло, а не там, где они потеряны.

Отметим еще одно обстоятельство. Часто говорят (в соответствии с классическим подходом математической статистики), что нельзя проверять нулевые гипотезы без рассмотрения альтернативных. Однако при анализе данных технических, экономических, медицинских или иных исследований зачастую полностью ясна формулировка той гипотезы, которую желательно проверить (например, гипотезы полной однородности - см. формулу (2)), в то время как формулировка альтернативной гипотезы не очевидна (то ли это гипотеза о неверности равенства (2) хотя бы для одного значения $x$, то ли это альтернатива (8), то ли - альтернатива сдвига (12), и т. д.). В таких случаях целесообразно "обернуть" задачу - исходя из статистического критерия найти альтернативы, относительно которых он состоятелен. Именно это и проделано в настоящем пункте для критерия Вилкоксона.

Подведем итоги рассмотрения критерия Вилкоксона.

1. Критерий Вилкоксона (Манна-Уитни) является одним из самых распространенных непараметрических ранговых критериев, используемых для проверки однородности двух выборок. Его значение не меняется при любом монотонном преобразовании шкалы измерения (т.е. он пригоден для статистического анализа данных, измеренных в порядковой шкале).

2. Распределение статистики критерия Вилкоксона определяется функциями распределения *F(x)* и *G(x)* и объемами *m* и *n* двух выборок. При больших объемах выборок распределение статистики Вилкоксона является асимптотически нормальным с параметрами, выписанными выше (см. формулы (1), (3) и (5)).

3. При альтернативной гипотезе, когда функции распределения выборок *F(x)* и *G(x*) не совпадают, распределение статистики Вилкоксона зависит от величины $a = P(X < Y)$. Если *a* отличается от 1/2, то мощность критерия Вилкоксона стремится к 1, и отличает нулевую гипотезу $F = G$ от альтернативной. Если же $a = 1/2$, то это не всегда имеет место. В примере 2 приведены две *различные* функции распределения выборок *F(x)* и *G(x)* такие, что гипотеза однородности $F = G$ при проверке с помощью критерия Вилкоксона будет приниматься *чаще,* чем если она на самом деле верна.

4. Следовательно, в случае общей альтернативы критерий Вилкоксона не является состоятельным, т.е. не всегда позволяет обнаружить различие функций распределения. Однако это не лишает его практической ценности, точно так же, как несостоятельность критериев типа хи-квадрат при проверке согласия, независимости или однородности не мешает отклонять нулевую гипотезу во многих практически важных случаях. Однако принятие нулевой гипотезы с помощью критерия Вилкоксона может означать не совпадение *F* и *G*, а лишь выполнение равенства $a = 1/2$.

5. Иногда утверждают, что с помощью критерия Вилкоксона можно проверять равенство медиан функций распределения *F* и *G*. Это не так. В примерах 3 и 4 указаны *F* и *G* с $a = 1/2$, но с различными медианами. Во многих случаях это различие нельзя обнаружить с помощью критерия Вилкоксона, как это показано при численном анализе асимптотической дисперсии в примере 4.

6. Указанные выше недостатки критерия Вилкоксона исчезают для специального вида альтернативы - т.н. "альтернативы сдвига" $H_1: F(x) = G(x + r)$. В этом частном случае при справедливости альтернативной гипотезы мощность стремится к 1, различие медиан также всегда обнаруживается. Однако альтернатива сдвига не всегда естественна. Ее целесообразно принять, если одним и тем же прибором проводятся две серии измерений двух значений некоторой величины (физической, химической и т.п.). При этом функция распределения *G(x)* описывает результаты измерений с погрешностями одного значения, а $F(x) = G(x+r)$ - другого. Другими словами, меняется лишь измеряемое значение, а собственно распределение погрешностей - одно и то же, присущее используемому средству измерения (и обычно описанное в его техническом паспорте). Однако в большинстве статистических исследований нет никаких оснований считать, что при альтернативе функция распределения второй выборки лишь сдвигается, но не меняется каким-либо иным образом.

7. При всех своих недостатках критерий Вилкоксона прост в применении и часто позволяет обнаруживать различие групп (поскольку оно часто сводится к отличию $a = P(X < Y)$ от 1/2 ). Приведенные здесь критические замечания не следует понимать как призыв к полному отказу от использования критерия Вилкоксона. Однако для проверки гипотезы однородности в случае альтернативы общего вида можно порекомендовать состоятельные критерии, в частности, рассматриваемые в следующем пункте критерии Смирнова и типа омега-квадрат (Лемана-Розенблатта).

8. В литературе по прикладным статистическим методам соседствуют два стиля изложения. Один из них исходит из формулировок нулевой и альтернативных гипотез (или описания набора гипотез, из которого надо выбрать наиболее адекватную), для проверки которых строятся те или иные критерии. При другом стиле изложения упор делается на алгоритмическое описание критериев для проверки тех или иных гипотез, а об альтернативах даже не упоминается.

Например, в литературе по математической статистике часто говорится, что для проверки нормальности используются критерии асимметрии и эксцесса (они описаны, например, в главе 2.3 и в лучшем справочнике 1960-1980-х годов [1, табл. 4.7]). Однако эти критерии позволяют проверять некоторые соотношения между моментами распределения, но отнюдь не являются состоятельными критериями нормальности (не все отклонения от нормальности обнаруживают).

Впрочем, для прикладной статистики эти критерии большого практического значения не имеют, поскольку заранее известно, что распределения конкретных технических, экономических, медицинских и иных данных скорее всего отличны от нормальных.

Так что недостатки критерия Вилкоксона не является исключением, мощность ряда иных популярных в математической статистике критериев заслуживает тщательного изучения, при этом заранее можно сказать, что зачастую они не позволяют проверять те гипотезы, с которыми традиционно связаны. При применении подобных критериев к анализу реальных данных необходимо тщательно взвешивать их достоинства и недостатки.

### 3.1.4. Состоятельные критерии проверки однородности независимых выборок

В соответствии с методологией прикладной статистики естественно потребовать, чтобы рекомендуемый для массового использования в технических, экономических, медицинских и иных исследованиях критерий однородности был состоятельным. Напомним: это значит, что для любых отличных друг от друга функций распределения $F(x)$ и $G(x)$ (другими словами, при справедливости альтернативной гипотезы $H_1$) вероятность отклонения гипотезы $H_0$ должна стремиться к 1 при увеличении объемов выборок $m$ и $n$. Из перечисленных выше (в конце п. 3.1.2) критериев однородности состоятельными являются только критерии Смирнова и типа омега-квадрат.

Проведенное исследование мощности (методом статистических испытаний) первых четырех из перечисленных выше критериев (при различных вариантах функций распределения $F(x)$ и $G(x)$) подтвердило преимущество критериев Смирнова и омега-квадрат и при объемах выборок 6-12. Рассмотрим эти критерии подробнее.

**Критерий Смирнова однородности двух независимых выборок.** Он предложен членом-корреспондентом АН СССР Н.В. Смирновым в 1939 г. (см. справочник [1]). Единственное ограничение - функции распределения $F(x)$ и $G(x)$ должны быть непрерывными. Напомним, что согласно Л.Н. Большеву и Н.В. Смирнову [1] значение эмпирической функции распределения в точке $x$ равно доле результатов наблюдений в выборке, меньших $x$. Критерий Смирнова основан на использовании эмпирических функций распределения $F_m(x)$ и $G_n(x)$, построенных по первой и второй выборкам соответственно. Значение статистики Смирнова

$$D_{m,n} = \sup_x | F_m(x) - G_n(x) |$$

сравнивают с соответствующим критическим значением (см., например, [1]) и по результатам сравнения принимают или отклоняют гипотезу $H_0$ о совпадении (однородности) функций распределения. Практически значение статистики $D_{m,n}$ рекомендуется согласно монографии [1] вычислять по формулам

$$D^+_{m,n} = \max_{1\le r\le n}\left[\frac{r}{n} - F_m(y'_r)\right] = \max_{1\le r\le m}\left[G_n(x'_s) - \frac{s-1}{m}\right],$$

$$D^-_{m,n} = \max_{1\le r\le n}\left[F_m(y'_r) - \frac{r-1}{n}\right] = \max_{1\le r\le m}\left[\frac{s}{m} - G_n(x'_s)\right],$$

$$D_{m,n} = \max(D^+_{m,n}, D^-_{m,n}),$$

где $x'_1<x'_2<\ldots<x'_m$ - элементы первой выборки $x_1,x_2,\ldots,x_m$, переставленные в порядке возрастания, а $y'_1<y'_2<\ldots<y'_n$ - элементы второй выборки $y_1,y_2,\ldots,y_n$, также переставленные в порядке возрастания. Поскольку функции распределения $F(x)$ и $G(x)$ предполагаются непрерывными, то вероятность совпадения каких-либо выборочных значений равна 0.

Разработаны алгоритмы и программы для ЭВМ, позволяющие рассчитывать точные распределения, процентные точки и достигаемый уровень значимости для двухвыборочной статистики Смирнова $D_{m,n}$, разработаны подробные таблицы (см., например, методику [9], содержащую описание алгоритмов, тексты программ и подробные таблицы).

Однако у критерия Смирнова есть и недостатки. Его распределение сосредоточено в сравнительно небольшом числе точек, поэтому функция распределения растет большими скачками. В результате не удается выдержать заданный уровень значимости. Реальный уровень значимости может в несколько раз отличаться от номинального (подробному обсуждению неклассического феномена существенного отличия реального уровня значимости от номинального посвящена работа [10]).

**Критерий типа омега-квадрат (Лемана-Розенблатта).** Статистика критерия типа омега-квадрат для проверки однородности двух независимых выборок имеет вид:

$$A = \frac{mn}{m+n} \int_{-\infty}^{\infty} (F_m(x) - G_n(x))^2 \, dH_{m+n}(x) ,$$

где $H_{m+n}(x)$ – эмпирическая функция распределения, построенная по объединенной выборке. Легко видеть, что

$$H_{m+n}(x) = \frac{m}{m+n} F_m(x) + \frac{n}{m+n} G_n(x) .$$

Статистика $A$ типа омега-квадрат была предложена Э. Леманом в 1951 г., изучена М. Розенблаттом в 1952 г., а затем и другими исследователями. Она зависит лишь от рангов элементов двух выборок в объединенной выборке. Пусть $x_1, x_2, ..., x_m$ - первая выборка, $x_1' < x_2' < ... x_m'$ - соответствующий вариационный ряд, $y_1, y_2, ..., y_n$ -вторая выборка, $y_1' < y_2' < ... < y_n'$ - вариационный ряд, соответствующий второй выборке. Поскольку функции распределения независимых выборок непрерывны, то с вероятностью 1 все выборочные значения различны, совпадения отсутствуют. Статистика $A$ представляется в виде (см., например, [1]):

$$A = \frac{1}{mn(m+n)} [m \sum_{i=1}^{m} (r_i - i)^2 + n \sum_{j=1}^{n} (s_j - j)^2] - \frac{4mn-1}{6(m+n)},$$

где $r_i$ - ранг $x'_i$ и $s_j$ - ранг $y'_j$ в общем вариационном ряду, построенном по объединенной выборке.

Правила принятия решений при проверке однородности двух выборок на основе статистик Смирнова и типа омега-квадрат, т.е. таблицы критических значений в зависимости от уровней значимости и объемов значимости приведены, например, в таблицах [1].

**Рекомендации по выбору критерия однородности.** Для критерия типа омега-квадрат нет выраженного эффекта различия между номинальными и реальными уровнями значимости. Поэтому мы *рекомендуем для проверки однородности функций распределения (гипотеза $H_0$) применять статистику А типа омега-квадрат. Если методическое, табличное или программное обеспечение для статистики Лемана - Розенблатта отсутствует, рекомендуем использовать критерий Смирнова. Для проверки однородности математических ожиданий (гипотеза $H'_0$) целесообразно применять критерий Крамера-Уэлча.* По нашему мнению, статистики Стьюдента, Вилкоксона и др. допустимо использовать лишь в отдельных частных случаях, рассмотренных выше.

**Некоторые соображения о внедрении современных методов прикладной статистики в практику технических, экономических, медицинских и иных исследований.** Даже из проведенного выше разбора лишь одной из типичных статистических задач - задачи проверки однородности двух независимых выборок - можно сделать вывод о целесообразности широкого развертывания работ по критическому анализу сложившейся практики статистической обработки данных и по внедрению накопленного арсенала современных методов прикладной статистики. По нашему мнению, широкого внедрения заслуживают, в частности, методы многомерного статистического анализа, планирования эксперимента, статистики объектов нечисловой природы.

Очевидно, рассматриваемые работы должны быть плановыми, организационно оформленными, проводиться мощными самостоятельными организациями и подразделениями. Целесообразно создание службы статистических консультаций в системе научно-исследовательских учреждений и вузов технического, экономического, медицинского профиля.

### 3.1.5. Методы проверки однородности для связанных выборок

Начнем с практического примера. Приведем письмо главного инженера подмосковного химического комбината (некоторые названия изменены).

"Директору Института высоких статистических технологий
и эконометрики (Фамилия, имя, отчество)

Наш комбинат выпускает мастику по ГОСТ (следует номер) и является разработчиком указанного стандарта.

В результате исследовательских работ по подбору стандартного метода определения вязкости мастики на комбинате накоплен большой опыт сравнительных данных определения вязкости по двум методам:

- неразбавленной мастики - на нестандартном приборе фабрики им. Петрова;
- раствора мастики - на стандартном вискозиметре ВЗ-4.

Учитывая высокую компетентность сотрудников Вашего института, прошу Вас, в порядке оказания технической помощи нашему предприятию, поручить соответствующей лаборатории провести обработку представленных данных современными статистическими методами и выдать заключение о наличии (или отсутствии) зависимости между указанными выше методами определения вязкости мастики. Ваш5е заключение необходимо для решения спорного вопроса о целесообразности вновь ввести в ГОСТ (следует номер) метода определения вязкости мастики по вискозиметру ВЗ-4, который, по мнению некоторых потребителей, был необоснованно исключен из этого ГОСТ по изменению № 1.

Заранее благодарю Вас за оказанную помощь.

Приложение: статистика на 3 листах.

Главный инженер (Подпись) (Фамилия, имя, отчество)"

*Комментарий.* Вязкость мастики - один из показателей качества мастики. Измерять этот показатель можно по-разному. И, как оказалось, разные способы измерения дают разные результаты. Ничего необычного в этом нет. Однако поставщику и потребителю следует согласовать способы измерения показателей качества. Иначе достаточно часто поставщик (производитель) будет утверждать, что он выполнил условия контракта, а потребитель заявлять, что нет. Такая конфликтная ситуация иногда называется арбитражной, поскольку для ее решения стороны могут обращаться в арбитражный суд. Простейший метод согласования способов измерения показателей состоит в том, чтобы выбрать один из них и внести в государственный стандарт, который тем самым будет содержать не только описание продукции, перечень ее показателей качества и требований к ним, но и способы измерения этих показателей.

**Заключение по статистическим данным, представленным химическим комбинатом.** Для каждой из 213 партий мастики представлены два числа - результат измерения вязкости на нестандартном приборе фабрики им. Петрова и результат измерения вязкости на стандартном вискозиметре ВЗ-4. Требуется установить, дают ли два указанных метода сходные результаты. Если они дают сходные результаты, то нет необходимости вводить в соответствующий ГОСТ указание о методе определения вязкости. Если же методы дают существенно различные результаты, то подобное указание ввести необходимо.

Для применения эконометрических методов в рассматриваемой задаче необходимо описать вероятностную модель. Считаем, что статистические данные имеют вид $(x_i, y_i), i = 1,2,...,213,$ где $x_i$

-результат измерения на нестандартном приборе фабрики им. Петрова в *i*-ой партии, а $y_i$ - результат измерения вязкости на стандартном вискозиметре ВЗ-4 в той же *i*-ой партии. Пусть $a_i$ - истинное значение показателя качества в *i*-ой партии. Естественно считать, что указанные выше случайные вектора независимы в совокупности. При этом они не являются одинаково распределенными, поскольку отличаются истинными значениями показателей качества $a_i$. Принимаем, что *при каждом i случайные величины* $x_i - a_i$ *и* $y_i - a_i$ *независимы и одинаково распределены*. Это условие и означает *однородность в связанных выборках.* Параметры связи - величины $a_i$ . Их наличие не позволяет объединить первые координаты в одну выборку, вторую - во вторую, как делалось в случае проверки однородности двух независимых выборок.

В предположении непрерывности функций распределения из условия однородности в связанных выборках вытекает, что

$$P(x_i < y_i) = P(x_i \geq y_i) = \frac{1}{2}.$$

Рассмотрим случайные величины $Z_i = x_i - y_i, \quad i = 1,2,...,213$. Из последнего соотношения вытекает, что при справедливости гипотезы однородности для связанных выборок эти случайные величины имеют нулевые медианы. Другими словами, проверка того, что метода измерения вязкости дают схожие результаты, эквивалентна проверке равенства 0 медиан величин $Z_i$.

Для проверки гипотезы о том, что медианы величин $Z_i$ нулевые, применим широко известный критерий знаков (см., например, справочник [1, с.89-91]). Согласно этому критерию необходимо подсчитать, в скольких партиях $x_i < y_i$ и в скольких $x_i \geq y_i$ . Для представленных химическим комбинатом данных $x_i < y_i$ в 187 случаях из 213 и $x_i \geq y_i$ в 26 случаях из 213.

Если рассматриваемая гипотеза верна, то число *W* осуществлений события $\{x_i < y_i\}$ имеет биномиальное распределение с параметрами *p* = 1/2 и *n* = 213. Математическое ожидание *M(W)*=106,5, а среднее квадратическое отклонение $\sigma = \sqrt{np(1-p)} = 7{,}3$. Следовательно, интервал $M(W) \pm 3\sigma$ - это интервал 84≤*W*≤129. Найденное по данным химического комбината значение *W*=187 лежит далеко вне этого интервала. Поэтому рассматриваемую гипотезу необходимо отвергнуть (на любом используемом в прикладных работах уровне значимости, в частности, на уровне значимости 1%).

Таким образом, статистический анализ показывает, что два метода дают существенно различные результаты - по прибору фабрики им. Петрова результаты измерений, как правило, меньше, чем по вискозиметру ВЗ-4. Это означает, что в соответствующий ГОСТ целесообразно ввести указание на метод определения вязкости.

**Система вероятностных моделей при проверке гипотезы однородности для связанных выборок**. Как и в случае проверки однородности для независимых выборок, система вероятностных моделей состоит из трех уровней. Наиболее простая модель - на уровне однородности альтернативного признака - уже рассмотрена. Она сводится к проверке гипотезы о значении параметра биномиального распределения:

$$H_0 : p = \frac{1}{2}.$$

Речь идет о "критерии знаков". При справедливости гипотезы однородности число *W* осуществлений события $\{x_i < y_i\}$ имеет биномиальное распределение с вероятностью успеха *p* = 1/2 и числом испытаний *n*. Альтернативная гипотеза состоит в том, что вероятность успеха отличается от 1/2:

$$H_1 : p \neq \frac{1}{2}.$$

Гипотезу *p* = 1/2 можно проверять как непосредственно с помощью биномиального распределения (используя таблицы или программное обеспечение), так и опираясь на теорему

Муавра-Лапласа. Согласно этой теореме

$$\lim_{n\to\infty} P\{\frac{2W-n}{\sqrt{n}} \le x\} = \Phi(x)$$

при всех $x$, где $\Phi(x)$ - функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Из теоремы Муавра-Лапласа вытекает правило принятия решений на уровне значимости 5%: если

$$\left|\frac{2W-n}{\sqrt{n}}\right| \le 1{,}96,$$

то гипотезу однородности связанных выборок принимают, в противном случае отклоняют. Как обычно, при желании использовать другой уровень значимости применяют в качестве критического значения иной квантиль нормального распределения. Использование предельных теорем допустимо при достаточно больших объемах выборки. По поводу придания точного смысла термину "достаточно большой" продолжаются дискуссии. Обычно считается, что несколько десятков (два-три десятка) - это уже "достаточно много". Более правильно сказать, что ответ зависит от задачи, от ее сложности и практической значимости.

Второй уровень моделей проверки однородности связанных выборок - это уровень проверки однородности характеристик, прежде всего однородности математических ожиданий. Исходные данные - количественные результаты измерений (наблюдений, испытаний, анализов, опытов) двух признаков $x_j$ и $y_j$ ,$j$ = 1,2,…,$n$, а непосредственно анализируются их разности $Z_j = x_j - y_j$ , $j$ = 1,2,…,$n$. Предполагается, что эти разности независимы в совокупности и одинаково распределены, однако функция распределения неизвестна статистику. Необходимо проверить непараметрическую гипотезу

$$H_{01} : M(Z_j) = 0.$$

Альтернативная гипотеза также является непараметрической и имеет вид:

$$H_{11} : M(Z_j) \ne 0.$$

Как и в случае проверки гипотезы согласованности для независимых выборок с помощью критерия Крамера-Уэлча, в рассматриваемой ситуации естественно использовать статистику

$$Q = \sqrt{n}\,\frac{\overline{Z}}{s(Z)},$$

где

$$\overline{Z} = \frac{Z_1 + Z_2 + \ldots + Z_n}{n} -$$

среднее арифметическое разностей, а

$$s(Z) = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(Z_j - \overline{Z})^2} -$$

выборочное среднее квадратическое отклонение. Из Центральной Предельной Теоремы теории вероятностей и теорем о наследовании сходимости, полученных в монографии [4] и описанных в главе 1.4, вытекает, что

$$\lim_{n\to\infty} P\{Q \le x\} = \Phi(x)$$

при всех $x$, где $\Phi(x)$ - функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Отсюда вытекает правило принятия решений на уровне значимости 5%: если

$$|Q| \le 1{,}96,$$

то гипотезу однородности математических ожиданий связанных выборок принимают, в противном случае отклоняют. Как обычно, при желании использовать другой уровень значимости применяют в качестве критического значения иной квантиль нормального распределения.

Повторим, что использование предельных теорем допустимо при достаточно больших объемах выборки.

Третий уровень моделей проверки однородности связанных выборок - это уровень проверки однородности (совпадения) функций распределения. Необходимо проверить непараметрическую гипотезу наиболее всеохватного вида:

$$H_{03} : F(x) = G(x), x \in R^1,$$

где

$$F(x) = P(x_i \le x), G(x) = P(y_i \le x).$$

При этом предполагается, что все участвующие в вероятностной модели случайные величины независимы (в совокупности) между собой.

Отметим одно важное свойство функции распределения случайной величины *Z*. Если случайные величины *X* и *Y* независимы и одинаково распределены, то для функции распределения $H(x)=P(Z \le x)$ случайной величины $Z = X – Y$ выполнено, как нетрудно видеть, соотношение

$$H(-x)=1-H(x).$$

Это соотношение означает симметрию функции распределения относительно 0. Плотность такой функции распределения является четной функцией, ее значения в точках x и *(-x)* совпадают.

Проверка гипотезы однородности связанных выборок в наиболее общем случае сводится к проверке симметрии функции распределения разности $Z = X – Y$ относительно 0.

### 3.1.6. Проверка гипотезы симметрии

Рассмотрим методы проверки гипотезы симметрию функции распределения относительно 0. Сначала обсудим, какого типа отклонения от гипотезы симметрии можно ожидать при альтернативных гипотезах?

Как и в случае проверки однородности независимых выборок, в зависимости от вида альтернативной гипотезы выделяют два подуровня моделей. Рассмотрим сначала альтернативу сдвига

$$H_{13} : G(x) = F(x + a).$$

В этом случае распределение *Z* при альтернативе отличается сдвигом от симметричного относительно 0. Для проверки гипотезы однородности может быть использован критерий знаковых рангов, разработанный Вилкоксоном (см., например, справочник [2, с.46-53]).

Он строится следующим образом. Пусть $R(Z_j)$ является рангом $|Z_j|$ в ранжировке от меньшего к большему абсолютных значений разностей $|Z_1|, |Z_2|,\dots,|Z_n|$, $j=1,2,\dots,n$. Положим для $j=1,2,\dots,n$

$$Q(Z_j) = \begin{cases} 1, Z_j > 0, \\ 0, Z_j < 0. \end{cases}$$

Статистика критерия знаковых рангов имеет вид

$$W^+ = \sum_{j=1}^{n} R(Z_j) Q(Z_j).$$

Таким образом, нужно просуммировать ранги положительных разностей в вариационном ряду, построенном стандартным образом по абсолютным величинам всех разностей.

Для практического использования статистики критерия знаковых рангов Вилкоксона либо обращаются к соответствующим таблицам и программному обеспечению, либо применяют асимптотические соотношения. При выполнении нулевой гипотезы статистика

$$W^{++} = \frac{W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

имеет асимптотическое (при $n \to \infty$) стандартное нормальное распределение с математическим ожиданием 0 и дисперсией 1. Следовательно, правило принятия решений на уровне значимости 5%: имеет обычный вид: если

$$\left|W^{++}\right| \le 1{,}96,$$

то гипотезу однородности связанных выборок по критерию знаковых рангов Вилкоксона принимают, в противном случае отклоняют. Как обычно, при желании использовать другой уровень значимости применяют в качестве критического значения иной квантиль нормального распределения. Повторим еще раз, что использование предельных теорем допустимо при достаточно больших объемах выборки.

Альтернативная гипотеза общего вида записывается как

$$H_{14} : H(-x_0) \ne 1 - H(x_0)$$

при некотором $x_0$ . Таким образом, проверке подлежит гипотеза симметрии относительно 0, которую можно переписать в виде

$$H(x) + H(-x) - 1 = 0 .$$

Для построенной по выборке $Z_j = x_j - y_j$ , $j = 1,2,\ldots,n$, эмпирической функции распределения $H_n(x)$ последнее соотношение выполнено лишь приближенно:

$$H_n(x) + H_n(-x) - 1 \approx 0.$$

Как измерять отличие от 0? По тем же соображениям, что и в предыдущем пункте, целесообразно использовать статистику типа омега-квадрат. Соответствующий критерий был предложен в работе [17]. Он имеет вид

$$\omega_n^2 = \sum_{j=1}^{n} (H_n(Z_j) + H_n(-Z_j) - 1)^2 .$$

В работе [11] найдено предельное распределение этой статистики:

$$\lim_{n\to\infty} P(\omega_n^2 < x) = S_0(x).$$

В табл.1 приведены критические значения статистики типа омега-квадрат для проверки симметрии распределения (и тем самым для проверки однородности связанных выборок), соответствующие наиболее распространенным значениям уровней значимости (расчеты проведены Г.В. Мартыновым).

Табл.1. Критические значения статистики $\omega_n^2$ для проверки симметрии распределения

| Значение функции распределения $S_0(x)$ | Уровень значимости $\alpha = 1 - S_0(x)$ | Критическое значение $x$ статистики $\omega_n^2$ |
|---|---|---|
| 0,90 | 0,10 | 1,20 |
| 0,95 | 0,05 | 1,66 |
| 0,99 | 0,01 | 2,80 |

Как следует из табл.1, правило принятия решений при проверке однородности связанных выборок в наиболее общей постановке и при уровне значимости 5% формулируется так. Вычислить статистику $\omega_n^2$ . Если $\omega_n^2 \le 1{,}66$, то принять гипотезу однородности. В противном случае - отвергнуть.

*Пример.* Пусть величины $Z_j$ , $j$=1,2,…,20, таковы:

20, 18, (-2), 34, 25, (-17), 24, 42, 16, 26,
13, (-23), 35, 21, 19, 8, 27, 11, (-5), 7.

Соответствующий вариационный ряд $Z(1) < Z(2) < \ldots < Z(20)$ имеет вид:

(-23)<(-17)<(-5)<(-2)<7<8<11<13<16<18<
<19<20<21<24<25<26<27<34<35<42.

Для расчета значения статистики $\omega_n^2$ построим табл.2 из 7 столбцов и 20 строк, не считая заголовков столбцов (сказуемого таблицы). В первом столбце указаны номера (ранги) членов вариационного ряда, во втором - сами эти члены, в третьем - значения эмпирической функции распределения при значениях аргумента, совпадающих с членами вариационного ряда. В следующем столбце приведены члены вариационного ряда с обратным знаком, а затем указываются соответствующие значения эмпирической функции распределения. Например, поскольку минимальное наблюдаемое значение равно (-23), то $H_n(x)=0$ при $x<-23$, а потому для членов вариационного ряда с 14-го по 20-й в пятом столбце стоит 0. В качестве другого примера рассмотрим минимальный член вариационного ряда, т.е. (-23). Меняя знак, получаем 23. Это число стоит между 13-м и 14-м членами вариационного ряда, $21<23<24$. На этом интервале эмпирическая функция распределения совпадает со своим значением в левом конце, поэтому следует записать в пятом столбце значение 0,65. Остальные ячейки пятого столбца заполняются аналогично. На основе третьего и пятого столбцов элементарно заполняется шестой столбец, а затем и седьмой. Остается найти сумму значений, стоящих в седьмом столбце. Подобная таблица удобна как для ручного счета, так и при использовании электронных таблиц типа *Excel*.

Табл.2. Расчет значения статистики $\omega_n^2$ для проверки симметрии распределения

| $j$ | $Z(j)$ | $H_n(Z(j))$ | $-Z(j)$ | $H_n(-Z(j))$ | $H_n(Z(j))+H_n(-Z(j))-1$ | $(H_n(Z(j))+H_n(-Z(j))-1)^2$ |
|---|---|---|---|---|---|---|
| 1 | -23 | 0,05 | 23 | 0,65 | -0,30 | 0,09 |
| 2 | -17 | 0,10 | 17 | 0,45 | -0,45 | 0,2025 |
| 3 | -5 | 0,15 | 5 | 0,20 | -0,65 | 0,4225 |
| 4 | -2 | 0,20 | 2 | 0,20 | -0,60 | 0,36 |
| 5 | 7 | 0,25 | -7 | 0,10 | -0,65 | 0,4225 |
| 6 | 8 | 0,30 | -8 | 0,10 | -0,60 | 0,36 |
| 7 | 11 | 0.35 | -11 | 0,10 | -0,55 | 0,3025 |
| 8 | 13 | 0,40 | -13 | 0,10 | -0,50 | 0,25 |
| 9 | 16 | 0,45 | -16 | 0,10 | -0,45 | 0,2025 |
| 10 | 18 | 0,50 | -18 | 0,05 | -0,45 | 0,2025 |
| 11 | 19 | 0,55 | -19 | 0,05 | -0,40 | 0,16 |
| 12 | 20 | 0,60 | -20 | 0,05 | -0,35 | 0,1225 |
| 13 | 21 | 0,65 | -21 | 0,05 | -0,30 | 0,09 |
| 14 | 24 | 0,70 | -24 | 0 | -0,30 | 0,09 |
| 15 | 25 | 0,75 | -25 | 0 | -0,25 | 0,0625 |
| 16 | 26 | 0,80 | -26 | 0 | -0,20 | 0,04 |
| 17 | 27 | 0,85 | -27 | 0 | -0,15 | 0,0225 |
| 18 | 34 | 0,90 | -34 | 0 | -0,10 | 0,01 |
| 19 | 35 | 0,95 | -35 | 0 | -0,05 | 0,0025 |
| 20 | 42 | 1,00 | -42 | 0 | 0 | 0 |

Результаты расчетов (суммирование значений по седьмому столбцу табл.2) показывают, что значение статистики $\omega_n^2=3,055$. В соответствии с табл.1 это означает, что на любом используемом в прикладных эконометрических исследованиях уровнях значимости отклоняется гипотеза симметрии распределения относительно 0 (а потому и гипотеза однородности в связанных выборках).

В настоящей главе затронута лишь небольшая часть непараметрических методов анализа числовых эконометрических данных. В частности, обратим внимание на непараметрические оценки плотности, которые используются для описания данных, проверки однородности, в задачах

восстановления зависимостей и других областях эконометрики. Непараметрические оценки плотности рассмотрены в главе 2.1.

## Литература

1. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983. - 416 с.
2. Холлендер М., Вульф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983. - 518 с.
3. Боровков А.А. Математическая статистика. – М.: Наука, 1984. - 472 с.
4. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука, 1979. – 296 с.
5. Орлов А.И., Друянова Г.Б. Непараметрическое оценивание коэффициентов вариации технических характеристик и показателей качества. – Журнал «Надежность и контроль качества», 1987, No.7, с.10-16.
6. Крамер Г. Математические методы статистики / Пер. с англ. / 2-е изд. - М.: Мир, 1975. – 648 с.
7. Гаек Я., Шидак З. Теория ранговых критериев / Пер. с англ. - М.: Наука, 1971. – 376 с.
8. Смолянский М.Л. Таблицы неопределенных интегралов. - М.: ГИФМЛ, 1961. - 108 с.
9. Методика. Проверка однородности двух выборок параметров продукции при оценке ее технического уровня и качества. – М.: ВНИИ стандартизации, 1987. – 116 с.
10. Камень Ю.Э., Камень Я.Э., Орлов А.И. Реальные и номинальные уровни значимости в задачах проверки статистических гипотез / Заводская лаборатория. 1986. Т.52. № 12. С.55-57.
11. Орлов А.И. О проверке симметрии распределения. – Журнал «Теория вероятностей и ее применения». 1972. Т.17. No.2. С.372-377.

## Контрольные вопросы и задачи

1. Почему непараметрические методы анализа числовых данных предпочтительнее параметрических?
2. Указать доверительные границы для математических ожиданий (с доверительной вероятностью 0,95) и проверить гипотезу о равенстве математических ожиданий с помощью критерия Крамера-Уэлча (уровень значимости $\alpha$=0.05):

| **N** | $\mathbf{n_1}$ | $\overline{X}$ | $\mathbf{s_x}$ | $\mathbf{n_2}$ | $\overline{Y}$ | $\mathbf{s_y}$ |
|---|---|---|---|---|---|---|
| 1 | 100 | 13,7 | 7,3 | 200 | 12,1 | 2,5 |
| 2 | 200 | 10 | 5,3 | 400 | 12 | 1,7 |

3. Проверить гипотезу об однородности функций распределения с помощью критерия Вилкоксона (на уровне значимости $\alpha$=0.05):

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Первая выборка | 33 | 27 | 12 | 27 | 39 | 42 | 47 | 48 | 50 | 32 |
| Вторая выборка | 11 | 20 | 30 | 31 | 22 | 18 | 17 | 25 | 28 | 29 |

4. Для каждого из $N$ = 20 объектов даны значения $X_j$ и $Y_j$ , $j$ = 1,2,…,$N$, результатов измерений (наблюдений, испытаний, анализов, опытов) двух признаков. Необходимо проверить, есть ли значимое различие между значениями двух признаков или же это различие может быть объяснено случайными отклонениям значений признаков. Другими словами, требуется проверить однородность (т.е. отсутствие различия) связанных выборок.

Табл. Исходные данные для задачи 4.

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_j$ | 74 | 79 | 65 | 69 | 71 | 66 | 71 | 73 | 72 | 68 |
| $Y_j$ | 73 | 65 | 71 | 69 | 70 | 69 | 78 | 70 | 60 | 62 |
| $j$ | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_j$ | 70 | 69 | 76 | 74 | 72 | 69 | 74 | 72 | 77 | 75 |
| $Y_j$ | 61 | 67 | 73 | 67 | 73 | 64 | 67 | 65 | 63 | 70 |

Проверку однородности на уровне значимости 0,05 проведите с помощью трех критериев:

А) критерия знаков (основанного на проверке гипотезы $p = 0{,}5$ для биномиального распределения с использованием теоремы Муавра-Лапласа);

Б) критерия для проверки равенства 0 математического ожидания (критерий основан на асимптотической нормальности выборочного среднего арифметического, деленного на выборочное среднее квадратическое отклонение);

В) критерия типа омега-квадрат для проверки гипотезы симметрии функции распределения (разности результатов измерений, наблюдений, испытаний, анализов, опытов для двух признаков) относительно 0.

5. Какова роль альтернативных гипотез, в частности, гипотезы сдвига, в выборе критерия для проверки нулевой гипотезы?

**Темы докладов, рефератов, исследовательских работ**

1. Асимптотическая нормальность оценок параметров как основа для проверки гипотез о параметрах.
2. Сравнение двухвыборочных критериев Крамера-Уэлча и Стьюдента.
3. Достоинства и недостатки двухвыборочного критерия Вилкоксона по сравнению с другими непараметрическими критериями однородности.
4. Для данных задачи 3 рассчитайте значения статистик Смирнова и типа омега-квадрат (Лемана – Розенблатта) и проверьте однородность двух выборок.

*Примечание.* Для уровня значимости 0,05 критическим значением для критерия Смирнова является 0,7 (т.е. гипотеза однородности отклоняется, если значение статистики Смирнова не менее 0,7). Для того же уровня значимости критическим значением для критерия типа омега-квадрат (Лемана – Розенблатта) является 0,461.

5. Многообразие непараметрических критериев проверки гипотез (по монографиям [1, 2, 7]).
6. Сравнение мощностей непараметрических критериев однородности.

## 3.2. Многомерный статистический анализ

В многомерном статистическом анализе выборка состоит из элементов многомерного пространства. Отсюда и название этого раздела прикладной статистики. Из многих задач многомерного статистического анализа рассмотрим основные – корреляцию, восстановление зависимости, классификацию, уменьшение размерности, индексы.

### 3.2.1. Коэффициенты корреляции

Термин "корреляция" означает "связь". В эконометрике этот термин обычно используется в сочетании "коэффициенты корреляции". Рассмотрим линейный и непараметрические парные коэффициенты корреляции.

Обсудим способы измерения связи между двумя случайными переменными. Пусть исходными данными является набор случайных векторов $(x_i, y_i) = (x_i(\omega), y_i(\omega)),\ i = 1,2,...n.$ Выборочным коэффициентом корреляции, более подробно, выборочным линейным парным коэффициентом корреляции К. Пирсона, как известно, называется число

$$r_n = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Если $r_n$ = 1, то $y_i = ax_i + b,$ причем $a$>0. Если же $r_n$ = - 1, то $y_i = ax_i + b,$ причем $a$<0. Таким образом, близость коэффициента корреляции к 1 (по абсолютной величине) говорит о достаточно тесной линейной связи.

Если случайные вектора $(x_i, y_i) = (x_i(\omega), y_i(\omega)),\ i = 1,2,...n,$ независимы и одинаково распределены, то выборочный коэффициент корреляции сходится к теоретическому при безграничном возрастании объема выборки:

$$r_n \to \rho = \frac{M(x_1 - M(x_1))(y_1 - M(y_1))}{\sqrt{D(x_1)}\sqrt{D(y_1)}}$$

(сходимость по вероятности).

Более того, выборочный коэффициент корреляции является асимптотически нормальным. Это означает, что

$$\lim_{n\to\infty} P\left(\frac{r_n - \rho}{\sqrt{D_0(r_n)}} < x\right) = \Phi(x),$$

где $\Phi(x)$ - функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1, а $D_0(r_n)$ - асимптотическая дисперсия выборочного коэффициента корреляции. Она имеет довольно сложное выражение, приведенное в монографии [1, с.393]:

$$D_0(r_n) = \frac{\rho^2}{4n}\left(\frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}} + \frac{4\mu_{22}}{\mu_{11}^2} - \frac{4\mu_{31}}{\mu_{11}\mu_{20}} - \frac{4\mu_{13}}{\mu_{11}\mu_{02}}\right).$$

Здесь под $\mu_{km}$ понимаются теоретические центральные моменты порядка $k$ и $m$, а именно,

$$\mu_{km} = M(x_1 - M(x_1))^k (y_1 - M(y_1))^m.$$

Коэффициенты корреляции типа $r_n$ используются во многих алгоритмах многомерного статистического анализа. В теоретических рассмотрениях часто считают, что случайные вектора $(x_i, y_i) = (x_i(\omega), y_i(\omega)),\ i = 1,2,...n,$ имеют двумерное нормальное распределение. Распределения реальных данных, как правило, отличны от нормальных (см. главу 2.1). Почему же распространено представление о двумерном нормальном распределении? Дело в том, что теория в этом случае проще. В частности, равенство 0 теоретического коэффициента корреляции эквивалентно независимости случайных величин. Поэтому проверка независимости сводится к проверке статистической гипотезы о равенстве 0 теоретического коэффициента корреляции. Эта

гипотеза принимается, если $|r_n| < C(n,\alpha)$, где $C(n,\alpha)$ - некоторое граничное значение, зависящее от объема выборки $n$ и уровня значимости $\alpha$.

Если предположение о двумерной нормальности не выполнено, то из равенства 0 теоретического коэффициента корреляции не вытекает независимость случайных величин. Нетрудно построить пример случайного вектора, для которого коэффициент корреляции равен 0, но координаты зависимы. Кроме того, для проверки гипотез о коэффициенте корреляции нельзя пользоваться таблицами, рассчитанными в предположении нормальности. Можно построить правила принятия решений на основе асимптотической нормальности выборочного коэффициента корреляции. Но есть и другой путь – перейти к непараметрическим коэффициентам корреляции, одинаково пригодным при любом непрерывном распределении случайного вектора.

Для расчета непараметрического *коэффициента ранговой корреляции Спирмена* необходимо сделать следующее. Для каждого $x_i$ рассчитать его ранг $r_i$ в вариационном ряду, построенном по выборке $x_1, x_2, ....x_n$. Для каждого $y_i$ рассчитать его ранг $q_i$ в вариационном ряду, построенном по выборке $y_1, y_2, ....y_n$. Для набора из $n$ пар $(r_i, q_i)$, $i = 1,2,...,n$, вычислить линейный коэффициент корреляции. Он называется коэффициентом ранговой корреляции, поскольку определяется через ранги. В качестве примера рассмотрим данные из табл.1 (см. монографию [2]).

Таблица 1.

Данные для расчета коэффициентов корреляции

| $i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x_i$ | 5 | 10 | 15 | 20 | 25 |
| $y_i$ | 6 | 7 | 30 | 81 | 300 |
| $r_i$ | 1 | 2 | 3 | 4 | 5 |
| $q_i$ | 1 | 2 | 3 | 4 | 5 |

Для данных табл.1 коэффициент линейной корреляции равен 0,83, непосредственной линейной связи нет. А вот коэффициент ранговой корреляции равен 1, поскольку увеличение одной переменной однозначно соответствует увеличению другой переменной. Во многих экономических задачах, например, при выборе инвестиционных проектов, достаточно именно монотонной зависимости одной переменной от другой.

Поскольку суммы рангов и их квадратов нетрудно подсчитать, то *коэффициент ранговой корреляции Спирмена* равен

$$\rho_n = 1 - \frac{6\sum_{i=1}^{n}(r_i - q_i)^2}{n^3 - n}.$$

Отметим, что *коэффициент ранговой корреляции Спирмена* остается постоянным при любом строго возрастающем преобразовании шкалы измерения результатов наблюдений. Другими словами, он является адекватным в порядковой шкале (см. главу 2.1), как и другие ранговые статистики, например, статистики Вилкоксона, Смирнова, типа омега-квадрат для проверки однородности независимых выборок (глава 3.1).

Широко используется также коэффициент ранговой корреляции $\tau$ Кендалла, коэффициент ранговой конкордации Кендалла и Б. Смита и др. Наиболее подробное обсуждение этой тематики содержится в монографии [3], необходимые для практических расчетов таблицы имеются в справочнике [4]. Дискуссия о выборе вида коэффициентов корреляции продолжается до настоящего времени [2].

### 3.2.2. Восстановление линейной зависимости между двумя переменными

Начнем с задачи точечного и доверительного оценивания линейной функции одной переменной.

Исходные данные – набор $n$ пар чисел $(t_k, x_k)$, $k = 1,2,\ldots,n,$ где $t_k$ – независимая переменная (например, время), а $x_k$ – зависимая (например, индекс инфляции, курс доллара США, объем месячного производства или размер дневной выручки торговой точки). Предполагается, что переменные связаны зависимостью

$$x_k = a\,(t_k - t_{cp}) + b + e_k,\ k = 1,2,\ldots,n,$$

где $a$ и $b$ – параметры, неизвестные статистику и подлежащие оцениванию, а $e_k$ – погрешности, искажающие зависимость. Среднее арифметическое моментов времени

$$t_{cp} = (t_1 + t_2 + \ldots + t_n)/n$$

введено в модель для облегчения дальнейших выкладок.

Обычно оценивают параметры $a$ и $b$ линейной зависимости методом наименьших квадратов. Затем восстановленную зависимость используют, например, для точечного и интервального прогнозирования.

Как известно, метод наименьших квадратов был разработан великим немецким математиком К. Гауссом в 1794 г. Согласно этому методу для расчета наилучшей функции, приближающей линейным образом зависимость $x$ от $t$, следует рассмотреть функцию двух переменных

$$f(a,b) = \sum_{i=1}^{n} (x_i - a(t_i - t_{cp}) - b)^2.$$

Оценки метода наименьших квадратов - это такие значения $a^*$ и $b^*$, при которых функция $f(a,b)$ достигает минимума по всем значениям аргументов.

Чтобы найти эти оценки, надо вычислить частные производные от функции $f(a,b)$ по аргументам $a$ и $b$, приравнять их 0, затем из полученных уравнений найти оценки: Имеем:

$$\frac{\partial f(a,b)}{\partial a} = \sum_{i=1}^{n} 2(x_i - a(t_i - t_{cp}) - b)(-(t_i - t_{cp})),$$

$$\frac{\partial f(a,b)}{\partial b} = \sum_{i=1}^{n} 2(x_i - a(t_i - t_{cp}) - b)(-1).$$

Преобразуем правые части полученных соотношений. Вынесем за знак суммы общие множители 2 и (-1). Затем рассмотрим слагаемые. Раскроем скобки в первом выражении, получим, что каждое слагаемое разбивается на три. Во втором выражении также каждое слагаемое есть сумма трех. Значит, каждая из сумм разбивается на три суммы. Имеем:

$$\frac{\partial f(a,b)}{\partial a} = (-2)(\sum_{i=1}^{n} x_i (t_i - t_{cp}) - a\sum_{i=1}^{n} (t_i - t_{cp})^2 - b\sum_{i=1}^{n} (t_i - t_{cp})),$$

$$\frac{\partial f(a,b)}{\partial b} = (-2)(\sum_{i=1}^{n} x_i - a\sum_{i=1}^{n} (t_i - t_{cp}) \ - bn).$$

Приравняем частные производные 0. Тогда в полученных уравнениях можно сократить множитель (-2). Поскольку

$$\sum_{i=1}^{n} (t_i - t_{cp}) = 0, \qquad (1)$$

уравнения приобретают вид

$$\sum_{i=1}^{n} x_i (t_i - t_{cp}) - a\sum_{i=1}^{n} (t_i - t_{cp})^2 = 0,$$

$$\sum_{i=1}^{n} x_i - bn = 0.$$

Следовательно, оценки метода наименьших квадратов имеют вид

$$a^* = \frac{\sum_{i=1}^{n} x_i (t_i - t_{cp})}{\sum_{i=1}^{n} (t_i - t_{cp})^2}, \quad b^* = x_{cp} = \frac{x_1 + x_2 + \ldots + x_n}{n} \ . \ (2)$$

В силу соотношения (1) оценку a* можно записать в более симметричном виде:

Эту оценку нетрудно преобразовать и к виду

$$a^* = \frac{\sum_{i=1}^{n} (x_i - x_{cp})(t_i - t_{cp})}{\sum_{i=1}^{n} (t_i - t_{cp})^2}. \qquad (3)$$

$$a^* = \frac{\sum_{i=1}^{n} x_i t_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} t_i^{\,2} - \frac{1}{n}\left(\sum_{i=1}^{n} t_i\right)^2}. \qquad (4)$$

Следовательно, восстановленная функция, с помощью которой можно прогнозировать и интерполировать, имеет вид

$$x^*(t) = a^*(t - t_{cp}) + b^*.$$

Обратим внимание на то, что использование $t_{cp}$ в последней формуле ничуть не ограничивает ее общность. Сравним с моделью вида

$$x_k = c\,t_k + d + e_k,\ k = 1,2,\ldots,n.$$

Ясно, что

$$c = a,\ \ d = b - at_{cp}.$$

Аналогичным образом связаны оценки параметров:

$$c^* = a^*,\ \ d^* = b^* - a^* t_{cp}.$$

Для получения оценок параметров и прогностической формулы нет необходимости обращаться к какой-либо вероятностной модели. Однако для того, чтобы изучать погрешности оценок параметров и восстановленной функции, т.е. строить доверительные интервалы для $a^*$, $b^*$ и $x^*(t)$, подобная модель необходима.

**Непараметрическая вероятностная модель**. Пусть значения независимой переменной $t$ детерминированы, а погрешности $e_k$, $k = 1,2,\ldots,n$, - независимые одинаково распределенные случайные величины с нулевым математическим ожиданием и дисперсией $\sigma^2$, неизвестной статистику.

В дальнейшем неоднократно будем использовать Центральную Предельную Теорему (ЦПТ) теории вероятностей для величин $e_k$, $k = 1,2,\ldots,n$ (с весами), поэтому для выполнения ее условий необходимо предположить, например, что погрешности $e_k$, $k = 1,2,\ldots,n$, финитны или имеют конечный третий абсолютный момент. Однако заострять внимание на этих внутриматематических "условиях регулярности" нет необходимости.

**Асимптотические распределения оценок параметров.** Из формулы (2) следует, что

$$b^* = \frac{a}{n}\sum_{i=1}^{n} (t_i - t_{cp}) + b + \frac{1}{n}\sum_{i=1}^{n} e_i = b + \frac{1}{n}\sum_{i=1}^{n} e_i \quad . \quad (5)$$

Согласно ЦПТ оценка $b^*$ имеет асимптотически нормальное распределение с математическим ожиданием $b$ и дисперсией $\sigma^2 / n$, оценка которой приводится ниже.

Из формул (2) и (5) вытекает, что

Последнее слагаемое во втором соотношении при суммировании по $i$ обращается в 0, поэтому из

$$x_i - x_{cp} = a(t_i - t_{cp}) + b + e_i - b - \frac{1}{n}\sum_{i=1}^{n} e_i,$$

$$(x_i - x_{cp})(t_i - t_{cp}) = a(t_i - t_{cp})^2 + e_i(t_i - t_{cp}) - \frac{(t_i - t_{cp})}{n}\sum_{i=1}^{n} e_i.$$

формул (2-4) следует, что

$$a^* = a + \sum_{i=1}^{n} c_i e_i,\ \ c_i = \frac{(t_i - t_{cp})}{\sum_{i=1}^{n} (t_i - t_{cp})^2}. \quad (6)$$

Формула (6) показывает, что оценка $a*$ является асимптотически нормальной с математическим ожиданием $a$ и дисперсией

$$D(a*) = \sum_{i=1}^{n} c_i^2 D(e_i) = \frac{\sigma^2}{\sum_{i=1}^{n} (t_i - t_{cp})^2} \quad .$$

Отметим, что многомерная нормальность имеет быть, когда каждое слагаемое в формуле (6)

мало сравнительно со всей суммой, т.е.

$$\lim_{n\to\infty} \max |t_i - t_{cp}| / \{\sum_{i=1}^{n} (t_i - t_{cp})^2\}^{1/2} = 0 \quad .$$

Из формул (5) и (6) и исходных предположений о погрешностях вытекает также несмещенность оценок параметров.

Несмещенность и асимптотическая нормальность оценок метода наименьших квадратов позволяют легко указывать для них асимптотические доверительные границы (аналогично границам в предыдущей главе) и проверять статистические гипотезы, например, о равенстве определенным значениям, прежде всего 0. Предоставляем читателю возможность выписать формулы для расчета доверительных границ и сформулировать правила проверки упомянутых гипотез.

**Асимптотическое распределение прогностической функции.** Из формул (5) и (6) следует, что

$$M(x*(t)) = M\{a*(t - t_{cp}) + b*\} = M(a*)(t - t_{cp}) + M(b*) = a(t - t_{cp}) + b = x(t),$$

т.е. рассматриваемая оценка прогностической функции является несмещенной. Поэтому

$$D(x*(t)) = D(a*)(t - t_{cp})^2 + 2M\{(a* - a)(b* - b)(t - t_{cp})\} + D(b*).$$

При этом, поскольку погрешности независимы в совокупности и $M(e_i) = 0$, то

$$M\{(a* - a)(b* - b)(t - t_{cp})\} = \frac{1}{n} \sum_{i=1}^{n} c_i (t - t_{cp}) M(e_i^2) = \frac{1}{n} (t - t_{cp}) \sigma^2 \sum_{i=1}^{n} c_i = 0 \ .$$

Таким образом,

$$D(x*(t)) = \sigma^2 \{\frac{1}{n} + \frac{(t - t_{cp})^2}{\sum_{i=1}^{n} (t_i - t_{cp})^2}\} \quad .$$

Итак, оценка $x*(t)$ является несмещенной и асимптотически нормальной. Для ее практического использования необходимо уметь оценивать остаточную дисперсию $M(e_i^2) = \sigma^2$.

**Оценивание остаточной дисперсии**. В точках $t_k$ , $k = 1,2,...,n,$ имеются исходные значения зависимой переменной $x_k$ и восстановленные значения $x*(t_k)$. Рассмотрим остаточную сумму квадратов

$$SS = \sum_{i=1}^{n} (x*(t_i) - x(t_i))^2 = \sum_{i=1}^{n} \{(a* - a)(t_i - t_{cp}) + (b* - b) - e_i)^2.$$

В соответствии с формулами (5) и (6)

$$SS = \sum_{i=1}^{n} \{(t_i - t_{cp}) \sum_{j=1}^{n} c_j e_j + \tfrac{1}{n} \sum_{j=1}^{n} e_j - e_i\}^2 = \sum_{i=1}^{n} \{\sum_{j=1}^{n} \{c_j (t_i - t_{cp}) + \frac{1}{n}\} e_j - e_i\}^2 = \sum_{i=1}^{n} SS_i.$$

Найдем математическое ожидание каждого из слагаемых:

$$M(SS_i) = \sum_{j=1}^{n} \{c_j (t_i - t_{cp}) + \frac{1}{n}\}^2 \sigma^2 - 2\{c_i (t_i - t_{cp}) + \frac{1}{n}) \sigma^2 + \sigma^2.$$

Из сделанных ранее предположений вытекает, что при $n \to \infty$ имеем $M(SS_i) \to \sigma^2, i = 1,2,...,n,$ следовательно, по закону больших чисел статистика SS/n является состоятельной оценкой остаточной дисперсии $\sigma^2$.

Получением состоятельной оценкой остаточной дисперсии завершается последовательность задач, связанных с рассматриваемым простейшим вариантом метода наименьших квадратов. Не представляет труда выписывание верхней и нижней границ для прогностической функции:

$$x_{верх}(t) = a*(t - t_{ср}) + b* + \delta(t), \quad x_{нижн}(t) = a*(t - t_{ср}) + b* - \delta(t),$$

где погрешность $\delta(t)$ имеет вид

$$\delta(t) = U(p)\sigma*\{\frac{1}{n} + \frac{(t - t_{ср})^2}{\sum_{i=1}^{n}(t_i - t_{ср})^2}\}^{1/2}, \quad \sigma* = \left(\frac{SS}{n}\right)^{1/2}.$$

Здесь $p$ - доверительная вероятность, $U(p)$, как и в главе 3.1 - квантиль нормального распределения порядка $(1+p)/2$, т.е.

$$\Phi(U(p)) = \frac{1+p}{2}.$$

При $p$= 0,95 (наиболее применяемое значение) имеем $U(p)$ = 1,96. Для других доверительных вероятностей соответствующие значения квантилей можно найти в статистических таблицах (см., например, наилучшее в этой сфере издание [4]).

**Сравнение параметрического и непараметрического подходов.** Во многих литературных источниках рассматривается параметрическая вероятностная модель метода наименьших квадратов. В ней предполагается, что погрешности имеют нормальное распределение. Это предположение позволяет математически строго получить ряд выводов. Так, распределения статистик вычисляются точно, а не в асимптотике, соответственно вместо квантилей нормального распределения используются квантили распределения Стьюдента, а остаточная сумма квадратов $SS$ делится не на $n$, а на $(n-2)$. Ясно, что при росте объема данных различия стираются.

Рассмотренный выше непараметрический подход не использует нереалистическое предположение о нормальности погрешностей (см. главу 2.1). Платой за это является асимптотический характер результатов. В случае простейшей модели метода наименьших квадратов оба подхода дают практически совпадающие рекомендации. Это не всегда так, не всегда два подхода дают близкие результаты. Напомним, что в задаче обнаружения выбросов методы, опирающиеся на нормальное распределение, нельзя считать обоснованными, и обнаружено было это их свойство с помощью непараметрического подхода (см. главу 2.3).

**Общие принципы.** Кратко сформулируем несколько общих принципов построения, описания и использования методов прикладной статистики. Во-первых, должны быть четко сформулированы исходные предпосылки, т.е. полностью описана используемая вероятностно-статистическая модель. Во-вторых, не следует принимать предпосылки, которые редко выполняются на практике. В-третьих, алгоритмы расчетов должны быть корректны с точки зрения математико-статистической теории. В-четвертых, алгоритмы должны давать полезные для практики выводы.

Применительно к задаче восстановления зависимостей это означает, что целесообразно применять непараметрический подход, что и сделано выше. Однако предположение нормальности, хотя и очень сильно сужает возможности применения, с чисто математической точки зрения позволяет продвинуться дальше. Поэтому для первоначального изучения ситуации, так сказать, "в лабораторных условиях", нормальная модель может оказаться полезной.

**Пример оценивания по методу наименьших квадратов.** Пусть даны $n$ = 6 пар чисел $(t_k, x_k)$, $k$ = 1,2,…,6, представленных во втором и третьем столбцах табл.2. В соответствии с формулами (2) и (4) выше для вычисления оценок метода наименьших квадратов достаточно найти суммы выражений, представленных во втором, третьем, четвертом и пятом столбцах табл.2.

Таблица 2.

Расчет по методу наименьших квадратов
при восстановлении линейной функции одной переменной

| $i$ | $t_i$ | $x_i$ | $t_i^2$ | $t_i x_i$ | $a^* t_i$ | $\hat{x}_i$ | $x_i - \hat{x}_i$ | $(x_i - \hat{x}_i)^2$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 12 | 1 | 12 | 3,14 | 12,17 | -0,17 | 0,03 |
| 2 | 3 | 20 | 9 | 60 | 9,42 | 18,45 | 1,55 | 2,40 |
| 3 | 4 | 20 | 16 | 80 | 12,56 | 21,59 | -1,59 | 2,53 |
| 4 | 7 | 32 | 49 | 224 | 21,98 | 31,01 | 0,99 | 0,98 |
| 5 | 9 | 35 | 81 | 315 | 28,26 | 37,29 | -2,29 | 5,24 |
| 6 | 10 | 42 | 100 | 420 | 31,40 | 40,43 | 1,57 | 2,46 |
| Σ | 34 | 161 | 256 | 1111 | | | 0,06 | 13,64 |
| $\frac{\Sigma}{n}$ | 5,67 | 26,83 | 42,67 | 185,17 | | | | |

В соответствии с формулой (2) $b^*$=26,83, а согласно формуле (4)

$$a^* = \frac{1111 - \frac{1}{6}161 \times 34}{256 - \frac{1}{6}(34)^2} = \frac{1111 - 912,33}{256 - 192,67} = \frac{198,67}{63,33} = 3,14.$$

Следовательно, прогностическая формула имеет вид

$$x^*(t) = 3,14(t - 5,67) + 26,83 = 3,14t - 3,14 \times 5,67 + 26,83 =$$
$$= 3,14t - 17,80 + 26,83 = 3,14t + 9,03.$$

Следующий этап анализа данных - оценка точности приближения функции методом наименьших квадратов. Сначала рассматриваются т.н. восстановленные значения

$$\hat{x}_i = x^*(t_i), \quad i = 1,2,...,n.$$

Это те значения, которые полученная в результате расчетов прогностическая функция принимает в тех точках, в которых известны истинные значения зависимой переменной $x_i$.

Вполне естественно сравнить восстановленные и истинные значения. Это и сделано в шестом - восьмом столбцах табл. 2. Для простоты расчетов в шестом столбце представлены произведения $a^* t_i$, седьмой отличается от шестого добавлением константы 9,03 и содержит восстановленные значения. Восьмой столбец - это разность третьего и седьмого.

Непосредственный анализ восьмого столбца табл.2 показывает, что содержащиеся в нем числа сравнительно невелики по величине по сравнению с третьим столбцом (на порядок меньше по величине). Кроме того, знаки "+" и "-" чередуются. Эти два признака свидетельствуют о правильности расчетов. При использовании метода наименьших квадратов знаки не всегда чередуются. Однако если сначала идут только плюсы, а потом только минусы (или наоборот, сначала только минусы, а потом только плюсы), то это верный показатель того, что в вычислениях допущена ошибка.

Верно следующее утверждение.

**Теорема.**

$$\sum_{i=1}^{n} (x_i - \hat{x}_i) = 0.$$

Доказательство этой теоремы оставляем читателю в качестве упражнения.

Однако сумма по восьмому столбцу дает 0,06, а не 0. Незначительное отличие от 0 связано с ошибками округления при вычислениях. Близость суммы значений зависимой переменной и суммы восстановленных значений - практический критерий правильности расчетов.

В последнем девятом столбце табл.2 приведены квадраты значений из восьмого столбца. Их сумма - это остаточная сумма квадратов $SS$ = 13,64. В соответствии со сказанным выше оценками дисперсии погрешностей и их среднего квадратического отклонения являются

$$(\sigma^2)^* = \frac{SS}{n} = \frac{13,4}{6} = 2,27; \quad \sigma^* = \sqrt{\frac{SS}{n}} = \sqrt{\frac{13,4}{6}} = 1,49.$$

Рассмотрим распределения оценок параметров. Оценка $b^*$ имеет асимптотически нормальное распределение с математическим ожиданием $b$ и дисперсией, которая оценивается

как 2,27/6=0,38 (здесь считаем, что 6 - "достаточно большое" число, что, конечно, можно оспаривать). Оценкой среднего квадратического отклонения является 0,615. Следовательно, при доверительной вероятности 0,95 доверительный интервал для параметра *b* имеет вид (26,83 - 1,96·0,615; 26,83 + 1,96·0,615) = (25,625; 28,035).

В формулах для дисперсий участвует величина

$$\sum_{i=1}^{n}(t_i - t_{cp})^2 = \sum_{i=1}^{n}(t_i^2 - 2t_i t_{cp} + t_{cp}^2) = \sum_{i=1}^{n} t_i^2 - 2t_{cp}\sum_{u=1}^{n} t_i + nt_{cp}^2 = \sum_{i=1}^{n} t_i^2 - nt_{cp}^2.$$

Подставив численные значения, получаем, что

$$\sum_{i=1}^{n} t_i^2 - nt_{cp}^2 = 256 - 6(5{,}67)^2 = 63{,}1.$$

Дисперсия для оценки *a** коэффициента при линейном члене прогностической функции оценивается как 2,27/63,1=0,036, а среднее квадратическое отклонение - как 0,19. Следовательно, при доверительной вероятности 0,95 доверительный интервал для параметра *a* имеет вид (3,14 - 1,96·0,19; 3,14 + 1,96·0,19) = (2,77; 3,51).

Прогностическая формула с учетом погрешности имеет вид (при доверительной вероятности 0,95)

$$x^*(t) = 3{,}14t + 9{,}03 \pm 1{,}96 \times 1{,}49\sqrt{\frac{1}{6} + \frac{(t-5{,}67)^2}{63{,}1}}.$$

В этой записи сохранено происхождение различных составляющих. Упростим:

$$x^*(t) = 3{,}14t + 9{,}03 \pm 2{,}92\sqrt{\frac{1}{6} + \frac{(t-5{,}67)^2}{63{,}1}}.$$

Например, при $t = 12$ эта формула дает

$$x^*(12) = 46{,}71 \pm 2{,}615.$$

Следовательно, нижняя доверительная граница - это 44,095, а верхняя доверительная граница - это 49,325.

Насколько далеко можно прогнозировать? Обычный ответ таков - до тех пор, пока сохраняется тот стабильный комплекс условий, при котором справедлива рассматриваемая зависимость. Изобретатель метода наименьших квадратов Карл Гаусс исходил из задачи восстановления орбиты астероида (малой планеты) Церера. Движение подобных небесных тел может быть рассчитано на сотни лет. А вот параметры комет (например, срок возвращения) не поддаются столь точному расчету, поскольку за время пребывания в окрестности Солнца сильно меняется масса кометы. В социально-экономической области горизонты надежного прогнозирования еще менее определены. В частности, они сильно зависят от решений центральной власти.

Чтобы выявить роль погрешностей в прогностической формуле, рассмотрим формальный предельный переход $t \to \infty$. Тогда слагаемые 9,03; 1/6; 5,67 становятся бесконечно малыми, и

$$x^*(t) \approx 3{,}14t \pm \frac{2{,}92}{\sqrt{63{,}1}}t = (3{,}14 \pm 0{,}37)t.$$

Таким образом, погрешности составляют около

$$\frac{100 \times 0{,}37}{3{,}14}\% = 11{,}8\%$$

от тренда (математического ожидания) прогностической функции. В социально-экономических исследованиях подобные погрешности считаются вполне приемлемыми.

### 3.2.3. Основы линейного регрессионного анализа

Метод наименьших квадратов, рассмотренный в простейшем случае, допускает различные обобщения. Например, метод наименьших квадратов дает алгоритм расчетов, если исходные данные – по-прежнему набор *n* пар чисел ($t_k$ , $x_k$), $k = 1,2,\ldots,n$, где $t_k$ – независимая переменная (например, время), а $x_k$ – зависимая (например, индекс инфляции), а восстанавливать надо не линейную зависимость, а квадратическую:

$$x(t) = at^2 + bt + c.$$

Следует рассмотреть функцию трех переменных

$$f(a,b,c) = \sum_{k=1}^{n} (x_k - at_k^2 - bt_k - c)^2.$$

Оценки метода наименьших квадратов - это такие значения параметров *a*, b** и *c**, при которых функция *f(a,b,c)* достигает минимума по всем значениям аргументов. Чтобы найти эти оценки, надо вычислить частные производные от функции *f(a,b,c)* по аргументам *a, b* и *c,* приравнять их 0, затем из полученных уравнений найти оценки: Имеем:

$$\frac{\partial f(a,b,c)}{\partial a} = \sum_{k=1}^{n} \frac{\partial}{\partial a}(x_k - at_k^2 - bt_k - c)^2 = \sum_{k=1}^{n} 2(-t_k^2)(x_k - at_k^2 - bt_k - c)^2.$$

Приравнивая частную производную к 0, получаем линейное уравнение относительно трех неизвестных параметров *a,b,c:*

$$a\sum_{k=1}^{n} t_k^4 + b\sum_{k=1}^{n} t_k^3 + c\sum_{k=1}^{n} t_k^2 = \sum_{k=1}^{n} t_k^2 x_k.$$

Приравнивая частную производную по параметру *b* к 0, аналогичным образом получаем уравнение

$$a\sum_{k=1}^{n} t_k^3 + b\sum_{k=1}^{n} t_k^2 + c\sum_{k=1}^{n} t_k = \sum_{k=1}^{n} t_k x_k.$$

Наконец, приравнивая частную производную по параметру *c* к 0, получаем уравнение

$$a\sum_{k=1}^{n} t_k^2 + b\sum_{k=1}^{n} t_k + cn = \sum_{k=1}^{n} x_k.$$

Решая систему трех уравнений с тремя неизвестными, находим оценки метода наименьших квадратов.

Другие задачи, рассмотренные в предыдущем подразделе (доверительные границы для параметров и прогностической функции и др.), также могут быть решены. Соответствующие алгоритмы более громоздки. Для их записи полезен аппарат матричной алгебры (см., например, одну из лучших в этой области монографий [5]). Для реальных расчетов используют соответствующие компьютерные программы.

Раздел прикладной статистики, посвященный восстановлению зависимостей, называется регрессионным анализом. Термин «линейный регрессионный анализ» используют, когда рассматриваемая функция линейно зависит от оцениваемых параметров (от независимых переменных зависимость может быть произвольной). Теория оценивания неизвестных параметров хорошо развита именно в случае линейного регрессионного анализа. Если же линейности нет и нельзя перейти к линейной задаче, то, как правило, хороших свойств от оценок ожидать не приходится.

Продемонстрируем подходы в случае зависимостей различного вида. Если зависимость имеет вид многочлена (полинома)

$$x(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + ... + a_m t^m,$$

то коэффициенты многочлена могут быть найдены путем минимизации функции

$$f(a_0, a_1, a_2, a_3, ..., a_m) = \sum_{k=1}^{n} (x_k - a_0 - a_1 t_k - a_2 t_k^2 - a_3 t_k^3 - ... - a_m t_k^m)^2.$$

Функция от *t* не обязательно должна быть многочленом. Можно, например, добавить периодическую составляющую, соответствующую сезонным колебаниям. Хорошо известно, например, что инфляция (рост потребительских цен) имеет четко выраженный годовой цикл. А именно, в среднем цены быстрее всего растут зимой, в декабре - январе, а медленнее всего (иногда в среднем даже падают) летом, в июле - августе. Пусть для определенности

$$x(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + ... + a_m t^m + A\sin Bt,$$

тогда неизвестные параметры могут быть найдены путем минимизации функции

$$f(a_0, a_1, a_2, a_3, ..., a_m, A, B) = \sum_{k=1}^{n} (x_k - a_0 - a_1 t_k - a_2 t_k^2 - a_3 t_k^3 - ... - a_m t_k^m - A\sin Bt_k)^2.$$

Пусть *I*(*t*) -индекс инфляции в момент *t*. Принцип стабильности условий приводит к гипотезе о постоянстве темпов роста средних цен, т.е. индекса инфляции. Таким образом, естественная модель для индекса инфляции - это

$$I(t) = Ae^{Bt}.$$

Эта модель не является линейной, метод наименьших квадратов непосредственно применять нельзя. Однако если прологарифмировать обе части предыдущего равенства:

$$\ln I(t) = \ln A + Bt,$$

то получим линейную зависимость, рассмотренную выше.

Независимых переменных может быть не одна, а несколько. Пусть, например, по исходным данным $(x_k, y_k, z_k), k = 1,2,...,n,$ требуется оценить неизвестные параметры *a* и *b* в зависимости

$$z = ax + by + \varepsilon,$$

где $\varepsilon$ - погрешность. Это можно сделать, минимизируя функцию

$$f(a,b) = \sum_{k=1}^{n} (z_k - ax_k - by_k)^2.$$

Зависимость от *x* и *y* не обязательно должна быть линейной. Предположим, что из каких-то соображений известно, что зависимость должна иметь вид

$$z = ax + by + cx^2 y + dxy + ey^3 + \varepsilon,$$

тогда для оценки пяти параметров необходимо минимизировать функцию

$$f(a,b,c,d,e) = \sum_{k=1}^{n} (z_k - ax_k - by_k - cx_k^2 y_k - dx_k y_k - ey_k^3)^2.$$

Более подробно рассмотрим пример из микроэкономики. В одной из оптимизационных моделей поведения фирмы используется т.н. производственная функция *f*(*K*,*L*), задающая объем выпуска в зависимости от затрат капитала *K* и труда *L*. В качестве конкретного вида производственной функции часто используется так называемая функция Кобба-Дугласа

$$f(K,L) = K^{\alpha} L^{\beta}.$$

Однако откуда взять значения параметров $\alpha$ и $\beta$? Естественно предположить, что они - одни и те же для предприятий отрасли. Поэтому целесообразно собрать информацию $(f_k, K_k, L_k), k = 1,2,...,n,$ где $f_k$ - объем выпуска на *k*-ом предприятии, $K_k$- объем затрат капитала на *k*-ом предприятии, $L_k$ - объем затрат труда на *k*-ом предприятии (в кратком изложении не пытаемся дать точных определений используемым понятиям из экономики предприятия). По собранной информации естественно попытаться оценить параметры $\alpha$ и $\beta$. Но они входят в зависимость нелинейно, поэтому сразу применить метод наименьших квадратов нельзя. Помогает логарифмирование:

$$\ln f(K,L) = \alpha \ln K + \beta \ln L.$$

Следовательно, целесообразно сделать *замену переменных*

$$x_k = \ln K_k, y_k = \ln L_k, z_k = \ln f_k, k = 1,2,3,...,n,$$

а затем находить оценки параметров $\alpha$ и $\beta$, минимизируя функцию

$$g(\alpha, \beta) = \sum_{k=1}^{n} (z_k - \alpha x_k - \beta y_k)^2.$$

Найдем частные производные:

$$\frac{\partial g(\alpha,\beta)}{\partial \alpha} = \sum_{k=1}^{n} 2(z_k - \alpha x_k - \beta y_k)(-x_k),$$

$$\frac{\partial g(\alpha,\beta)}{\partial \beta} = \sum_{k=1}^{n} 2(z_k - \alpha x_k - \beta y_k)(-y_k).$$

Приравняем частные производные к 0, сократим на 2, раскроем скобки, перенесем свободные члены вправо. Получим систему двух линейных уравнений с двумя неизвестными:

$$\alpha\sum_{k=1}^{n}x_k^2+\beta\sum_{k=1}^{n}x_k y_k=\sum_{k=1}^{n}x_k z_k,$$

$$\alpha\sum_{k=1}^{n}x_k y_k+\beta\sum_{k=1}^{n}y_л^2=\sum_{k=1}^{n}y_k z_k.$$

Таким образом, для вычисления оценок метода наименьших квадратов необходимо найти пять сумм

$$\sum_{k=1}^{n}x_k^2,\quad \sum_{k=1}^{n}x_k y_k,\quad \sum_{k=1}^{n}y_k^2,\quad \sum_{k=1}^{n}x_k z_k,\quad \sum_{k=1}^{n}y_k z_k.$$

Для упорядочения расчета этих сумм может быть использована таблица типа той, что применялась выше. Отметим, что рассмотренная в предыдущем подразделе постановка переходит в разбираемую сейчас при $y_k=1,\quad k=1,2,...,n.$

Подходящая замена переменных во многих случаях позволяет перейти к линейной зависимости. Например, если

$$y=\frac{1}{a+bx},$$

то замена $z=1/y$ приводит к линейной зависимости $z = a + bx$. Если $y=(a+bx)^2$, то замена $z=\sqrt{y}$ приводит к линейной зависимости $z = a + bx$.

**Основной показатель качества регрессионной модели.** Одни и те же данные можно обрабатывать различными способами. На первый взгляд, показателем отклонений данных от модели может служить остаточная сумма квадратов *SS*. Чем этот показатель меньше, тем приближение лучше, значит, и модель лучше описывает реальные данные. Однако это рассуждение годится только для моделей с одинаковым числом параметров. Ведь если добавляется новый параметр, по которому можно минимизировать, то и минимум, как правило, оказывается меньше.

В качестве основного показателя качества регрессионной модели используют оценку остаточной дисперсии

$$\hat{\sigma}^2(m)=\frac{SS}{n-m},$$

скорректированную на число $m$ параметров, оцениваемых по наблюдаемым данным. В случае задачи восстановления линейной функции одной переменной, рассмотренной в предыдущем подразделе, оценка остаточной дисперсии имеет вид

$$\hat{\sigma}^2=\frac{SS}{n-2},$$

поскольку число оцениваемых параметров $m$=2.

Почему эта формула отличается от приведенной в предыдущем подразделе? Там в знаменателе $n$, а здесь - ($n$-2). Дело в том, что там была рассмотрена непараметрическая теория при большом объеме данных (при $n\to\infty$). А при безграничном возрастании $n$ разница между $n$ и ($n$-2) сходит на нет.

Однако *при подборе вида модели* знаменатель дроби, оценивающей остаточную дисперсию, приходится корректировать на число параметров. Если этого не делать, то придется заключить, что всегда многочлен второй степени лучше соответствует данным, чем линейная функция, многочлен третьей степени лучше приближает исходные данные, чем многочлен второй степени, и т.д. В конце концов доходим до многочлена степени ($n$-1) с $n$ коэффициентами, который проходит через все заданные точки. Но его прогностические возможности, скорее всего, существенно меньше, чем у линейной функции. *Излишнее усложнение статистических моделей вредно.*

Типовое поведение скорректированной оценки остаточной дисперсии

$$v(m)=\hat{\sigma}^2(m)$$

в зависимости от параметра $m$ в случае расширяющейся системы моделей выглядит так. Сначала наблюдаем заметное убывание. Затем оценка остаточной дисперсии колеблется около некоторой константы (теоретического значения дисперсии погрешности).

Поясним ситуацию на примере модели восстановления зависимости, выраженной многочленом:

$$x(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \ldots + a_m t^m.$$

Пусть эта модель справедлива при $m = m_0$. При $m < m_0$ в скорректированной оценке остаточной дисперсии учитываются не только погрешности измерений, но и соответствующие (старшие) члены многочлена (предполагаем, что коэффициенты при них отличны от 0). При $m \geq m_0$ имеем

$$\lim_{n\to\infty} v(m) = \sigma^2.$$

Следовательно, скорректированная оценка остаточной дисперсии будет колебаться около указанного предела. Поэтому в качестве оценки неизвестной статистику степени многочлена (полинома) можно использовать первый локальный минимум скорректированной оценки остаточной дисперсии, т.е.

$$m^* = \min\{m : v(m-1) > v(m), \quad v(m) \leq v(m+1)\}.$$

В работе [6] найдено предельное распределение этой оценки степени многочлена.

**Теорема.** При справедливости некоторых условий регулярности

$$\lim_{n\to\infty} P(m^* < m_0) = 0, \quad \lim_{n\to\infty} P(m^* = m_0 + u) = \lambda(1-\lambda)^u, \ u = 0,1,2,....,$$

где

$$\lambda = \Phi(1) - \Phi(-1) = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} \exp\left\{-\frac{x^2}{2}\right\} dx \approx 0{,}68268.$$

Таким образом, предельное распределение оценки $m^*$ степени многочлена (полинома) является геометрическим. Это означает, в частности, что оценка не является состоятельной. При этом вероятность получить меньшее значение, чем истинное, исчезающе мала. Далее имеем:

$$P(m^* = m_0) \to 0{,}68268, \quad P(m^* = m_0 + 1) \to 0{,}68268(1 - 0{,}68268) = 0{,}21663,$$

$$P(m^* = m_0 + 2) \to 0{,}68268(1 - 0{,}68268)^2 = 0{,}068744,$$

$$P(m^* = m_0 + 3) \to 0{,}68268(1 - 0{,}68268)^3 = 0{,}021814...$$

Разработаны и иные методы оценивания неизвестной степени многочлена, например, путем многократного применения процедуры проверки адекватности регрессионной зависимости с помощью статистики Фишера (см. работу [7]). Предельное поведение оценок - таково же, как в приведенной выше теореме, только значение параметра $\lambda$ иное.

**Пример практического использования линейного регрессионного анализа.** Руководитель маркетинговой службы новгородского завода ГАРО А.А. Пивень применил его для построения математической модели рынка легковых подъемников. Требуется выявить факторы (показатели), оказывающие наибольшее влияние на объем продаж подъемников, найти зависимость объема продаж от этих факторов и использовать эту зависимость для прогнозирования объема продаж.

Зависимая переменная – объем продаж V, независимые переменные:

- грузоподъемность (X1),
- цена (X2)
- наличие напольной рамы (X3),
- наличие синхронизации (X4),
- количество двигателей (X5),
- суммарная мощность двигателей (X6),
- высота подхвата в нижнем положении (X7),
- максимальная высота подъема (X8),
- скорость подъема (X9),
- гарантийный срок (X10),
- срок службы (X11),

- время на рынке (X12),
- внешний вид (X13),
- срок поставки (X14),
- уровень сервисного обслуживания (X15),
- наличие системы смазки (X16),
- масса (X17).

Для восстановления зависимости использовалась линейная регрессионная модель. По результатам пошагового анализа из рассмотрения последовательно исключались независимые переменные (параметры подъемника), имеющие (в линейной модели) коэффициенты, незначимо отличающиеся от нуля, иными словами, мало отличающиеся в сравнении с их дисперсией. Для этого использовался пакет STATISTICA 6.0, конкретно модуль «Множественная регрессия» (*Multiple regression*).

В результате расчетов получена зависимость объема продаж подъемника П3-Т от 12 факторов:

$$V = - 1769.77 - 65.09\,X1 - 0.03X2 + 68.79X3 + 147.54X4 + 156.28X5 + 2.53X7 + 1.06X8 + 25.75X12 - 132.26X13 - 12.41X14 + 107.78X15 + 397X16 .$$

Влияние остальных пяти факторов оказалось незначимым.

Исходя из расчетов, прогнозное значение продаж подъемников на второй год продаж составит ориентировочно 1010 шт. С вероятностью 95% можно утверждать, что объем продаж будет лежать в границах [695, 1332] шт.

**Оценивание условного математического ожидания.** Рассмотрим общее понятие регрессии как условного математического ожидания. Пусть случайный вектор $(x(\omega), y(\omega))$ имеет плотность *p(x,y).* Как известно из любого курса теории вероятностей, плотность условного распределения $y(\omega)$ при условии $x(\omega) = x_0$ имеет вид

$$p(y \mid x) = p(y \mid x(\omega) = x_0) = \frac{p(x,y)}{\int\limits_{-\infty}^{+\infty} p(x,y)dy}.$$

Условное математическое ожидание, т.е. регрессионная зависимость *y* от *x*, имеет вид

$$f(x) = \int\limits_{-\infty}^{+\infty} yp(y \mid x)dy = \frac{\int\limits_{-\infty}^{+\infty} yp(x,y)dy}{\int\limits_{-\infty}^{+\infty} p(x,y)dy}.$$

Таким образом, для нахождения оценок регрессионной зависимости достаточно найти оценки совместной плотности распределения вероятности $p_n(x,y)$ такие, что

$$p_n(x,y) \to p(x,y)$$

при $n \to \infty$. Тогда непараметрическая оценка регрессионной зависимости

$$f_n(x) = \frac{\int\limits_{-\infty}^{+\infty} yp_n(x,y)dy}{\int\limits_{-\infty}^{+\infty} p_n(x,y)dy}$$

при $n \to \infty$ является состоятельной оценкой регрессии как условного математического ожидания

$$f_n(x) \to f(x).$$

Общий подход к построению непараметрических оценок плотности распределения вероятностей развит в главе 2.1 выше.

Регрессионному анализу (т.е. методам восстановления зависимостей) посвящена огромная литература. Он хорошо представлен в программных продуктах по анализу данных, особенно та его часть, которая связана с методом наименьших квадратов. Обзор современных методов и моделей дан в учебнике [6].

### 3.2.4. Основы теории классификации

При внедрении современных статистических методов в практику фундаментальных и прикладных научно-технических, социально-экономических, медицинских и иных исследований, при разработке соответствующих программных продуктов невозможно обойтись без классификации самих этих методов. Естественно исходить из вида обрабатываемых данных. В соответствии с современными воззрениями делим прикладную статистику на четыре области: - статистика случайных величин (одномерная статистика); многомерный статистический анализ; статистика временных рядов и случайных величин; статистика объектов нечисловой природы. В первой области элемент выборки - число, во второй - вектор, в третьей - функция, в четвертой - объект нечисловой природы.

Как известно, математический аппарат статистики объектов нечисловой природы базируется на использовании расстояний (мер близости, показателей различия) в пространствах таких объектов. Это вызвано отсутствием в таких пространствах операций суммирования, на которых основано большинство методов других областей статистики. Любые методы, использующие только расстояния (меры близости, показатели различия) между объектами, следует относить к статистике объектов нечисловой природы, поскольку такие методы могут работать с объектами произвольного пространства, если в нем задана метрика или ее аналоги. Таким образом, весьма многие методы прикладной статистики следует включать в статистику объектов нечисловой природы.

В настоящем пункте рассматривается важное направление прикладной статистики – математические методы классификации. Значительную их часть следовало бы отнести к статистике объектов нечисловой природы, а именно, методы классификации, основанные на расстояниях между объектами. Однако исторически теория классификации рассматривается в основном в рамках многомерного статистического анализа, поскольку многие ее методы используют специфику конечномерного евклидова пространства.

**Основные направления в математической теории классификации.** Какие научные исследования относить к этой теории? Исходя из потребностей специалиста, применяющего математические методы классификации, целесообразно принять, что сюда входят исследования, во-первых, отнесенные самими авторами к этой теории; во вторых, связанные с ней общностью тематики, хотя бы их авторы и не упоминали термин «классификация». Это предполагает ее сложную внутреннюю структуру.

В литературных источниках наряду с термином «классификация» в близких смыслах используются термины «группировка», «распознавание образов», «диагностика», «дискриминация», «сортировка» и др. Терминологический разнобой связан прежде всего с традициями научных кланов, к которым относятся авторы публикаций, а также с внутренним делением самой теории классификации.

В научных исследованиях по современной теории классификации можно выделить два относительно самостоятельных направления. Одно из них опирается на опыт таких наук, как биология, география, геология, и таких прикладных областей, как ведение классификаторов продукции и библиотечное дело. Типичные объекты рассмотрения - классификация химических элементов (таблица Д.И. Менделеева), биологическая систематика, универсальная десятичная классификация публикаций (УДК), классификатор товаров на основе штрих-кодов.

Другое направление опирается на опыт технических исследований, экономики, маркетинговых исследований, социологии, медицины. Типичные задачи - техническая и медицинская диагностика, а также, например, разбиение на группы отраслей промышленности, тесно связанных между собой, выделение групп однородной продукции. Обычно используются такие термины, как «распознавание образов» или «дискриминантный анализ». Это направление обычно опирается на математические модели; для проведения расчетов интенсивно используется ЭВМ. Однако относить его к математике столь же нецелесообразно, как астрономию или квантовую механику. Рассматриваемые математические модели можно и нужно изучать на формальном уровне, и такие исследования проводятся. Но направление в целом сконцентрировано на решении конкретных задач прикладных областей и вносит вклад в технические или экономические науки, медицину, социологию, но, как правило, не в

математику. Использование математических методов как инструмента исследования нельзя относить к чистой математике.

В 60-х годах XX века внутри прикладной статистики достаточно четко оформилась область, посвященная методам классификации. Несколько модифицируя формулировки М. Дж. Кендалла и А. Стьюарта 1966 г. (см. русский перевод [8, с.437]), в теории классификации выделим три подобласти: дискриминация (дискриминантный анализ), кластеризация (кластер-анализ), группировка. Опишем эти подобласти.

В дискриминантном анализе классы предполагаются заданными - плотностями вероятностей или обучающими выборками. Задача состоит в том, чтобы вновь поступающий объект отнести в один из этих классов. У понятия «дискриминация» имеется много синонимов: диагностика, распознавание образов с учителем, автоматическая классификация с учителем, статистическая классификация и т.д.

При кластеризации и группировке целью является выявление и выделение классов. Синонимы: построение классификации, распознавание образов без учителя, автоматическая классификация без учителя, типология, таксономия и др. Задача кластер-анализа состоит в выяснении по эмпирическим данным, насколько элементы "группируются" или распадаются на изолированные "скопления", "кластеры" (от *cluster* (англ.) - гроздь, скопление). Иными словами, задача - выявление естественного разбиения на классы, свободного от субъективизма исследователя, а цель - выделение групп однородных объектов, сходных между собой, при резком отличии этих групп друг от друга.

При группировке, наоборот, «мы хотим разбить элементы на группы независимо от того, естественны ли границы разбиения или нет» [8, с.437]. Цель по-прежнему состоит в выявлении групп однородных объектов, сходных между собой (как в кластер-анализе), однако «соседние» группы могут не иметь резких различий (в отличие от кластер-анализа). Границы между группами условны, не являются естественными, зависят от субъективизма исследователя. Аналогично при лесоустройстве проведение просек (границ участков) зависит от специалистов лесного ведомства, а не от свойств леса.

Задачи кластеризации и группировки принципиально различны, хотя для их решения могут применяться одни и те же алгоритмы. Важная для практической деятельности проблема состоит в том, чтобы понять, разрешима ли задача кластер-анализа для конкретных данных или возможна только их группировка, поскольку совокупность объектов достаточно однородна и не разбивается на резко разделяющиеся между собой кластеры.

Как правило, в математических задачах кластеризации и группировки основное - выбор метрики, расстояния между объектами, меры близости, сходства, различия. Хорошо известно, что для любого заданного разбиения объектов на группы и любого $\varepsilon > 0$ можно указать метрику такую, что расстояния между объектами из одной группы будут меньше $\varepsilon$, а между объектами из разных групп - больше $1/\varepsilon$. Тогда любой разумный алгоритм кластеризации даст именно заданное разбиение.

Понимание и обсуждение постановок задач осложняется использованием одного и того же термина в разных смыслах. Термином "классификация" (и термином "диагностика") обозначают, по крайней мере, три разные вещи: процедуру построения классификации (и выделение классов, используемых при диагностике), построенную классификацию (систему выделенных классов) и процедуру ее использования (правила отнесения вновь поступающего объекта к одному из ранее выделенных классов). Другими словами, имеем естественную триаду: построение – изучение – использование классификации.

Как уже отмечалось, для построения системы диагностических классов используют разнообразные методы кластерного анализа и группировки объектов. Наименее известен второй член триады (отсутствующий у Кендалла и Стьюарта [8]) – изучение отношений эквивалентности, полученных в результате построения системы диагностических классов. Статистический анализ полученных, в частности экспертами, отношений эквивалентности - часть статистики бинарных отношений и тем самым - статистики объектов нечисловой природы (см. главу 3.4).

Диагностика в узком смысле слова (процедура использования классификации, т.е. отнесения вновь поступающего объекта к одному из выделенных ранее классов) - предмет

дискриминантного анализа. Отметим, что с точки зрения статистики объектов нечисловой природы дискриминантный анализ является частным случаем общей схемы регрессионного анализа, соответствующим ситуации, когда зависимая переменная принимает конечное число значений, а именно - номера классов, а вместо квадрата разности стоит функция потерь от неправильной классификации. Однако есть ряд специфических постановок, выделяющих задачи диагностики среди всех регрессионных задач.

**О построении диагностических правил.** Начнем с краткого обсуждения одного распространенного заблуждения. Иногда рекомендуют сначала построить систему диагностических классов, а потом в каждом диагностическом классе отдельно проводить регрессионный анализ (в классическом смысле) или применять иные методы многомерного статистического анализа. Однако обычно забывают, что при этом нельзя опираться на вероятностную модель многомерного нормального распределения, так как распределение результатов наблюдений, попавших в определенный кластер, будет отнюдь не нормальным, а усеченным нормальным (усечение определяется границами кластера).

Процедуры построения диагностических правил делятся на вероятностные и детерминированные. К первым относятся так называемые задачи расщепления смесей. В них предполагается, что распределение вновь поступающего случайного элемента является смесью вероятностных законов, соответствующих диагностическим классам. Как и при выборе степени полинома в регрессии (см. предыдущий подраздел), при анализе реальных социально-экономических данных встает вопрос об оценке числа элементов смеси, т.е. числа диагностических классов. Были изучены результаты применения обычно рекомендуемого критерия Уилкса для оценки числа элементов смеси. Оказалось (см. статью [9]), что оценка с помощью критерия Уилкса не является состоятельной, асимптотическое распределение этой оценки – геометрическое, как и в случае задачи восстановления зависимости в регрессионном анализе. Итак, продемонстрирована несостоятельность обычно используемых оценок. Для получения состоятельных оценок достаточно связать уровень значимости в критерии Уилкса с объемом выборки, как это было предложено и для задач регрессии [7].

Как уже отмечалось, задачи построения системы диагностических классов целесообразно разбить на два типа: с четко разделенными кластерами (задачи кластер-анализа) и с условными границами, непрерывно переходящими друг в друга классами (задачи группировки). Такое деление полезно, хотя в обоих случаях могут применяться одинаковые алгоритмы. Сколько же существует алгоритмов построения системы диагностических правил? Иногда называют то или иное число. На самом же деле их бесконечно много, в чем нетрудно убедиться.

Действительно, рассмотрим один определенный алгоритм - алгоритм средней связи. Он основан на использовании некоторой меры близости $d(x,y)$ между объектами $x$ и $y$. Как он работает? На первом шаге каждый объект рассматривается как отдельный кластер. На каждом следующем шаге объединяются две ближайших кластера. Расстояние между объектами рассчитывается как средняя связь (отсюда и название алгоритма), т.е. как среднее арифметическое расстояний между парами объектов, один из которых входит в первый кластер, а другой - во второй. В конце концов все объекты объединяются вместе, и результат работы алгоритма представляет собой дерево последовательных объединений (в терминах теории графов), или "Дендрограмму". Из нее можно выделить кластеры разными способами. Один подход - исходя из заданного числа кластеров. Другой - из соображений предметной области. Третий - исходя из устойчивости (если разбиение долго не менялось при возрастании порога объединения – значит, оно отражает реальность). И т.д.

К алгоритму средней связи естественно сразу добавить алгоритм ближайшего соседа (когда расстоянием между кластерами называется минимальное из расстояний между парами объектов, один из которых входит в первый кластер, а другой - во второй). А также и алгоритм дальнего соседа (когда расстоянием между кластерами называется максимальное из расстояний между парами объектов, один из которых входит в первый кластер, а другой - во второй).

Каждый из трех описанных алгоритмов (средней связи, ближайшего соседа, дальнего соседа), как легко проверить, порождает бесконечное (континуальное) семейство алгоритмов кластер-анализа. Дело в том, что величина $d^a(x,y)$, $a>0$, также является мерой близости между $x$ и $y$ и порождает новый алгоритм. Если параметр $a$ пробегает отрезок, то получается бесконечно много алгоритмов классификации.

Каким из них пользоваться при обработке данных? Дело осложняется тем, что практически в любом пространстве данных мер близости различных видов существует весьма много. Именно в связи с обсуждаемой проблемой следует указать на принципиальное различие между кластер-анализом и задачами группировки.

Если классы реальны, естественны, существуют на самом деле, четко отделены друг от друга, то любой алгоритм кластер-анализа их выделит. Следовательно, *в качестве критерия естественности классификации следует рассматривать устойчивость относительно выбора алгоритма кластер-анализа.*

Проверить устойчивость можно, применив к данным несколько подходов, например, столь непохожие алгоритмы, как «ближнего соседа» и «дальнего соседа». Если полученные результаты содержательно близки, то они адекватны действительности. В противном случае следует предположить, что естественной классификации не существует, задача кластер-анализа не имеет решения, и можно проводить только группировку.

Как уже отмечалось, часто применяется т.н. агломеративный иерархический алгоритм "Дендрограмма", в котором вначале все элементы рассматриваются как отдельные кластеры, а затем на каждом шагу объединяются два наиболее близких кластера. Для работы «Дендрограммы» необходимо задать правило вычисления расстояния между кластерами. Оно вычисляется через расстояние $d(x,y)$ между элементами $x$ и $y$. Поскольку $d^a(x,y)$ при $0<a<1$ также расстояние, то, как правило, существует бесконечно много различных вариантов этого алгоритма. Представим себе, что они применяются для обработки одних и тех же реальных данных. Если при всех $a$ получается одинаковое разбиение элементов на кластеры, т.е. результат работы алгоритма устойчив по отношению к изменению $a$ (в смысле общей схемы устойчивости, рассмотренной в главе 1.4), то имеем «естественную» классификацию. В противном случае результат зависит от субъективно выбранного исследователем параметра $a$, т.е. задача кластер-анализа неразрешима (предполагаем, что выбор $a$ нельзя специально обосновать). Задача группировки в этой ситуации имеет много решений. Из них можно выбрать одно по дополнительным критериям.

Следовательно, получаем эвристический критерий: если решение задачи кластер-анализа существует, то оно находится с помощью любого алгоритма. Целесообразно использовать наиболее простой.

**Проблема поиска естественной классификации.** Существуют различные точки зрения на эту проблему. Естественная классификация обычно противопоставляется искусственной. На Всесоюзной школе-семинаре «Использование математических методов в задачах классификации» (г. Пущино, 1986 г.), в частности, были высказаны мнения, что естественная классификация:

- закон природы;
- основана на глубоких закономерностях, тогда как искусственная классификация - на неглубоких;
- для конкретного индивида та, которая наиболее быстро вытекает из его тезауруса;
- удовлетворяет многим целям; цель искусственной классификации задает человек;
- классификация с точки зрения потребителя продукции;
- классификация, позволяющая делать прогнозы;
- имеет критерием устойчивость.

Приведенные высказывания уже дают представление о больших расхождениях в понимании «естественной классификации». Этот термин следует признать нечетким, как, впрочем, и многие другие термины, и профессиональные - социально-экономические, научно-технические, и используемые в обыденном языке. Нетрудно подробно обоснована нечеткость естественного языка и тот факт, что "мы мыслим нечетко", что, однако, не слишком мешает нам решать производственные и жизненные проблемы. Кажущееся рациональным требование выработать сначала строгие определения, а потом развивать науку - невыполнимо. Следовать ему - значит отвлекать силы от реальных задач. При системном подходе к теории классификации становится ясно, что строгие определения можно надеяться получить на последних этапах построения теории. Мы же сейчас находимся чаще всего на первых этапах. Поэтому, не давая определения понятиям «естественная классификация» и «естественная диагностика», обсудим, как проверить

на «естественность» классификацию (набор диагностических классов), полученную расчетным путем.

Можно выделить два критерия «естественности», по поводу которых имеется относительное согласие:

А. Естественная классификация должна быть реальной, соответствующей действительному миру, лишенной внесенного исследователем субъективизма;

Б. Естественная классификация должна быть важной или с научной точки зрения (давать возможность прогноза, предсказания новых свойств, сжатия информации и т.д.), или с практической.

Пусть классификация проводится на основе информации об объектах, представленной в виде матрицы «объект-признак» или матрицы попарных расстояний (мер близости). Пусть алгоритм классификации дал разбиение на кластеры. Как можно получить доводы в пользу естественности этой классификации? Например, уверенность в том, что она - закон природы, может появиться только в результате ее длительного изучения и практического применения. Это соображение относится и к другим из перечисленных выше критериев, в частности к Б (важности). Сосредоточимся на критерии А (реальности).

Понятие «реальности» кластера требует специального обсуждения. (оно начато в работе [9]). Рассмотрим существо различий между понятиями «классификация» и «группировка». Пусть, к примеру, необходимо деревья, растущие в определенной местности, разбить на группы находящихся рядом друг с другом. Ясна интуитивная разница между несколькими отдельными рощами, далеко отстоящими друг от друга и разделенными полями, и сплошным лесом, разбитым просеками на квадраты с целью лесоустройства.

Однако формально определить эту разницу столь же сложно, как определить понятие «куча зерен», чем занимались еще в Древней Греции. Ясно, что одно зерно не составляет кучи, два зерна не составляют кучи,… Если к тому, что не составляет кучи, добавить еще одно зерно, то куча не получится. Значит - по принципу математической индукции - никакое количество зерен не составляет кучи. Но ясно, что миллиард зерен - большая куча зерен - подсчитайте объем!.

Переформулируем сказанное в терминах "кластер-анализа" и "методов группировки". Выделенные с помощью первого подхода кластеры реальны, а потому могут рассматриваться как кандидаты в "естественные". Группировка дает "искусственные" классы, которые не могут быть "естественными".

Выборку из унимодального распределения можно, видимо, рассматривать как "естественный", "реальный" кластер. Применим к ней какой-либо алгоритм классификации ("средней связи", "ближайшего соседа" и т.п.). Он даст какое-то разбиение на классы, которые, разумеется, не являются "реальными", поскольку отражают прежде всего свойства алгоритма, а не исходных данных. Как отличить такую ситуацию от противоположной, когда имеются реальные кластеры и алгоритм классификации более или менее точно их выделяет? Как известно, "критерий истины – практика", но слишком много времени необходимо для применения подобного критерия. Поэтому представляет интерес критерий, оценивающий "реальность" выделяемых с помощью алгоритма классификации кластеров одновременно с его применением.

Такой показатель существует - это критерий устойчивости. Устойчивость - понятие широкое. Общая схема формулирования и изучения проблем устойчивости рассмотрена в главе 1.4. В частности, поскольку значения признаков всегда измеряются с погрешностями, то "реальное" разбиение должно быть устойчиво (т.е. не меняться или меняться слабо) при малых отклонениях исходных данных. Алгоритмов классификации существует бесконечно много, и "реальное" разбиение должно быть устойчиво по отношению к переходу к другому алгоритму. Другими словами, если "реальное" разбиение на классы возможно, то оно находится с помощью любого алгоритма автоматической классификация. Следовательно, критерием естественности классификации может служить совпадение результатов работы двух достаточно различающихся алгоритмов, например "ближайшего соседа" и "дальнего соседа".

Выше рассмотрены два типа "глобальных" критериев "естественности классификации", касающихся разбиения в целом. «Локальные» критерии относятся к отдельным кластерам. Простейшая постановка такова: достаточно ли однородны два кластера (две совокупности) для

их объединения:? Если объединение возможно, то кластеры не являются "естественными". Преимущество этой постановки в том, что она допускает применение статистических критериев однородности двух выборок. В одномерном случае (классификация по одному признаку) разработано большое число подобных критериев — Крамера-Уэлча, Смирнова, омега-квадрат (Лемана - Розенблатта), Вилкоксона, Ван-дер-Вардена, Лорда, Стьюдента и др. (см. главу 3.1 и справочник [4]). Имеются критерии и для многомерных данных. Для одного из видов объектов нечисловой природы - люсианов - статистические методы выделения "реальных" кластеров развиты в работе [10].

Что касается глобальных критериев, то для изучения устойчивости по отношению к малым отклонениям исходных данных естественно использовать метод статистических испытаний и проводить расчеты по "возмущенным" данным. Некоторые теоретические утверждения, касающиеся влияния «возмущений» на кластеры различных типов, получены в работе [9].

Опишем практический опыт реализации анализа устойчивости. Несколько алгоритмов классификации были применены к данным, полученным при проведении маркетинга образовательных услуг и приведенным в работе [11]. Для анализа данных были использованы широко известные алгоритмы "ближайшего соседа", "дальнего соседа" и алгоритм кластер-анализа из работы [12]. С содержательной точки зрения полученные разбиения отличались мало. Поэтому есть основания считать, что с помощью этих алгоритмов действительно выявлена «реальная» структура данных.

Идея устойчивости как критерия "реальности" иногда реализуется неадекватно. Так, для однопараметрических алгоритмов иногда предлагают выделять разбиения, которым соответствуют наибольшие интервалы устойчивости по параметру, т.е. наибольшие приращения параметра между очередными объединениями кластеров. Для данных работы [11] это предложение не дало полезных результатов - были получены различные разбиения: три алгоритма - три разбиения. И с теоретической точки зрения предложение этого специалиста несостоятельно. Покажем это.

Действительно, рассмотрим алгоритм "ближайшего соседа", использующий меру близости $d(x,y)$, и однопараметрическое семейство алгоритмов с мерой близости $d^a(x,y)$, $a>0$, также являющихся алгоритмами "ближайшего соседа". Тогда дендрограммы, полученные с помощью этих алгоритмов, совпадают при всех $a$, поскольку при их реализации происходит лишь сравнение мер близости между объектами. Другими словами, дендрограмма, полученная с помощью алгоритма «ближайшего соседа», является адекватной в порядковой шкале (измерения меры близости $d(x,y))$, т.е. сохраняется при любом строго возрастающем преобразовании этой меры. Однако выделенные по обсуждаемому методу "устойчивые разбиения" меняются. В частности, при достаточно большом $a$ "наиболее объективным" в соответствии с рассматриваемым предложением будет, как нетрудно показать, разбиение на два кластера! Таким образом, разбиение, выдвинутое им как "устойчивое", на самом деле оказывается весьма неустойчивым.

### 3.2.5. Статистические методы классификации

Рассмотрим с позиций прикладной статистики несколько конкретных вопросов теории классификации.

**Вероятностная теория кластер-анализа.** Как и для прочих статистических методов, свойства алгоритмов кластер-анализа необходимо изучать на вероятностных моделях. Это касается, например, условий естественного объединения двух кластеров.

Вероятностные постановки нужно применять, в частности, при перенесении результатов, полученных по выборке, на генеральную совокупность. Вероятностная теория кластер-анализа и методов группировки различна для исходных данных типа таблиц «объект Ч признак» и матриц сходства. Для первых параметрическая вероятностно-статистическая теория называется "расщеплением смесей". Непараметрическая теория основана на непараметрических оценках плотностей вероятностей и их мод. Основные результаты, связанные с непараметрическими оценками плотности, обсуждались в главе 2.1.

Если исходные данные - матрица сходства $\|d(x,y)\|$, то необходимо признать, что развитой вероятностно-статистической теории пока нет. Подходы к ее построению намечены в работе [9]. Одна из основных проблем - проверка "реальности" кластера, его объективного существования независимо от расчетов исследователя. Проблема "реальности" кластера давно обсуждается специалистами различных областей. Типичное рассуждение таково. Предположим, что результаты наблюдений можно рассматривать как выборку из некоторого распределения с монотонно убывающей плотностью при увеличении расстояния от некоторого центра. Примененный к подобным данным какой-либо алгоритм кластер-анализа порождает некоторое разбиение. Ясно, что оно - чисто формальное, поскольку выделенным таксонам (кластерам) не соответствуют никакие "реальные" классы. Другими словами, задача кластер-анализа не имеет решения, а алгоритм дает лишь группировку. При обработке реальных данных мы не знаем вида плотности. Проблема состоит в том, чтобы определить, каков результат работы алгоритма (реальные кластеры или формальные группы).

Частный случай этой проблемы - проверка обоснованности объединения двух кластеров, которые мы рассматриваем как два множества объектов, а именно, множества $\{a_1, a_2, \ldots, a_k\}$ и $\{b_1, b_2, \ldots, b_m\}$. Пусть, например, используется алгоритм типа "Дендрограмма". Естественной представляется следующая идея. Пусть есть две совокупности мер близости. Одна - меры близости между объектами, лежащими внутри одного кластера, т.е. $d(a_i,a_j)$, $1\le i<j\le k$, $d(b_\alpha,b_\beta)$, $1\le\alpha<\beta\le m$. Другая совокупность - меры близости между объектами, лежащими в разных кластерах, т.е. $d(a_i,b_\alpha)$, $1\le i\le k$, $1\le\alpha\le m$. Эти две совокупности мер близости предлагается рассматривать как независимые выборки и проверять гипотезу о совпадении их функций распределения. Если гипотеза не отвергается, объединение кластеров считается обоснованным; в противном случае - объединять нельзя, алгоритм прекращает работу.

В рассматриваемом подходе есть две некорректности (см. также работу [9, разд.4]). Во-первых, меры близости не являются независимыми случайными величинами. Во-вторых, не учитывается, что объединяются не заранее фиксированные кластеры (с детерминированным составом), а полученные в результате работы некоторого алгоритма, и их состав (в частности, количество элементов) оказывается случайным. От первой из этих некорректностей можно частично избавиться. Справедливо следующее утверждение.

*Теорема 1.* Пусть $a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_m$ - независимые одинаково распределенные случайные величины (со значениями в произвольном пространстве). Пусть случайная величина $d(a_1,a_2)$ имеет все моменты. Тогда при $k,m\to\infty$ распределение статистики

$$\frac{8\sqrt{3}U-3(k+m)(k+m-1)(k(k+1)+m(m+1))}{2(k+m)\sqrt{km(k^2+m^2)}}$$

(где $U$ - сумма рангов элементов первой выборки в объединенной выборке; первая выборка составлена из внутрикластерных расстояний (мер близости) $d(a_i,a_j)$, $1\le i<j\le k$, и $d(b_\alpha,b_\beta)$, $1\le\alpha<\beta\le m$, а вторая - из межкластерных расстояний $d(a_i,b_\alpha)$, $1\le i\le k$, $1\le\alpha\le m$) сходится к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1.

На основе теоремы 1 очевидным образом формулируется правило проверки обоснованности объединения двух кластеров. Другими словами, мы проверяем статистическую гипотезу, согласно которой объединение двух кластеров образует однородную совокупность. Если величина $U$ слишком мала, статистическая гипотеза однородности отклоняется (на заданном уровне значимости), и возможность объединения отбрасывается. Таким образом, хотя расстояния между объектами в кластерах зависимы, но эта зависимость слаба, и доказана математическая теорема о допустимости применения критерия Вилкоксона для проверки возможности объединения кластеров.

**О вычислительной сходимости алгоритмов кластер-анализа.** Алгоритмы кластер-анализа и группировки зачастую являются итерационными. Например, формулируется правило улучшения решения задачи кластер-анализа шаг за шагом, но момент остановки вычислений не обсуждается. Примером является известный алгоритм "Форель", в котором постепенно улучшается положение центра кластера. В этом алгоритме на каждом шаге строится шар определенного заранее радиуса, выделяются элементы кластеризуемой совокупности, попадающие в этот шар, и новый центр кластера строится как центр тяжести выделенных

элементов. При анализе алгоритма «Форель» возникает проблема: завершится ли процесс улучшения положения центра кластера через конечное число шагов или же он может быть бесконечным. Она получила название «проблема остановки». Для широкого класса так называемых "эталонных алгоритмов" проблема остановки была решена в работе [9]: процесс улучшения остановится через конечное число шагов.

Отметим, что алгоритмы кластер-анализа могут быть модифицированы разнообразными способами. Например, описывая алгоритм "Форель" в стиле статистики объектов нечисловой природы, заметим, что вычисление центра тяжести для совокупности многомерных точек – это нахождение эмпирического среднего для меры близости, равной квадрату евклидова расстояния. Если взять более естественную меру близости – само евклидово расстояние, то получим алгоритм кластер-анализа "Медиана", отличающийся от "Форели" тем, что новый центр строится не с помощью средних арифметических координат элементов, попавших в кластер, а с помощью медиан.

Проблема остановки возникает не только при построении диагностических классов. Она принципиально важна, в частности, и при оценивании параметров вероятностных распределений методом максимального правдоподобия. Обычно не представляет большого труда выписать систему уравнений максимального правдоподобия и предложить решать ее каким-либо численным методом. Однако когда остановиться, сколько итераций сделать, какая точность оценивания будет при этом достигнута? Общий ответ, видимо, невозможно найти, но обычно нет ответа и для конкретных семейств распределения вероятностей. Именно поэтому нет оснований рекомендовать решать системы уравнений максимального правдоподобия. Вместо них целесообразно использовать т.н. одношаговые оценки (подробнее см. об этих оценках главу 2.2). Эти оценки задаются конечными формулами, но асимптотически столь же хороши (на профессиональном языке - эффективны), как и оценки максимального правдоподобия.

**О сравнении алгоритмов диагностики по результатам обработки реальных данных.** Перейдем к этапу применения диагностических правил, когда классы, к одному из которых нужно отнести вновь поступающий объект, уже выделены.

В прикладных исследованиях применяют различные методы дискриминантного анализа, основанные на вероятностно-статистических моделях, а также с ними не связанные, т.е. эвристические, использующие детерминированные методы анализа данных. Независимо от "происхождения", каждый подобный алгоритм должен быть исследован как на параметрических и непараметрических вероятностно-статистических моделях порождения данных, так и на различных массивах реальных данных. Цель исследования - выбор наилучшего алгоритма в определенной области применения, включение его в стандартные программные продукты, методические материалы, учебные программы и пособия. Но для этого надо уметь сравнивать алгоритмы по качеству. Как это делать?

Часто используют такой показатель качества алгоритма диагностики, как "вероятность правильной классификации" (при обработке конкретных данных - "частота правильной классификации"). Чуть ниже мы покажем, что этот показатель качества некорректен, а потому пользоваться им не рекомендуется. Целесообразно применять другой показатель качества алгоритма диагностики - оценку специального вида т.н. "расстояния Махаланобиса" между классами. Изложение проведем на примере разработки программного продукта для специалистов по диагностике материалов. Прообразом является диалоговая система «АРМ материаловеда», разработанная Институтом высоких статистических технологий и эконометрики для ВНИИ эластомерных материалов.

При построении информационно-исследовательской системы диагностики материалов (ИИСДМ) возникает задача сравнения прогностических правил «по силе». Прогностическое правило - это алгоритм, позволяющий по характеристикам материала прогнозировать его свойства. Если прогноз дихотомичен («есть» или «нет»), то правило является алгоритмом диагностики, при котором материал относится к одному из двух классов. Ясно, что случай нескольких классов может быть сведен к конечной последовательности выбора между двумя классами.

Прогностические правила могут быть извлечены из научно-технической литературы и практики. Каждое из них обычно формулируется в терминах небольшого числа признаков, но наборы признаков сильно меняются от правила к правилу. Поскольку в ИИСДМ должно

фиксироваться лишь ограниченное число признаков, то возникает проблема их отбора. Естественно отбирать лишь те их них, которые входят в наборы, дающие наиболее «надежные» прогнозы. Для придания точного смысла термину «надежный» необходимо иметь способ сравнения алгоритмов диагностики по прогностической "силе".

Результаты обработки реальных данных с помощью некоторого алгоритма диагностики в рассматриваемом случае двух классов описываются долями: правильной диагностики в первом классе $\kappa$; правильной диагностики во втором классе $\lambda$; долями классов в объединенной совокупности $\pi_i$, $i = 1.2$; $\pi_1 + \pi_2 = 1$.

При изучении качества алгоритмов классификации их сравнивают по результатам дискриминации вновь поступающей контрольной выборки. Именно по контрольной выборке определяются величины $\kappa, \lambda, \pi_1, \pi_2$. Однако иногда вместо контрольной используют обучающую выборку, т.е. указанные величины определяются ретроспективно, в результате анализа уже имеющихся данных. Обычно это связано с трудоемкостью получения данных. Тогда к и л зависимы. Однако в случае, когда решающее правило основано на использовании дискриминантной поверхности, параметры которой оцениваются по обучающим выборкам, величины к и л асимптотически (при безграничном росте объемов выборок) независимы [9], что позволяет использовать приводимые ниже результаты и в этом случае.

Нередко как показатель качества алгоритма диагностики (прогностической «силы») используют долю правильной диагностики

$$\mu = \pi_1 \kappa + \pi_2 \lambda.$$

Однако показатель $\mu$ определяется, в частности, через характеристики $\pi_1$ и $\pi_2$, частично заданные исследователем (например, на них влияет тактика отбора образцов для изучения). В аналогичной медицинской задаче величина $\mu$ оказалась больше для тривиального прогноза, согласно которому у всех больных течение заболевания будет благоприятно. Тривиальный прогноз сравнивался с алгоритмом выделения больных с прогнозируемым тяжелым течением заболевания. Он был разработан группы под руководством академика АН СССР И.М. Гельфанда. Применение этого алгоритма с медицинской точки зрения вполне оправдано [13].

Другими словами, по доле правильной классификации алгоритм академика И.М. Гельфанда оказался хуже тривиального - объявить всех больных легкими, не требующими специального наблюдения. Этот вывод очевидно нелеп. И причина появления нелепости вполне понятна. Хотя доля тяжелых больных невелика, но смертельные исходы сосредоточены именно в этой группе больных. Поэтому целесообразна гипердиагностика - рациональнее часть легких больных объявить тяжелыми, чем сделать ошибку в противоположную сторону. Применение теории статистических решений в рассматриваемой постановке вряд ли возможно, поскольку оценить количественно потери от смерти больного нельзя по этическим соображениям. Поэтому, на наш взгляд, долю правильной диагностики $\mu$ нецелесообразно использовать как показатель качества алгоритма диагностики.

Применение теории статистических решений требует знания потерь от ошибочной диагностики, а в большинстве научно-технических и экономических задач определить потери, как уже отмечалось, сложно. В частности, из-за необходимости оценивать человеческую жизнь в денежных единицах. По этическим соображениям это, на наш взгляд, недопустимо. Сказанное не означает отрицания пользы страхования, но, очевидно, страховые выплаты следует рассматривать лишь как способ первоначального смягчения потерь от утраты близких.

Для выявления информативного набора признаков целесообразно использовать *метод пересчета на модель линейного дискриминантного анализа,* согласно которому статистической оценкой прогностической "силы" является

$$\delta^* = \Phi(d^*/2), \quad d^* = \Phi^{-1}(\kappa) + \Phi^{-1}(\lambda),$$

где $\Phi(x)$ - функция стандартного нормального распределения вероятностей с математическим ожиданием 0 и дисперсией 1, а $\Phi^{-1}(y)$ - обратная ей функция.

*Пример 1.* Если доли правильной классификации $\kappa$ = 0,90 и л = 0,80, то Ц$^{-1}$(к) = 1,28 и Ц$^{-1}$(л) = 0,84, откуда $d^*$ = 2,12 и прогностическая сила д$^*$ = Ц$^{-1}$(1,06) = 0,86. При этом доля

правильной классификации м может принимать любые значения между 0,80 и 0,90, в зависимости от доли элементов того или иного класса среди анализируемых данных.

Если классы описываются выборками из многомерных нормальных совокупностей с одинаковыми матрицами ковариаций, а для классификации применяется классический линейный дискриминантный анализ Р.Фишера, то величина $d*$ представляет собой состоятельную статистическую оценку так называемого расстояния Махаланобиса между рассматриваемыми двумя совокупностями (конкретный вид этого расстояния сейчас не имеет значения), независимо от порогового значения, определяющего конкретное решающее правило. В общем случае показатель $\delta^*$ вводится как эвристический.

Пусть алгоритм классификации применялся к совокупности, состоящей из $m$ объектов первого класса и $n$ объектов второго класса.

*Теорема 2*. Пусть $m, n \to \infty$. Тогда для всех $x$

$$P\left\{\frac{\delta* - \delta}{A(\kappa,\lambda)} < x\right\} \to \Phi(x),$$

где $\delta$ - истинная "прогностическая сила" алгоритма диагностики; $\delta*$ - ее эмпирическая оценка,

$$A^2(\kappa,\lambda) = \frac{1}{4}\left\{\left[\frac{\varphi(d*/2)}{\varphi(\Phi^{-1}(\kappa))}\right]^2 \frac{\kappa(1-\kappa)}{m} + \left[\frac{\varphi(d*/2)}{\varphi(\Phi^{-1}(\lambda))}\right]^2 \frac{\lambda(1-\lambda)}{n}\right\};$$

$\varphi(x) = \Phi'(x))$ - плотность стандартного нормального распределения вероятностей с математическим ожиданием 0 и дисперсией 1.

С помощью теоремы 2 по $\kappa$ и $\lambda$ обычным образом определяют доверительные границы для "прогностической силы" $\delta$.

*Пример 2.* В условиях примера 1 при $m = n = 100$ найдем асимптотическое среднее квадратическое отклонение $A(0{,}90;\ 0{,}80)$.

Поскольку ц(Ц$^{-1}$(к)) = ц(1,28) = 0,176, ц(Ц$^{-1}$(л)) = ц(0,84) = 0,280, ц($d$*/2) = ц(1,06) = 0,227, то подставляя в выражение для $A^2$ численные значения, получаем, что

$$A^2(0{,}90;\ 0{,}80) = \frac{0{,}0372}{m} + \frac{0{,}0265}{n}$$

(численные значения плотности стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1 и функции, обратной к функции этого распределения, можно было взять, например, из справочника [4]).

При $m = n = 100$ имеем А(0,90; 0,80) = 0,0252. При доверительной вероятности г = 0,95 имеем $u(0{,}95)$ = Ц$^{-1}$(1,0,975) = 1,96, а потому нижняя доверительная граница для прогностической силы д есть $д_Н$ = 0,86 – 1,96 Ч 0,0252 = 0,81, а верхняя доверительная граница такова: $д_В$ = 0,86 + 1,96 Ч 0,0252 = 0,91. Аналогичный расчет при $m = n = 1000$ дает $д_Н$ = 0,845, $д_В$ = 0,875.

Как проверить обоснованность пересчета на модель линейного дискриминантного анализа? Допустим, что классификация состоит в вычислении некоторого прогностического индекса $y$ и сравнении его с заданным порогом $c$. Объект относят к первому классу, если $y \le c$, ко второму, если $y > c$. Прогностический индекс – это обычно линейная функция от характеристик рассматриваемых объектов. Другими словами, от координат векторов, описывающих объекты.

Возьмем два значения порога $c_1$ и $c_2$. Если пересчет на модель линейного дискриминантного анализа обоснован, то , как можно показать, "прогностические силы" для обоих правил совпадают: $\delta(c_1) = \delta(c_2)$. Выполнение этого равенства можно проверить как статистическую гипотезу.

Пусть $\kappa_1$ - доля объектов первого класса, для которых $y \le c_1$, а $\kappa_2$ - доля объектов первого класса, для которых $c_1 < y \le c_2$. Аналогично пусть $\lambda_2$ - доля объектов второго класса, для которых $c_1 < y \le c_2$, а $\lambda_3$ - доля объектов второго класса, для которых $y > c_2$. Тогда можно рассчитать две оценки одного и того же расстояния Махаланобиса. Они имеют вид:

$$d*(c_1) = \Phi^{-1}(\kappa_1) + \Phi^{-1}(\lambda_2 + \lambda_3), \quad d*(c_2) = \Phi^{-1}(\kappa_1 + \kappa_2) + \Phi^{-1}(\lambda_3).$$

*Теорема 3.* Если истинные прогностические силы двух правил диагностики совпадают, $\delta(c_1) = \delta(c_2)$, то при $m \to \infty, n \to \infty$ при всех $x$

$$P\left\{\frac{d^*(c_1) - d^*(c_2)}{B} < x\right\} \to \Phi(x),$$

где

$$B^2 = \frac{1}{m}T(\kappa_1;\kappa_2) + \frac{1}{n}T(\lambda_3;\lambda_2);$$

$$T(x;y) = \frac{x(1-x)}{\varphi^2(\Phi^{-1}(x))} + \frac{(x+y)(1-x-y)}{\varphi^2(\Phi^{-1}(x+y))} - \frac{2x(1-x-y)}{\varphi(\Phi^{-1}(x))\varphi(\Phi^{-1}(x+y))}.$$

Из теоремы 3 вытекает метод проверки рассматриваемой гипотезы: при выполнении неравенства

$$\left|\frac{d^*(c_1) - d^*(c_2)}{B}\right| \le \Phi^{-1}(1 - \frac{\alpha}{2})$$

она принимается на уровне значимости, асимптотически равном $\alpha$, в противном случае - отвергается.

*Пример 3.* Пусть данные примеров 1 и 2 соответствуют порогу $c_1$. Пусть порогу $c_2$ соответствуют к' = 0,95 и л' = 0,70. Тогда в обозначениях теоремы 3 $к_1$ = 0,90, $к_2$ = 0,05, $л_2$ = 0,10, $л_3$ = 0,70. Далее $d^*(c_1)$ = 2,12 (пример 1), $d^*(c_2)$ = 2,17, $T(к_1, к_2)$ = 2,22, $T(л_3, л_2)$ = 0,89. Гипотеза о совпадении прогностических сил на двух порогах принимается на уровне значимости б = 0,05 тогда и только тогда, когда

$$\frac{0{,}05^2}{\frac{2{,}22}{m} + \frac{0{,}89}{n}} \le 1{,}96^2,$$

т.е. когда

$$\frac{2{,}22}{m} + \frac{0{,}89}{n} \ge 0{,}00065.$$

Так, гипотеза принимается при $m = n = 1000$ и отвергается при $m = n = 5000$.

**Подходы к построению прогностических правил.** Для решения задач диагностики используют два подхода – параметрический и непараметрический. Первый из них обычно основан на использовании того или иного индекса и сравнения его с порогом. Индекс может быть построен по статистическим данным, например, как в уже упомянутом линейном дискриминантном анализе Фишера. Часто индекс представляет собой линейную функцию от характеристик, выбранных специалистами предметной области, коэффициенты которой подбирают эмпирически. Непараметрический подход связан с леммой Неймана-Пирсона в математической статистике и с теорией статистических решений. Он опирается на использование непараметрических оценок плотностей распределений вероятностей, описывающих диагностические классы.

Обсудим ситуацию подробнее. Математические методы диагностики, как и статистические методы в целом, делятся на параметрические и непараметрические. Первые основаны на предположении, что классы описываются распределениями из некоторых параметрических семейств. Обычно рассматривают многомерные нормальные распределения, при этом зачастую принимают гипотезу о том, что ковариационные матрицы для различных классов совпадают. Именно в таких предположениях сформулирован классический дискриминантный анализ Фишера. Как известно, обычно нет оснований считать, что наблюдения извлечены из нормального распределения.

Поэтому более корректными, чем параметрические, являются непараметрические методы диагностики. Исходная идея таких методов основана на лемме Неймана-Пирсона, входящей в стандартный курс математической статистики. Согласно этой лемме решение об отнесении вновь поступающего объекта (сигнала, наблюдения и др.) к одному из двух классов принимается на основе отношения плотностей *f(x)/g(x),* где *f(x)* - плотность распределения, соответствующая первому классу, а *g(x)* - плотность распределения, соответствующая второму классу. Если плотности распределения неизвестны, то применяют их непараметрические оценки,

построенные по обучающим выборкам. Пусть обучающая выборка объектов из первого класса состоит из $n$ элементов, а обучающая выборка для второго класса - из $m$ объектов. Тогда рассчитывают значения непараметрических оценок плотностей $f_n(x)$ и $g_m(x)$ для первого и второго классов соответственно, а диагностическое решение принимают по их отношению. Таким образом, для решения задачи диагностики достаточно научиться строить непараметрические оценки плотности для выборок объектов произвольной природы.

Методы построения непараметрических оценок плотности распределения вероятностей в пространствах произвольной природы рассмотрены в главе 2.1.

### 3.2.6. Методы снижения размерности

В многомерном статистическом анализе каждый объект описывается вектором, размерность которого произвольна (но одна и та же для всех объектов). Однако человек может непосредственно воспринимать лишь числовые данные или точки на плоскости. Анализировать скопления точек в трехмерном пространстве уже гораздо труднее. Непосредственное восприятие данных более высокой размерности невозможно. Поэтому вполне естественным является желание перейти от многомерной выборки к данным небольшой размерности, чтобы «на них можно было посмотреть».

Кроме стремления к наглядности, есть и другие мотивы для снижения размерности. Те факторы, от которых интересующая исследователя переменная не зависит, лишь мешают статистическому анализу. Во-первых, на сбор информации о них расходуются ресурсы. Во-вторых, как можно доказать, их включение в анализ ухудшает свойства статистических процедур (в частности, увеличивает дисперсию оценок параметров и характеристик распределений). Поэтому желательно избавиться от таких факторов.

Обсудим с точки зрения снижения размерности пример использования регрессионного анализа для прогнозирования объема продаж, рассмотренный в подразделе 3.2.3. Во-первых, в этом примере удалось сократить число независимых переменных с 17 до 12. Во-вторых, удалось сконструировать новый фактор – линейную функцию от 12 упомянутых факторов, которая лучше всех иных линейных комбинаций факторов прогнозирует объем продаж. Поэтому можно сказать, что в результате размерность задачи уменьшилась с 18 до 2. А именно, остался один независимый фактор (приведенная в подразделе 3.2.3 линейная комбинация) и один зависимый – объем продаж.

При анализе многомерных данных обычно рассматривают не одну, а множество задач, в частности, по-разному выбирая независимые и зависимые переменные. Поэтому рассмотрим задачу снижения размерности в следующей формулировке. Дана многомерная выборка. Требуется перейти от нее к совокупности векторов меньшей размерности, максимально сохранив структуру исходных данных, по возможности не теряя информации, содержащихся в данных. Задача конкретизируется в рамках каждого конкретного метода снижения размерности.

**Метод главных компонент** является одним из наиболее часто используемых методов снижения размерности. Основная его идея состоит в последовательном выявлении направлений, в которых данные имеют наибольший разброс. Пусть выборка состоит из векторов, одинаково распределенных с вектором $X = (x(1), x(2), \dots, x(n))$. Рассмотрим линейные комбинации

$$Y(л(1), л(2), \dots, л(n)) = л(1)x(1) + л(2)x(2) + \dots + л(n)x(n),$$

где

$$л^2(1) + л^2(2) + \dots + л^2(n) = 1.$$

Здесь вектор $л = (л(1), л(2), \dots, л(n))$ лежит на единичной сфере в $n$-мерном пространстве.

В методе главных компонент прежде всего находят направление максимального разброса, т.е. такое л, при котором достигает максимума дисперсия случайной величины $Y(л) = Y(л(1), л(2), \dots, л(n))$. Тогда вектор л задает первую главную компоненту, а величина $Y(л)$ является проекцией случайного вектора $X$ на ось первой главной компоненты.

Затем, выражаясь терминами линейной алгебры, рассматривают гиперплоскость в $n$-мерном пространстве, перпендикулярную первой главной компоненте, и проектируют на эту гиперплоскость все элементы выборки. Размерность гиперплоскость на 1 меньше, чем размерность исходного пространства.

В рассматриваемой гиперплоскости процедура повторяется. В ней находят направление наибольшего разброса, т.е. вторую главную компоненту. Затем выделяют гиперплоскость, перпендикулярную первым двум главным компонентам. Ее размерность на 2 меньше, чем размерность исходного пространства. Далее – следующая итерация.

С точки зрения линейной алгебры речь идет о построении нового базиса в $n$-мерном пространстве, ортами которого служат главные компоненты.

Дисперсия, соответствующая каждой новой главной компоненте, меньше, чем для предыдущей. Обычно останавливаются, когда она меньше заданного порога. Если отобрано $k$ главных компонент, то это означает, что от $n$-мерного пространства удалось перейти к $k$-мерному, т.е. сократить размерность с $n$-до $k$, практически не исказив структуру исходных данных.

Для визуального анализа данных часто используют проекции исходных векторов на плоскость первых двух главных компонент. Обычно хорошо видна структура данных, выделяются компактные кластеры объектов и отдельно выделяющиеся вектора.

Метод главных компонент является одним из методов **факторного анализа** [14]**.** Различные алгоритмы факторного анализа объединены тем, что во всех них происходит переход к новому базису в исходном $n$-мерном пространстве. Важным является понятие «нагрузка фактора», применяемое для описания роли исходного фактора (переменной) в формировании определенного вектора из нового базиса.

Новая идея по сравнению с методом главных компонент состоит в том, что на основе нагрузок происходит разбиение факторов на группы. В одну группу объединяются факторы, имеющие сходное влияние на элементы нового базиса. Затем из каждой группы рекомендуется оставить одного представителя. Иногда вместо выбора представителя расчетным путем формируется новый фактор, являющийся центральным для рассматриваемой группы. Снижение размерности происходит при переходе к системе факторов, являющихся представителями групп. Остальные факторы отбрасываются.

Описанная процедура может быть осуществлена не только с помощью факторного анализа. Речь идет о кластер-анализе признаков (факторов, переменных). Для разбиения признаков на группы можно применять различные алгоритмы кластер-анализа. Достаточно ввести расстояние (меру близости, показатель различия) между признаками. Пусть $X$ и $Y$ – два признака. Различие $d(X,Y)$ между ними можно измерять с помощью выборочных коэффициентов корреляции:

$$d_1(X,Y) = 1 - r_n(X,Y), \quad d_2(X,Y) = 1 - \mathrm{c}_n(X,Y),$$

где $r_n(X,Y)$ – выборочный линейный коэффициент корреляции Пирсона, $\mathrm{c}_n(X,Y)$ – выборочный коэффициент ранговой корреляции Спирмена.

**Многомерное шкалирование**. На использовании расстояний (мер близости, показателей различия) $d(X,Y)$ между признаками $X$ и $Y$ основан обширный класс методов многомерного шкалирования [15, 16]. Основная идея этого класса методов состоит в представлении каждого объекта точкой геометрического пространства (обычно размерности 1, 2 или 3), координатами которой служат значения скрытых (латентных) факторов, в совокупности достаточно адекватно описывающих объект. При этом отношения между объектами заменяются отношениями между точками – их представителями. Так, данные о сходстве объектов – расстояниями между точками, данные о превосходстве – взаимным расположением точек [17].

В практике используется ряд различных моделей многомерного шкалирования. Во всех них встает проблема оценки истинной размерности факторного пространства. Рассмотрим эту проблему на примере обработки данных о сходстве объектов с помощью метрического шкалирования.

Пусть имеется $n$ объектов $O(1)$, $O(2)$, …, $O(n)$, для каждой пары объектов $O(i)$, $O(j)$ задана мера их сходства $s(i,j)$. Считаем, что всегда $s(i,j) = s(j,i)$. Происхождение чисел $s(i,j)$ не имеет значения для описания работы алгоритма. Они могли быть получены либо непосредственным измерением, либо с использованием экспертов, либо путем вычисления по совокупности описательных характеристик, либо как-то иначе.

В евклидовом пространстве рассматриваемые $n$ объектов должны быть представлены конфигурацией $n$ точек, причем в качестве меры близости точек-представителей выступает евклидово расстояние $d(i,j)$ между соответствующими точками. Степень соответствия между

совокупностью объектов и совокупностью представляющих их точек определяется путем сопоставления матриц сходства $||s(i,j)||$ и расстояний $||d(i,j)||$. Метрический функционал сходства имеет вид

$$S = \sum_{i<j} | s(i,j) - d(i,j) |^2 .$$

Геометрическую конфигурацию надо выбирать так, чтобы функционал S достигал своего наименьшего значения [17].

*Замечание.* В неметрическом шкалировании вместо близости самих мер близости и расстояний рассматривается близость упорядочений на множестве мер близости и множестве соответствующих расстояний. Вместо функционала $S$ используются аналоги ранговых коэффициентов корреляции Спирмена и Кендалла. Другими словами, неметрическое шкалирование исходит из предположения, что меры близости измерены в порядковой шкале.

Пусть евклидово пространство имеет размерность $m$. Рассмотрим минимум среднего квадрата ошибки

$$\alpha_m = \frac{2}{n(n-1)} \min S ,$$

где минимум берется по всем возможным конфигурациям $n$ точек в $m$-мерном евклидовом пространстве. Можно показать, что рассматриваемый минимум достигается на некоторой конфигурации. Ясно, что при росте $m$ величина $б_m$ монотонно убывает (точнее, не возрастает). Можно показать, что при $m \geq n - 1$ она равна 0 (если $s(i,j)$ – метрика). Для увеличения возможностей содержательной интерпретации желательно действовать в пространстве возможно меньшей размерности. При этом, однако, размерность необходимо выбрать так, чтобы точки представляли объекты без больших искажений. Возникает вопрос: как рационально выбирать размерность, т.е. натуральное число $m$?

В рамках детерминированного анализа данных обоснованного ответа на этот вопрос, видимо, нет. Следовательно, необходимо изучить поведение $б_m$ в тех или иных вероятностных моделях. Если меры близости $s(i,j)$ являются случайными величинами, распределение которых зависит от «истинной размерности» $m_0$ (и, возможно, от каких-либо еще параметров), то можно в классическом математико-статистическом стиле ставить задачу оценки $m_0$, искать состоятельные оценки и т.д.

Начнем строить вероятностные модели. Примем, что объекты представляют собой точки в евклидовом пространстве размерности $k$, где $k$ достаточно велико. То, что «истинная размерность» равна $m_0$, означает, что все эти точки лежат на гиперплоскости размерности $m_0$. Примем для определенности, что совокупность рассматриваемых точек представляет собой выборку из кругового нормального распределения с дисперсией $у^2(0)$. Это означает, что объекты $O(1)$, $O(2)$, …, $O(n)$ являются независимыми в совокупности случайными векторами, каждый из которых строится как $ж(1)e(1) + ж(2)e(2) + … + ж(m_0)e(m_0)$, где $e(1)$, $e(2)$, … , $e(m_0)$ – ортонормальный базис в подпространстве размерности $m_0$, в котором лежат рассматриваемые точки, а ж(1), ж(2), … , $ж(m_0)$ – независимые в совокупности одномерные нормальные случайные величины с математическим ожиданием ) и дисперсией $у^2(0)$.

Рассмотрим две модели получения мер близости $s(i,j)$. В первой из них $s(i,j)$ отличаются от евклидова расстояния между соответствующими точками из-за того, что точки известны с искажениями. Пусть $c(1)$, $c(2)$, … , $c(n)$ – рассматриваемые точки. Тогда

$$s(i,j) = d(c(i) + e(i), c(j) + e(j)), \ i,j = 1, 2, … , n,$$

где $d$ – евклидово расстояние между точками в $k$-мерном пространстве, вектора e(1), e(2), … , e($n$) представляют собой выборку из кругового нормального распределения в $k$-мерном пространстве с нулевым математическим ожиданием и ковариационной матрицей $у^2(1)I$, где $I$ – единичная матрица. Другими словами, $e(i) = з(1)e(1) + з(2)e(2) + … + з(k)e(k)$, где $e(1)$, $e(2)$, …, $e(k)$ – ортонормальный базис в $k$-мерном пространстве, а {з($i,t$), $i$ = 1, 2, … , n, $t$ = 1, 2, … , k} – совокупность независимых в совокупности одномерных случайных величин с нулевым математическим ожиданием и дисперсией $у^2(1)$.

Во второй модели искажения наложены непосредственно на сами расстояния:

$$s(i,j) = d(c(i), c(j)) + e(i,j), \ i,j = 1, 2, … , n, i \neq j,$$

где {e(*i,j*), *i,j* = 1, 2, … , *n*} – независимые в совокупности нормальные случайные величины с математическим ожиданием ) и дисперсией $y^2(1)$.

В работе [18] показано, что для обеих сформулированных моделей минимум среднего квадрата ошибки $б_m$ при $n \to \infty$ сходится по вероятности к

$$f(m) = f_1(m) + y^2(1)(k - m),\ \ m = 1, 2, \ldots, k,$$

где

$$f_1(m) = \begin{cases} \sigma^2(0)(m_0 - m), & m < m_0, \\ 0, & m \geq m_0. \end{cases}$$

Таким образом, функция $f(m)$ линейна на интервалах [1, $m_0$] и [$m_0$, k], причем на первом интервале она убывает быстрее, чем на втором. Отсюда следует, что статистика

$$m^* = Arg \min_m \{\alpha_{m+1} - 2\alpha_m + \alpha_{m-1})$$

является состоятельной оценкой истинной размерности $m_0$.

Итак, из вероятностной теории вытекает рекомендация – в качестве оценки размерности факторного пространства использовать $m^*$. Отметим, что подобная рекомендация была сформулировано как эвристическая одним из основателей многомерного шкалирования Дж. Краскалом [15]. Он исходил из опыта практического использования многомерного шкалирования и вычислительных экспериментов. Вероятностная теория позволила обосновать эту эвристическую рекомендацию.

### 3.2.7. Индексы и их применение

Индекс (лат. *index* – показатель, список) - статистический относительный показатель, характеризующий соотношение во времени (динамический индекс) или в пространстве (территориальный индекс) социально-экономических явлений. Речь идет о ценах на товары и услуги, объемах производства, себестоимости, объемах продаж и др. Индексы делятся на индивидуальные и сводные. Так, индивидуальный динамический индекс описывает изменение тех или иных явлений во времени. Например, изменения цены на отдельный товар, объема выплавки стали, урожайности картофеля. Для вычисления индивидуального индекса значение измеряемой величины в текущем периоде делят на ее значение в базисном периоде. Сводный индекс служит для сопоставления непосредственно несоизмеримых, разнородных явлений. Например, объемов продаж различных продовольственных товаров (в килограммах). Для требуемого сопоставления необходимо составные элементы несоизмеримых явлений сделать соизмеримыми, выразив их общей мерой: стоимостью, трудовыми затратами и т.д. Сводные индексы обычно имеют один из трех видов:

$$I_1 = \frac{\sum x_1 f_0}{\sum x_0 f_0}, \quad I_2 = \frac{\sum x_1 f_1}{\sum x_0 f_1}, \quad I_3 = \frac{\sum x_1 f_1}{\sum x_0 f_0},$$

где $x$ – индексируемая величина, $f$ – веса индексов, 0 и 1 – знаки соответственно базисного и текущего периодов [19, с.154]. Таким образом, индексы, как и коэффициенты корреляции, зависят от двух переменных – индексируемой величины $x$ и весов индексов $f$.

В качестве примера построения и использования индексов рассмотрим индекс потребительских цен, он же - индекс инфляции.

Под инфляцией понимаем рост (изменение) цен [6]. При анализе экономических процессов, протяженных во времени, необходимо переходить к сопоставимым ценам. Это невозможно сделать без расчета индекса роста цен, т.е. индекса инфляции. Проблема состоит в том, что цены на разные товары растут с различной скоростью, и необходимо эти скорости усреднять.

Рассмотрим конкретного покупателя товаров и услуг, т.е. конкретного экономического субъекта: физическое лицо, домохозяйство или фирму. Он покупает не один товар, а много. Обозначим через $n$ количество типов товаров или услуг (далее кратко - товаров), которые он хочет и может купить. Пусть

$$Q_i = Q_i(t),\ i=1,2,\ldots,n,$$

- объемы покупок этих товаров в момент времени $t$ по ценам:

$$r_i = r_i(t),\ i=1,2,...,n$$

(имеется в виду цена за единицу измерения соответствующего товара, например, за штуку или килограмм...).

Подход к измерению роста цен основан на выборе и фиксации потребительской корзины $(Q_1(t),\ Q_2(t),\ ...,\ Q_n(t))$, не меняющейся со временем, т.е. $(Q_1(t),\ Q_2(t),\ ...,\ Q_n(t)) \equiv (Q_1,\ Q_2,\ ...,\ Q_n)$. Затем необходимо сравнить стоимости потребительской корзины $(Q_1,\ Q_2,\ ...,\ Q_n)$ в старых $r_i(t_1)$, $i=1,2,...,n,$ и новых $r_i(t_2)$, $i=1,2,...,n,$ ценах.

**Определение.** *Индексом инфляции называется*

$$I(t_1,t_2) = \frac{\sum\limits_{1\le i\le n} r_i(t_2)Q_i}{\sum\limits_{1\le i\le n} r_i(t_1)Q_i}.$$

Таким образом, каждой потребительской корзине соответствует свой индекс инфляции. Однако согласно теореме сложения для индекса инфляции [6] он является средним взвешенным арифметическим роста цен на отдельные товары. Поэтому индексы инфляции, рассчитанные по разным достаточно обширным и представительным потребительским корзинам, достаточно близки между собой (см. конкретные данные в [6]).

Институт высоких статистических технологий и эконометрики (ИВСТЭ) использовал для измерения инфляции минимальную потребительскую корзину физиологически необходимых продовольственных товаров [6]. Она была разработана на основе исходных данных Института питания Российской академии медицинских наук (РАМН). Данные о динамике индекса инфляции приведены в табл.3.

Таблица 3.

Индекс инфляции и стоимость потребительской корзины

| № п/п | Дата снятия цен | Стоимость потребительской корзины *S(t)* (руб.) | Индекс инфляции *I(31.3.91;t)* |
|---|---|---|---|
| 1 | 31.3.91 | 26.60 | 1.00 |
| 2 | 14.8.93 | 17,691.00 | 665.08 |
| 3 | 15.11.93 | 28,050.00 | 1054.51 |
| 4 | 14.3.94 | 40,883.00 | 1536.95 |
| 5 | 14.4.94 | 44,441.00 | 1670.71 |
| 6 | 28.4.94 | 47,778.00 | 1796.17 |
| 7 | 26.5.94 | 52,600.00 | 1977.44 |
| 8 | 8.9.94 | 58,614.00 | 2203.53 |
| 9 | 6.10.94 | 55,358.00 | 2081.13 |
| 10 | 10.11.94 | 72,867.00 | 2739.36 |
| 11 | 1.12.94 | 78,955.00 | 2968.23 |
| 12 | 29.12.94 | 97,897.00 | 3680.34 |
| 13 | 2.2.95 | 129,165.00 | 4855.83 |
| 14 | 2.3.95 | 151,375.00 | 5690.79 |
| 15 | 30.3.95 | 160,817.00 | 6045.75 |
| 16 | 27.4.95 | 159,780.00 | 6006.77 |
| 17 | 1.6.95 | 167,590.00 | 6300.38 |
| 18 | 29.6.95 | 170,721.00 | 6418.08 |
| 19 | 27.7.95 | 175,499.00 | 6597.71 |
| 20 | 31.8.95 | 173,676.00 | 6529.17 |
| 22 | 28.9.95 | 217,542.00 | 8178.27 |

| | | | |
|---|---|---|---|
| 23 | 26.10.95 | 243,479.00 | 9153.35 |
| 24 | 30.11.95 | 222,417.00 | 8361.54 |
| 25 | 28.12.95 | 265,716.00 | 9989.32 |
| 26 | 1.2.96 | 287,472.55 | 10,807.24 |
| 27 | 5.3.96 | 297,958.00 | 11,201.43 |
| 28 | 5.4.96 | 304,033.44 | 11,429.83 |
| 29 | 8.5.96 | 305,809.55 | 11,496.60 |
| 30 | 5.6.96 | 302,381.69 | 11,367.73 |
| 31 | 3.7.96 | 306,065.21 | 11,506.21 |
| 32 | 3.8.96 | 308,963.42 | 11,615.17 |
| 33 | 7.9.96 | 288,835.07 | 10,858.46 |
| 34 | 1.10.96 | 278,235.35 | 10,459.98 |
| 35 | 5.11.96 | 287,094.77 | 10,793.04 |
| 36 | 4.12.96 | 298,024.76 | 11,203.94 |
| 37 | 3.1.97 | 314,287.16 | 11,815.31 |
| 38 | 4.2.97 | 334,738.24 | 12,584.14 |
| 39 | 4.1.98 | 345.72 | 12.997 |
| 40 | 3.1.99 | 622.30 | 23.395 |
| 41 | 5.1.00 | 851.32 | 32.004 |
| 42 | 3.1.01 | 949.21 | 35.684 |
| 43 | 2.7.01 | 1072.61 | 40.323 |
| 44 | 3.1.02 | 1125,76 | 43,321 |
| 45 | 2.7.02 | 1247.77 | 46.908 |
| 46 | 3.1.03 | 1295.75 | 48.712 |
| 47 | 1.7.03 | 1398.11 | 52.558 |

*Примечание 1.* В таблице целая часть отделяется от дробной десятичной точкой, а запятая используется для деления числа по разрядам (на западный манер). Учитывается проведенная деноминация рубля. Если ее не учитывать, то за 12 лет (1991-2003) цены (в Москве) выросли примерно в 50 тысяч раз. Поскольку экономические связи между регионами ослабли, то темпы роста цен в регионах различаются, но, видимо, не более чем на 10-20%.

**Использования индекса инфляции в экономических расчетах при принятии решений.** Хорошо известно, что стоимость денежных единиц со временем меняется. Например, на один доллар США полвека назад можно было купить примерно в восемь раз больше материальных ценностей (например, продовольствия), чем сейчас (см. таблицу пересчета в учебнике [20]), а если сравнивать с временами Тома Сойера - в 100 раз больше. Причем стоимость денежных единиц с течением времени, как правило, падает. Этому есть две основные причины - банковский процент и инфляция. В экономике есть инструменты для учета изменения стоимости денежных единиц с течением времени. Один из наиболее известных - расчет *NPV* (*Net Present Value*) - чистой текущей стоимости. Однако бухгалтерский учет и построенный на данных баланса предприятия экономический анализ финансово-хозяйственной деятельности предприятия пока что, как правило, игнорируют сам факт наличия инфляции. Обсудим некоторые возможности использования индекса инфляции в экономических расчетах в процессе подготовки и принятия решений.

**Переход к сопоставимым ценам.** Индекс инфляции даст возможность перехода к сопоставимым ценам, расходам, доходам и другим экономическим величинам. Например, по данным табл.7 индекс инфляции за 4 года - с 14.03.91 г. по 16.03.95 г. - составил 5936. Это означает, что покупательной способности 1 рубля марта 1991 г. соответствует примерно 6000 (а точнее 5936) рублей марта 1995 г.

Рассмотрим приведение доходов к неизменным ценам. Пусть Иван Иванович Иванов получал в 1990 г. 300 руб. в месяц, а в мае 1995 г. - 1 миллион руб. в месяц. Увеличились его доходы или уменьшились?

Номинальная заработная плата выросла в 1000000/300 = 3333 раза. Однако индекс инфляции на 18 мая 1995 г. составлял 7080. Это значит, что 1 руб. 1990 г. соответствовал по покупательной способности 7080 руб. в ценах на 18.05.95 г. Следовательно, в ценах 1990 г. доход И.И. Иванова составлял 1000000/7080 = 142 руб. 24 коп., т.е. 47,4% от дохода в 1990 г.

Можно поступить наоборот, привести доход 1990 г. к ценам на 18 мая 1995 г. Для этого достаточно умножить его на индекс инфляции: доход 1990 г. соответствует 300 х 7080 = 2 миллиона 124 тыс. руб. в ценах мая 1995 г.

**Средняя зарплата.** По данным Госкомстата РФ средняя заработная плата составляла в 1990 г. 297 руб., в октябре 1993 г. - 93 тыс. руб., в январе 1995 г. - 303 тыс. руб. Поскольку зарплата тратится в основном в следующем месяце после получки, то рассмотрим индексы инфляции на 15.11.93 г. и 2.02.95 г., равные 1045 и 4811 соответственно. В ценах 1990 г. средняя зарплата составила 89 руб. и 62 руб.98 коп. соответственно, т.е. 30% и 21,2% от зарплаты 1990 г.

Средняя зарплата рассчитывается путем деления фонда оплаты труда на число работников. При этом объединяются доходы и низкооплачиваемых лиц и сравнительно высокооплачиваемых. Известно, что распределение доходов резко асимметрично, большому числу низкооплачиваемых работников соответствует малое число лиц с высокими доходами. За 1991-1995-е годы дифференциация доходов резко увеличилась. Это означает, что доходы основной массы трудящихся сдвинулись влево относительно средней зарплаты. По нашей оценке 50% получают не более 70% от средней зарплаты, т.е. не более 212100 руб. по состоянию на январь 1995 г., а наиболее массовой является оплата в 50% от средней, т.е. около 150 тыс. руб. в месяц.

Доходы отдельных слоев трудящихся снизились еще существеннее. Зарплата профессора Московского государственного института электроники и математики (технического университета) составляла в марте 1994 г. - 42 руб.92 коп. (в ценах 1990 г.), в июле 1995 г. - 43 руб. 01 коп., т.е. с 1990 г. (400 руб.) снизилась в 9,3 раза, дошла до уровня прежней студенческой стипендии. А студенческие стипендии снизились примерно в той же пропорции и составляли 4-5 руб. в ценах 1990 г.

Кроме того, необходимо учесть, что Госкомстат учитывает начисленную зарплату, а не выплаченную. В отдельные периоды отечественной истории выплата заработной платы откладывалась надолго.

**Минимальная зарплата и прожиточный минимум.** Минимальная зарплата в сентябре 1994 г. (22500 руб.) и в мае 1995 г. (43700 руб.) составляла 38% и 23,4% соответственно от стоимости минимальной физиологически необходимой продовольственной корзины. После подъема до 55 тыс. руб. она в сентябре 1995 г. составляла около 26,34% от стоимости корзины, т.е. реально уменьшилась в 1,44 раза по сравнению с сентябрем 1994 г. В дальнейшем уменьшение стало еще более заметным.

Минимальная зарплата вместе с единой тарифной сеткой во многом определяла зарплату работников бюджетной сферы. Учитывая снижение коэффициентов тарифной сетки, проведенное весной 1995 г., снижение в 1,5 раза покупательной способности минимальной зарплаты, необходимо заключить, что в сентябре 1995 г. доход бюджетников в 2 раза меньше, чем год назад.

Оценим прожиточный минимум. Бюджетные обследования 1990 года показали, что для лиц с низкими доходами расходы на продовольствие составляют около 50% всех расходов, т.е. на промтовары и услуги идет около 50% доходов. Это соотношение подтвердило и проведенное ИВСТЭ бюджетное обследование конца 1995 г. Исходя из него, среднедушевой прожиточный минимум можно оценить, умножая на 2,0 стоимость минимальной продовольственной корзины ИВСТЭ. Например, на 1 сентября 1995 г. - 418220 руб. Т.е. прожиточный уровень для семьи из трех человек - муж, жена и ребенок - должен был на 1 сентября 1995 г. составлять 1,25 миллиона руб. (в месяц). Например, муж должен получать 800 тыс. руб., жена - 450 тыс. руб. в месяц. Очевидно, доходы большинства трудящихся меньше прожиточного уровня.

Численные значения стоимостей потребительских корзин и индексов инфляции рассчитаны ИВСТЭ в основном по ценам на продукты в Москве и Подмосковье. Однако для

других регионов численные значения отличаются мало. Для Москвы индекс инфляции на 1.09.95 г. - 7759, а для Иванова на 1.08.95 г. - 7542. Поскольку потребительская корзина на 14.03.91 г. в Иванове была на 95 коп. дешевле, то и на 1.08.95 г. она несколько дешевле - 195337 руб., а прожиточный минимум равен 390673 руб. Приведенные выше численные значения для Москвы в качестве первого приближения можно использовать для различных регионов России.

Индексы инфляции с помощью описанной выше методики можно рассчитать для любого региона, профессиональной или социальной группы, отдельного предприятия или даже конкретной семьи. Эти значения могут быть эффективно использованы на трехсторонних переговорах между профсоюзами, работодателями и представителями государства.

**Проценты по вкладам в банк, плата за кредит и инфляция.** Рассмотрим банк, честно выполняющий свои обязательства. Пусть он дает 10% в месяц по депозитным вкладам. Тогда 1 руб., положенный в банк, через месяц превращается в 1,1 руб., а через 2 - по формуле сложных процентов – в $1,1^2 = 1,21$ руб., ..., через год - в $1,1^{12} = 3,14$ руб. Однако за год росли не только вклады, но и цены. Например, с 19.05.94 г. по 18.05.95 г. индекс инфляции составил 3,73. Значит, в ценах на момент оформления вкладов итог годового хранения равен 3,14 / 3,73 = 0,84 руб. Хранение оказалось невыгодным - реальная стоимость вклада уменьшилась на 16%, несмотря на, казалось бы, очень выгодные условия банка.

Пусть фирма получила кредит под 200% годовых. Значит, вместо 1 рубля, полученного в настоящий момент в кредит, через год ей надо отдать 3 рубля. Пусть она взяла кредит 19.05.94 г., а отдает 18.05.95 г. Тогда в ценах на момент взятия кредита она отдает 3/3,73 = 0,80 руб. за 1 руб. кредита. Таким образом, кредит частично превратился в подарок - возвращать надо на 20% меньше, чем получил, реальная ставка кредита отрицательна, она равна (- 20)%! Такова была типичная ситуация в России в течение ряда лет начиная с 1992 г., особенно в 1992-1994 гг. Но бесплатных подарков в бизнесе не бывает - за них надо платить по другим каналам, как правило, криминальным.

**Сколько стоит доллар?** В июле 1995 г. индекс инфляции около 7000, а курс доллара США - около 4500 руб. за доллар. Следовательно, доллар США стоит 4500 / 7000 = 0,64 руб. в ценах 1990 г., т.е. примерно соответствует официальному обменному курсу в 1980-х годах. В сентябре 1994 г. курс доллара был около 2000, а индекс инфляции - около 2200, т.е. доллар стоил около 0,90 руб. в ценах 1990 г. Реальная покупательная способность доллара упала за 10 месяцев в 1,42 раза.

В середине 2003 г. курс доллара был несколько больше 30 руб. (30 руб. 38 коп.), индекс инфляции составлял 52,56, следовательно, 1 доллар США по своей покупательной способности в России на июль 2003 г. соответствовал 58 копейкам начала 1991 г.

**Инфляция, показатели работы предприятия и ВВП.** Индексы инфляции используются для пересчета номинальных цен в неизменные (сопоставимые). Другими словами, для приведения доходов и расходов к ценам определенного момента времени. Потребительские корзины для промышленных предприятий, конечно, должны включать промышленные товары, а потому отличаться от потребительских корзин, ориентированных для изучения жизненного уровня.

Сколько стоит предприятие? Важно оценить основные фонды. Для этого нужно взять их стоимость в определенный момент времени и умножить на индекс инфляции (и учесть амортизационные отчисления).

Валовой внутренний продукт, валовой национальный продукт и другие характеристики экономического положения страны рассчитываются в текущих ценах. Для перехода к неизменным ценам, грубо говоря, надо поделить на индекс инфляции (т.е. умножить на дефлятор).

**Проблема учета инфляции при экономическом анализе финансово-хозяйственной деятельности предприятия.** Как известно, разработана и широко применяется развернутая система коэффициентов, используемых при экономическом анализе финансово-хозяйственной деятельности предприятия [21]. Она основана на данных бухгалтерского баланса. Естественно, опирается на два столбца баланса - данные на "начало периода" и данные на "конец периода". Записывают в эти столбцы номинальные значения. В настоящее время инфляцию полностью игнорируют. Это приводит к искажению реального положения предприятия. Денежные средства

преувеличиваются, а реальная стоимость основных фондов занижается. По официальной отчетности предприятие может считаться получившим хорошую прибыль, а по существу - не иметь средств для продолжения производственной деятельности.

Ясно, что учитывать инфляцию надо. Вопрос в другом - как именно. Потребительская корзина должна, видимо, состоять из тех товаров и услуг, которые предприятие покупает. Стоимость основных фондов может не убывать в соответствии с амортизацией, а возрастать согласно отраслевому темпу инфляции, и т.д.

## Литература


1. Крамер Г. Математические методы статистики. - М.: Мир, 1975. - 648 с.
2. Красильников В.В. Статистика объектов нечисловой природы. - Набережные Челны: Изд-во Камского политехнического института, 2001. - 144 с.
3. Кендэл М. Ранговые корреляции. - М.: Статистика, 1975. - 216 с.
4. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983. - 416 с.
5. Себер Дж. Линейный регрессионный анализ. - М.: Мир, 1980. - 456 с.
6. Орлов А.И. Эконометрика. Учебник для вузов. Изд. 2-е, исправленное и дополненное. - М.: Изд-во "Экзамен", 2003. – 576 с.
7. Орлов А.И. Оценка размерности модели в регрессии. – В сб.: Алгоритмическое и программное обеспечение прикладного статистического анализа. Ученые записки по статистике, т.36. - М.: Наука, 1980. - С.92-99.
8. Кендалл М.Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. - М.: Наука, 1976. – 736 с.
9. Орлов А.И. Некоторые вероятностные вопросы теории классификации. – В сб.: Прикладная статистика. Ученые записки по статистике, т.45. - М.: Наука, 1983. – С.166-179.
10. Орлов А.И. Парные сравнения в асимптотике Колмогорова. – В сб.: Экспертные оценки в задачах управления. - М.: Изд-во ИПУ, 1982. - С. 58-66.
11. Орлов А.И.; Гусейнов Г.А. Математические методы в изучении способных к математике школьников – В сб.: Исследования по вероятностно-статистическому моделированию реальных систем. - М.: ЦЭМИ АН СССР, 1977. - С.80-93.
12. Куперштох В.Л., Миркин Б.Г., Трофимов В.А. Сумма внутренних связей как показатель качества классификации // Автоматика и телемеханика. 1976. № 3. С.91-98.
13. Гельфанд И.М., Алексеевская М.А., Губерман Ш.А. и др. Прогнозирование исхода инфаркта миокарда с помощью программы "Кора-3" // Кардиология. 1977. Т.17. № 6. С.19-23.
14. Харман Г. Современный факторный анализ. – М.: Статистика, 1972. – 488 с.
15. Терехина А.Ю. Анализ данных методами многомерного шкалирования. – М.: Наука, 1986. - 168 с.
16. Перекрест В.Т. Нелинейный типологический анализ социально-экономической информации: Математические и вычислительные методы. – Л.: Наука, 1983. – 176 с.
17. Тюрин Ю.Н., Литвак Б.Г., Орлов А.И., Сатаров Г.А., Шмерлинг Д.С. Анализ нечисловой информации. - М.: Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1981. - 80 с.
18. Орлов А.И. Общий взгляд на статистику объектов нечисловой природы. - В сб.: Анализ нечисловой информации в социологических исследованиях. - М.: Наука, 1985. С.58-92.
19. Статистический словарь / Гл. ред. М.А.Королев. – М.: Финансы и статистика, 1989. - 623 с.
20. Макконнелл К.Р., Брю С.Л. Экономикс: Принципы, проблемы и политика. В 2 т.: Пер. с англ. 11-го изд. - М.: Республика, 1992.
21. Баканов М.И., Шеремет А.Д. Теория экономического анализа. – М.: Финансы и статистика, 2000. – 416 с.


## Контрольные вопросы и задачи

1. Имеются данные за несколько лет о торговом обороте *Y* западногерманского предприятия и его расходах на рекламу *X*. Данные представлены в табл. 4.

Таблица 4.

Расходы на рекламу и торговый оборот предприятия.

| Годы, $t$ | 68 | 69 | 70 | 71 | 72 | 73 | 74 | 75 |
|---|---|---|---|---|---|---|---|---|
| Расходы на рекламу $x(t)$, тыс. марок | 4 | 4 | 5 | 6 | 8 | 8 | 10 | 11 |
| Торговый оборот $y(t)$, млн.марок | 4 | 5 | 6 | 6 | 8 | 10 | 12 | 13 |

Вычислите линейный коэффициент корреляции между случайными величинами $X$ и $Y$. С помощью метода наименьших квадратов определите коэффициенты линейной регрессии $Y = aX + b$. Постройте график (заданные точки $(x_i,y_i)$ и прямую $y= a*x+b*$). Найдите доверительные границы для регрессионной зависимости (при доверительной вероятности г = 0,95). Нанесите доверительные границы на график. Сделайте точечный и интервальный прогноз для торгового оборота при расходах на рекламу, равных 15 (тыс. марок ФРГ).

Аналогичным образом изучите зависимости расходах на рекламу $X$ и торгового оборота $Y$ от времени $t$ (за начало отсчета целесообразно взять 1971 год).

2. Семь школьников выполняют несколько заданий по математике и физике, которые оцениваются баллами 1-5, затем вычисляется средний балл для каждого школьника по каждому предмету: по математике - $x_i$, по физике - $y_j$. Данные представлены в табл.5. Определите, существует ли корреляция (т.е. связь) между этими оценками, вычислив коэффициент ранговой корреляции Спирмена.

Таблица 5.

Средние баллы по математике и физике.

| Школьник | Средний балл по математике $x_i$ | Средний балл по физике $y_i$ |
|---|---|---|
| A | 1,8 | 3,2 |
| B | 3,0 | 2,8 |
| C | 3,5 | 4,0 |
| D | 4,0 | 5,0 |
| E | 5,0 | 3,6 |
| F | 3,8 | 2,4 |
| G | 2,0 | 1,2 |

3. Исходные данные (табл.6) – набор $n$ пар чисел $(t_k , x_k), k = 1,2,...,n,$ где $t_k$ – независимая переменная (например, время), а $x_k$ – зависимая (например, индекс инфляции). Предполагается, что переменные связаны зависимостью

$$x_k = a\, t_k + b + e_k,\ k = 1,2,...,n,$$

где $a$ и $b$ – параметры, неизвестные статистику и подлежащие оцениванию, а $e_k$ – погрешности, искажающие зависимость.

Таблица 6.

Исходные данные для расчетов по методу наименьших квадратов.

| $t_k$ | 1 | 3 | 4 | 7 | 9 | 10 |
|---|---|---|---|---|---|---|
| $x_k$ | 12 | 20 | 20 | 32 | 35 | 42 |

Методом наименьших квадратов оцените параметры $a$ и $b$ линейной зависимости. Выпишите восстановленную зависимость.

Вычислите восстановленные значения зависимой переменной, сравните их с исходными значениями (найдите разности) и проверьте условие точности вычислений (при отсутствии ошибок в вычислениях сумма исходных значений должна равняться сумме восстановленных).

Найдите остаточную сумму квадратов и оцените дисперсию погрешностей.

Выпишите точечный прогноз, а также верхнюю и нижнюю доверительные границы для него (для доверительной вероятности 0,95).

Рассчитайте прогнозное значение и доверительные границы для него для момента t = 12.

Как изменятся результаты, если доверительная вероятность будет увеличена? А если она будет уменьшена?

4. Как в методе наименьших квадратов используются преобразования переменных?

5. Как соотносятся задачи группировки и задачи кластер-анализа?

6. В табл.7 приведены попарные расстояния между десятью социально-психологическими признаками способных к математике школьников [11]. Примените к этим данным алгоритмы ближнего соседа, средней связи и дальнего соседа. Для каждого из трех алгоритмов выделите наиболее устойчивые разбиения на кластеры.

Таблица 7.

Попарные расстояния между социально-психологическими признаками.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 1028 | | | | | | | | |
| 3 | 1028 | 608 | | | | | | | |
| 4 | 1050 | 688 | 610 | | | | | | |
| 5 | 1012 | 686 | 636 | 634 | | | | | |
| 6 | 1006 | 566 | 538 | 616 | 562 | | | | |
| 7 | 1012 | 1026 | 748 | 692 | 774 | 732 | | | |
| 8 | 960 | 1088 | 1144 | 1122 | 1120 | 1130 | 1110 | | |
| 9 | 1026 | 878 | 874 | 830 | 836 | 802 | 904 | 1040 | |
| 10 | 990 | 744 | 674 | 744 | 718 | 580 | 814 | 1090 | 830 |

7. Расскажите о динамике индекса инфляции в России.

## Темы докладов, рефератов, исследовательских работ

1. Примеры практического использования методов многомерного статистического анализа.

2. Для непараметрической модели метода наименьших квадратов в случае линейной функции одной переменной разработайте алгоритмы

а) расчета доверительных границ для коэффициентов модели;

б) проверки гипотез относительно этих коэффициентов.

3. Докажите, что сумма исходных значений зависимой переменной должны быть равна сумме восстановленных значений.

4. Критерии качества регрессионной модели.

5. Использование непараметрических оценок плотности для восстановления зависимости.

6. Теоремы умножения и сложения для индекса инфляции.

7. Экспериментальная работа: соберите данные о ценах и рассчитайте индекс инфляции для своего региона (данные о потребительской корзине ИВСТЭ и ценах на базовый момент времени приведены в [6]).

8. Учет инфляции при проведении анализа финансово-хозяйственной деятельности предприятия.

## 3.3. Статистика временных рядов

Под временными рядами понимают детерминированные или случайные функции от времени. При этом время предполагается дискретным, в противном случае говорят о случайных процессах, а не о временных рядах.

### 3.3.1. Методы анализа и прогнозирования временных рядов

**Модели стационарных и нестационарных временных рядов.** Пусть $t = 0,\pm1,\pm2,\pm3,...$ Рассмотрим временной ряд $X(t)$. Пусть сначала временной ряд принимает числовые значения. Это могут быть, например, цены на батон хлеба в соседнем магазине или курс обмена доллара на рубли в ближайшем обменном пункте. Обычно в поведении временного ряда выявляют две основные тенденции - тренд и периодические колебания.

При этом под трендом понимают зависимость от времени линейного, квадратичного или иного типа, которую выявляют тем или иным способом сглаживания (например, экспоненциального сглаживания) либо расчетным путем, в частности, с помощью метода наименьших квадратов. Другими словами, тренд - это очищенная от случайностей основная тенденция временного ряда.

Временной ряд обычно колеблется вокруг тренда, причем отклонения от тренда часто обнаруживают правильность. Часто это связано с естественной или назначенной периодичностью, например, сезонной или недельной, месячной или квартальной (например, в соответствии с графиками выплаты заплаты и уплаты налогов). Иногда наличие периодичности и тем более ее причины неясны, и задача статистика - выяснить, действительно ли имеется периодичность.

Элементарные методы оценки характеристик временных рядов обычно достаточно подробно рассматриваются в курсах "Общей теории статистики" (см., например, учебники [1, 2]), поэтому нет необходимости подробно разбирать их здесь. О некоторых современных методах оценивания длины периода и самой периодической составляющей речь пойдет ниже в подразделе 3.3.2.

*Характеристики временных рядов.* Для более подробного изучения временных рядов используются вероятностно-статистические модели. При этом временной ряд $X(t)$ рассматривается как случайный процесс (с дискретным временем). Основными характеристиками $X(t)$ являются *математическое ожидание* $X(t)$, т.е.

$$a(t) = MX(t),$$

*дисперсия* $X(t)$, т.е.

$$\sigma^2(t) = DX(t)$$

и *автокорреляционная функция* временного ряда $X(t)$

$$\rho(t,s) = \frac{M(X(t) - a(t))(X(s) - a(s))}{\sigma(t)\sigma(s)},$$

т.е. функция двух переменных, равная коэффициенту корреляции между двумя значениями временного ряда $X(t)$ и $X(s)$.

В теоретических и прикладных исследованиях рассматривают широкий спектр моделей временных рядов. Выделим сначала *стационарные* модели. В них совместные функции распределения $F(t_1,t_2,...,t_k)$ для любого числа моментов времени $k$, а потому и все перечисленные выше характеристики временного ряда *не меняются со временем*. В частности, математическое ожидание и дисперсия являются постоянными величинами, автокорреляционная функция зависит только от разности $t$ - $s$. Временные ряды, не являющиеся стационарными, называются *нестационарными.*

*Линейные регрессионные модели с гомоскедастичными и гетероскедастичными, независимыми и автокоррелированными остатками.* Как видно из сказанного выше, основное - это "очистка" временного ряда от случайных отклонений, т.е. оценивание математического ожидания. В отличие от простейших моделей регрессионного анализа, рассмотренных в главе 3.2, здесь естественным образом появляются более сложные модели. Например, дисперсия

может зависеть от времени. Такие модели называют гетероскедастичными, а те, в которых нет зависимости от времени - гомоскедастичными. (Точнее говоря, эти термины могут относиться не только к переменной "время", но и к другим переменным.)

Далее, в главе 3.2 предполагалось, что погрешности независимы между собой. В терминах настоящей главы это означало бы, что автокорреляционная функция должна быть вырожденной - равняться 1 при равенстве аргументов и 0 при их неравенстве. Ясно, что для реальных временных рядов так бывает отнюдь не всегда. Если естественный ход изменений наблюдаемого процесса является достаточно быстрым по сравнению с интервалом между последовательными наблюдениями, то можно ожидать "затухания" автокорреляции" и получения практически независимых остатков, в противном случае остатки будут автокоррелированы.

*Идентификация моделей.* Под идентификацией моделей обычно понимают выявление их структуры и оценивание параметров. Поскольку структура - это тоже параметр, хотя и нечисловой, то речь идет об одной из типовых задач прикладной статистики - оценивании параметров.

Проще всего задача оценивания решается для линейных (по параметрам) моделей с гомоскедастичными независимыми остатками. Восстановление зависимостей во временных рядах может быть проведено на основе методов наименьших квадратов и наименьших модулей оценивания параметров в моделях линейной (по параметрам) регрессии. На случай временных рядов переносятся результаты, связанные с оцениванием необходимого набора регрессоров, в частности, легко получить предельное геометрическое распределение оценки степени тригонометрического полинома.

Однако на более общую ситуацию такого простого переноса сделать нельзя. Так, например, в случае временного ряда с гетероскедастичными и автокоррелированными остатками снова можно воспользоваться общим подходом метода наименьших квадратов, однако система уравнений метода наименьших квадратов и, естественно, ее решение будут иными. Формулы в терминах матричной алгебры, о которых упоминалось в главе 3.2, будут отличаться. Поэтому рассматриваемый метод называется "*обобщенный метод наименьших квадратов* (ОМНК)".

*Замечание.* Как уже отмечалось в главе 3.2, простейшая модель метода наименьших квадратов допускает весьма далекие обобщения, особенно в области системам одновременных эконометрических уравнений для временных рядов. Для понимания соответствующей теории и алгоритмов необходимо владение методами матричной алгебры. Поэтому мы отсылаем тех, кому это интересно, к литературе по системам эконометрических уравнений [3, 4] и непосредственно по временным рядам [5, 6], в которой особенно много интересуются спектральной теорией, т.е. выделением сигнала из шума и разложением его на гармоники. Подчеркнем еще раз, что за каждой главой настоящей книги стоит большая область научных и прикладных исследований, вполне достойная того, чтобы посвятить ей много усилий. Однако из-за ограниченности объема книги мы вынуждены изложение сделать конспективным.

**Системы эконометрических уравнений.** В качестве первоначального примера рассмотрим эконометрическую модель временного ряда, описывающего рост индекса потребительских цен (индекса инфляции). Пусть $I(t)$ - рост цен в месяц $t$ (подробнее об этой проблематике см. главу 7 в [7]). По мнению некоторых экономистов естественно предположить, что

$$I(t) = cI(t\text{-}1) + a + bS(t\text{-}4) + e, \qquad (1)$$

где $I(t\text{-}1)$ - рост цен в предыдущий месяц (а $c$ - некоторый коэффициент затухания, предполагающий, что при отсутствии внешний воздействий рост цен прекратится), $a$ - константа (она соответствует линейному изменению величины $I(t)$ со временем), $bS(t\text{-}4)$ - слагаемое, соответствующее влиянию эмиссии денег (т.е. увеличения объема денег в экономике страны, осуществленному Центральным Банком) в размере $S(t\text{-}4)$ и пропорциональное эмиссии с коэффициентом $b$, причем это влияние проявляется не сразу, а через 4 месяца; наконец, e - это неизбежная погрешность.

Модель (1), несмотря на свою простоту, демонстрирует многие характерные черты гораздо более сложных эконометрических моделей. Во-первых, обратим внимание на то, что

некоторые переменные определяются (рассчитываются) внутри модели, такие, как $I(t)$. Их называют *эндогенными (внутренними).* Другие задаются извне (это *экзогенные* переменные). Иногда, как в теории управления, среди экзогенных переменных, выделяют *управляемые* переменные - те, с помощью выбора значений которых можно привести систему в нужное состояние.

Во-вторых, в соотношении (1) появляются переменные новых типов - с лагами, т.е. аргументы в переменных относятся не к текущему моменту времени, а к некоторым прошлым моментам.

В-третьих, составление эконометрической модели типа (1) - это отнюдь не рутинная операция. Например, запаздывание именно на 4 месяца в связанном с эмиссией денег слагаемом $bS(t\text{-}4)$ - это результат достаточно изощренной предварительной статистической обработки. Далее, требует изучения вопрос зависимости или независимости величин $S(t\text{-}4)$ и $I(t)$ в различные моменты времени $t$. От решения этого вопроса зависит, как выше уже отмечалось, конкретная реализация процедуры метода наименьших квадратов.

С другой стороны, в модели (1) всего 3 неизвестных параметра, и постановку метода наименьших квадратов выписать нетрудно:

$$f(a,b,c) = \sum_{1\le t\le k} (I(t) - cI(t-1) - a - bS(t-4))^2.$$

**Проблема идентифицируемости.** Представим теперь модель тапа (1) с большим числом эндогенных и экзогенных переменных, с лагами и сложной внутренней структурой. Вообще говоря, ниоткуда не следует, что существует хотя бы одно решение у такой системы. Поэтому возникает не одна, а две проблемы. Есть ли хоть одно решение (проблема идентифицируемости)? Если да, то как найти наилучшее решение из возможных? (Это - проблема статистической оценки параметров.)

И первая, и вторая задача достаточно сложны. Для решения обеих задач разработано множество методов, обычно достаточно сложных, лишь часть из которых имеет научное обоснование. В частности, достаточно часто пользуются статистическими оценками, не являющимися состоятельными (строго говоря, их даже нельзя назвать оценками).

Коротко опишем некоторые распространенные приемы при работе с системами линейных эконометрических уравнений.

**Система линейных одновременных эконометрических уравнений.** Чисто формально можно все переменные выразить через переменные, зависящие только от текущего момента времени. Например, в случае уравнения (1) достаточно положить

$$H(t) = I(t\text{-}1),\ G(t) = S(t\text{-}4).$$

Тогда уравнение примет вид

$$I(t) = cH(t) + a + bG(t) + e. \qquad (2)$$

Отметим здесь же возможность использования регрессионных моделей с переменной структурой путем введения фиктивных переменных. Эти переменные при одних значениях времени (скажем, начальных) принимают заметные значения, а при других - сходят на нет (становятся фактически равными 0). В результате формально (математически) одна и та же модель описывает совсем разные зависимости.

**Косвенный, двухшаговый и трехшаговый методы наименьших квадратов.** Как уже отмечалось, разработана масса методов эвристического анализа систем эконометрических уравнений. Они предназначены для решения тех или иных проблем, возникающих при попытках найти численные решения систем уравнений.

Одна из проблем связана с наличием априорных ограничений на оцениваемые параметры. Например, доход домохозяйства может быть потрачен либо на потребление, либо на сбережение. Значит, сумма долей этих двух видов трат априори равна 1. А в системе эконометрических уравнений эти доли могут участвовать независимо. Возникает мысль оценить их методом наименьших квадратов, не обращая внимания на априорное ограничение, а потом подкорректировать. Такой подход называют косвенным методом наименьших квадратов.

Двухшаговый метод наименьших квадратов состоит в том, что оценивают параметры отдельного уравнения системы, а не рассматривают систему в целом. В то же время трехшаговый метод наименьших квадратов применяется для оценки параметров системы

одновременных уравнений в целом. Сначала к каждому уравнению применяется двухшаговый метод с целью оценить коэффициенты и погрешности каждого уравнения, а затем построить оценку для ковариационной матрицы погрешностей. После этого для оценивания коэффициентов всей системы применяется обобщенный метод наименьших квадратов.

Менеджеру и экономисту не следует становиться специалистом по составлению и решению систем эконометрических уравнений, даже с помощью тех или иных программных систем, но он должен быть осведомлен о возможностях этого направления эконометрики, чтобы в случае производственной необходимости квалифицированно сформулировать задание для специалистов по прикладной статистике.

От оценивания тренда (основной тенденции) перейдем ко второй основной задаче эконометрики временных рядов - оцениванию периода (цикла).

### 3.3.2. Оценивание длины периоды и периодической составляющей

Рассмотрим достаточно широкий класс практически полезных непараметрических оценок длины периода и периодической составляющей во временных рядах. Из общих результатов об асимптотическом поведении решений экстремальных статистических задач (см. главу 2.2) вытекает состоятельность этих оценок.

Во многих прикладных задачах рассматривают временной ряд (или случайный процесс)

$$y(t)=x(t)+e(t),$$

где $x(t)$ - детерминированная периодическая функция от времени $t$, т.е. $x(t)=x(t+T)$ при некотором $T$, где $T$ - длина периода (минимальная из возможных, поскольку $2T, 3T, 4T$ - тоже, как легко видеть, длины периодов), а $e(t)$ - “шумы”, случайные погрешности, искажающие периодический сигнал. Требуется оценить (минимальную) длину периода $T=T_0$ и периодическую составляющую $x(t)$. При этом не предполагается, что функция $x(t)$ входит в какое-либо параметрическое семейство, например, конечных сумм синусов и косинусов, т.е. рассматривается задача непараметрического оценивания (минимальной) длины периода и периодической составляющей сигнала.

Приведем примеры прикладных постановок.

1. По акустическим сигналам необходимо установить тип двигателя (и его национальную принадлежность). Предполагается, что двигатели различаются по длине периода и виду основного периодического сигнала. Процедура идентификации основана на оценивании длины периода и периодической составляющей регистрируемого сигнала. Очевидна важность такой задачи при быстрой технической диагностике. В частности, высокая производительность, а потому и высокая экономическая эффективность при ремонте впрямую зависят от умения решать поставленную задачу. Не менее важно по шуму двигателя подводной лодки определить ее тип и национальную принадлежность.

2. В предположении цикличности экономических процессов требуется по статистическим данным установить длину цикла и на основе вида периодической составляющей построить прогноз, например, прогноз урожайности, емкости рынка тех или иных товаров или экономической активности в целом. Часто говорят об экономических циклах, но почти никогда не дают строгого определения понятия цикла. (Под строгим определением понимаем такое, согласно которому можно отличить "цикл" от "не цикла", можно выделить начало и конец цикла, отделить один цикл от другого, короче, однозначно выделить цикл как самостоятельный объект экономического изучения.)

3. По мнению авторов работы [8], для среднесрочного прогнозирования развития социокультурной сферы (социально-политического “климата”, живописи, музыки, архитектуры, поэзии и т.д.) необходимо выявить ее цикличность с помощью объективных измерений на базе субъективных первичных данных (т.е. на базе оценок экспертов).

4. В исторических событиях, описываемых согласно распространенной в настоящее время т.н. скалигеровской хронологии, автор работы [9] обнаруживает цикличность. Эта цикличность полностью объясняется новой статистической хронологией (см., например, [10]), построенной с помощью специальных методов статистики объектов нечисловой природы (см.

главу 3.4), предназначенных для анализа текстов исторических хроник, и одновременно служит еще одним подтверждением этой хронологии.

**Описание метода оценивания.** Пусть рассматриваемые функции $y(t)$, $x(t)$, $e(t)$ определены на отрезке [0; $A$]. При фиксированном $T$ рассмотрим "куски" сигнала $y(t)$ на последовательных отрезках длины $T$, т.е. на отрезках [$0;T$], [$T;2T$], [$2T;3T$], ... Удобно ввести последовательность функций на отрезке [$0;T$], полученную сдвигами этих кусков к началу координат:

$$y_1(t)=y(t),\ y_2(t)=y(t+T),\ y_3(t)=y(t+2T),\ ...$$

Все они определены на отрезке [$0;T$]. Число этих функций равно числу полных периодов длины $T$, укладывающихся на отрезке [0;$A$], т.е. равно целой части числа $A/T$. Отметим еще раз, что если $T$ - период, то $2T$, $3T$, $4T$, ... - тоже периоды. В дальнейшем из всех периодов будем рассматривать и оценивать, как правило, только наименьший.

Если $T=T_0$ - истинный период (или кратный ему) и погрешности $e(t)$ отсутствуют, то все введенные в предыдущем абзаце функции совпадают между собой и с периодической составляющей:

$$x(t)=y_1(t)=y_2(t)=y_3(t)=...$$

при всех $t$ из [0;$T$]. При наличии погрешностей полного совпадения не будет. Однако отклонения определяются лишь шумами в различные моменты времени. При этом в качестве оценки периодической составляющей $x(t)$ естественно взять среднее арифметическое $y_{cp}(t)$ функций $y_1(t)$, $y_2(t)$, $y_3(t)$, ... (могут быть использованы и другие виды средних величин).

Если же $T$ отличается от истинного периода $T_0$ (и кратных ему величин), то различия функций $y_1(t)$, $y_2(t)$, $y_3(t)$, ... между собой определяются также и различием значений $x(t)$ в точках, отстоящих друг от друга на интервалы, длина которых кратна $T$.

В предположении отсутствия погрешностей (т.е. когда $e(t)$ тождественно равно 0) рассмотрим поведение функции $y_{cp}(t)$ на отрезке [0;$T$] при росте длины интервала $A$ наблюдения сигнала, а потому и при росте числа периодов - целой части числа $A/T$. Если $T = T_0$ или $T$ кратно $T_0$, то, как уже сказано, $y_{cp}(t)$ совпадает с периодической составляющей $x(t)$. Если число $T/T_0$ иррационально, то можно показать, что значения $t+mT(\mathrm{mod}T_0)$, где $m$ - натуральные числа такие, что $t+mT<A$, асимптотически (при росте $A$) равномерно заполняют отрезок [0;$T_0$], а потому при выполнении соответствующих условий регулярности, например, непрерывности периодической составляющей сигнала, функция $y_{cp}(t)$ приближается к константе - среднему значению периодического сигнала $x(t)$, т.е. интегралу от $x(t)$ по отрезку [0;$T_0$], деленному на $T_0$. При этом при конечных $A$ функция $y_{cp}(t)$ отлична от константы. (Здесь запись $t+mT(\mathrm{mod}T_0)$ означает теоретико-числовое сравнение по модулю $T_0$ , т.е. взятие дробной части от числа $(t+mT)/T_0$, что соответствует вычитанию соответствующего количества целых периодов $T_0$.

Если же число $T/T_0$ рационально, то наблюдаем промежуточный случай по сравнению с двумя описанными выше, в котором $y_{cp}(t)$, как можно показать, приближается к периодической функции с периодом $T=T_0/n$ при некотором натуральном $n$. Эта функция получена усреднением $n$ последовательных участков длины $T_0/n$ периодического сигнала $x(t)$. Она не является константой, хотя разброс ее значений меньше, чем для исходного периодического сигнала, поскольку $T_0$ - минимальная длина периода.

Из сказанного вытекает, что для оценивания $T$ целесообразно ввести два показателя: показатель разброса $F(T;Y)=F(T;\ y_1(t),\ y_2(t),\ y_3(t),\ ...)$ множества функций $\{y_1(t),\ y_2(t),\ y_3(t),\ ...\}$ на отрезке [0;$T$] и показатель размаха $G(T;Y)=G(T,\ y_{cp}(t))$ функции $y_{cp}(t)$ на отрезке [0;$T$].

Символ $Y$ означает здесь, что показатели разброса и размаха строятся по функции $y(t)$.) При этом показатель разброса нацелен на оценку различий в значениях семейства функций при одном и том же значении аргумента. А показатель размаха - на различие значений одной и той же функции при различных значениях аргумента. Ниже выписан ряд формул для этих показателей в случае непрерывного времени. Для дискретного времени их можно адаптировать двумя способами: либо заменив sup на max, а интеграл на сумму; либо расширив область определения используемых функций на весь отрезок, например, соединив соседние точки отрезками или использовав для заполнения пропусков сплайны более высокого порядка.

В качестве оценки длины периода по фиксированным показателям разброса $F(T;Y)$ и размаха $G(T;Y)$ представляется рациональным использовать то $T$, при котором отношение

$F(T;Y)/G(T;Y)$ **впервые** (при росте $T$, начиная с 0) достигает минимума. Впервые - поскольку величины, кратные периоду, сами являются периодами. Поскольку показатели разброса $F(T;Y)$ и размаха $G(T;Y)$ могут быть выбраны многими разными способами, можно указанным выше способом построить целое семейство алгоритмов оценивания длины периода. С каждым из которых может быть связано семейство методов оценивания периодической составляющей путем того или иного способа усреднения функций $y_1(t)$, $y_2(t)$, $y_3(t)$, ...

**Показатели разброса и размаха.** Ввести показатели разброса $F(T;Y)=F(T; y_1(t), y_2(t), y_3(t), ...)$ можно разными способами. Пусть $k=[A/T]$. Можно использовать различные функционалы супремумного типа (здесь и далее число слагаемых $k$ не будем указывать в обозначении функционалов). Первым рассмотрим максимальный разброс непосредственно между значениями функций:

$$F_1(T,Y) = \sup\{|\, y_i(t) - y_j(t)\,|, i, j = 1,2,...,k, 0 \le t \le T\}.$$

Второй функционал супремумного типа будет учитывать не произвольные отклонения, а только отклонения от "средней функции", т.е. иметь вид

$$F_2(T,Y) = \sup\{|\, y_i(t) - y_{cp}(t)\,|, i = 1,2,...,k, 0 \le t \le T\}.$$

Третий функционал показывает, какую зону "заметают" значения функций:

$$F_3(T,Y) = \sup\{y_i(t), i = 1,2,...,k, 0 \le t \le T\} - \inf\{y_i(t), i = 1,2,...,k, 0 \le t \le T\}.$$

Для применения функционалов интегрального типа целесообразно сделать замену переменной $q=t/T$ и перейти к функциям $Y_i(q) = y_i(t) = y_i(qT)$, $i = 1, 2, ..., k$, $Y_{cp}(q) = y_{cp}(t) = y_{cp}(qT)$, определенным на отрезке [0;1]. В качестве показателя разброса представляется полезным рассмотреть то или иное отклонение совокупности функций $Y_i(q)$, $i=1, 2, ..., k$, друг относительно друга. Можно сказать, что эти функции заполняют некую “трубку”, которая тоньше всего при истинном значении периода $T$, а внутри нее проходит периодическая составляющая $X(q)=x(t)=x(qT)$. Естественно рассмотреть различные функционалы интегрального типа. Например, можно проинтегрировать максимум модулей попарных разностей:

$$F_4(T,Y) = \int_0^1 \max\{|\, Y_i(q) - Y_j(q)\,|, i, j = 1,2,...,k\} dq.$$

Вместо максимума можно проинтегрировать сумму:

$$F_5(T,Y) = \int_0^1 \sum_{i,j=1}^{k} |\, Y_i(q) - Y_j(q)\,|\, dq.$$

Как и для функционалов супремумного типа, естественно рассмотреть показатели разброса относительно "средней функции":

$$F_6(T,Y) = \int_0^1 \max\{|\, Y_i(q) - Y_{cp}(q)\,|, i = 1,2,...,k\} dq,$$

$$F_7(T,Y) = \int_0^1 \sum_{i=1}^{k} |\, Y_i(q) - Y_{cp}(q)\,|\, dq.$$

Следующие четыре функционала, используемые как показатели разброса, аналогичны четырем предыдущим, но включают в себя расчет квадратов:

$$F_8(T,Y) = \int_0^1 [\max\{|\, Y_i(q) - Y_j(q)\,|, i, j = 1,2,...,k\}]^2 dq,$$

$$F_9(T,Y) = \int_0^1 \sum_{i,j=1}^{k} \{Y_i(q) - Y_j(q)\}^2 dq,$$

$$F_6(T,Y) = \int_0^1 [\max\{|\, Y_i(q) - Y_{cp}(q)\,|, i = 1,2,...,k\}]^2 dq,$$

$$F_7(T,Y) = \int_0^1 \sum_{i=1}^{k} \{Y_i(q) - Y_{cp}(q)\}^2 dq.$$

Список показателей разброса можно существенно расширить. В частности, естественно использовать также расстояния в функциональных пространствах $L^p$ при произвольных $p \geq 1$. А для оценивания периодической составляющей применять не только среднее арифметическое, но и другие виды средних величин.

Показатели размаха также можно ввести самыми различными способами. Например, можно рассмотреть такой показатель:

$$G_1(T,Y) = \sup\{|\, y_{cp}(t) - y_{cp}(s)\,|,\ 0 \le t \le T,\ 0 \le s \le T\} =$$

$$= \sup\{y_{cp}(t),\ 0 \le t \le T\} - \inf\{y_{cp}(t),\ 0 \le t \le T\}.$$

Пусть сделана замена переменной $q=t/T$ и осуществлен переход к функции $Y_{cp}(q)=y_{cp}(t)=y_{cp}(qT)$. Возможными показателями размаха являются:

$$G_2(T,Y) = \int_0^1 \int_0^1 |\, Y_{cp}(q) - Y_{cp}(r)\,|\, dqdr,$$

$$G_2(T,Y) = \int_0^1 \int_0^1 (Y_{cp}(q) - Y_{cp}(r))^2\, dqdr.$$

Введем среднее значение оценки периодической составляющей:

$$Y_{cp} = \int_0^1 Y_{cp}(q)dq.$$

К естественным показателям размаха относятся, например, такие:

$$G_4(T,Y) = \sup\{|\, Y_{cp}(q) - Y_{cp}\,|,\ 0 \le q \le 1\},$$

$$G_5(T,Y) = \int_0^1 |\, Y_{cp}(q) - Y_{cp}\,|\, dq,$$

$$G_6(T,Y) = \int_0^1 (Y_{cp}(q) - Y_{cp})^2\, dq.$$

Список показателей размаха, как и список показателей разброса, можно значительно расширить. В частности, естественно использовать расстояния в функциональных пространствах $L^p$ при произвольном $p \geq 1$. А для оценивания периодической составляющей применять не только среднее арифметическое, но и другие виды средних - медиану, среднее геометрическое и др. Вопрос о выборе наилучших (в каком-либо смысле) показателей размаха и разброса здесь не обсуждается. Некоторые из причин этого отказа от оптимизации системы показателей рассмотрены ниже.

**Алгоритмы оценивания.** С прикладной точки зрения остается численно минимизировать один или несколько из 66 описанных выше функционалов $F_i(T;Y)/G_j(T;Y)$, $i = 1, 2, ..., 11$, $j = 1, 2, ..., 6$.

Численная минимизация по одному параметру (возможной длине периода) для современных ЭВМ не вызывает проблем, даже если попросту перебирать возможные значения периода с шагом 0,001. По нескольким реальным или смоделированным сигналам можно установить, какой из функционалов позволяет оценить период и периодическую составляющую реально встречающихся сигналов наиболее точно. Возможно и одновременное использование всех или части функционалов, что в соответствии с методологией устойчивости (см. главу 1.4) позволяет установить чувствительность оценок к выбору метода оценивания, найти интервал их разброса. Проведенные в Институте высоких статистических технологий и эконометрики расчеты по реальным и смоделированным данным о временных рядах показали, что описанные выше алгоритмы позволяют оценивать длину периода и восстанавливать периодическую составляющую временного ряда достаточно точно с практической точки зрения.

В обширной литературе по временным рядам (см., например, монографии [5-7], дающие представление обо всем массиве литературы по этой тематике) проблеме оценивания периода не уделяется большого внимания. Фактически рекомендуют пользоваться либо периодограммой, либо автокорреляционной функцией. С помощью периодограммы (несостоятельной оценки спектральной плотности) можно выделить лишь синусоидальные составляющие, в то время как

в кратко рассмотренных выше прикладных задачах периодическая составляющая представляет интерес сама по себе, без разложения на гармоники. Вторая рекомендация более полезна. В качестве оценки периода можно взять наименьшее положительное число, в котором достигается локальный максимум автокорреляционной функции. Эмпирический коэффициент автокорреляции - еще один функционал типа тех, что перечислены выше.

При поверхностном взгляде на проблемы статистического оценивания, как и на иные проблемы прикладной математики, часто возникает желание обсудить "оптимальность" тех или иных процедур. При более глубоком анализе становятся очевидными два обстоятельства. Во-первых, оптимальность имеет быть лишь в рамках той или иной теоретической модели, при отклонениях от которой оптимальность оценки, как правило, пропадает. Например, выборочное среднее арифметическое как оценка математического ожидания случайной величины оптимальна тогда и только тогда, когда распределение результатов наблюдений - гауссово (доказательство этого утверждения приведено в монографии [11]). С другой стороны, для практически любой статистической процедуры можно подобрать свойство оптимальности так, чтобы эта процедура оказалась оптимальной (как подобрать - это уже дело профессионала). Так, например, метод наименьших модулей оптимален, если погрешности имеют распределение Лапласа, а метод наименьших квадратов - когда их распределение гауссово. Поскольку реальные распределения - не Лапласа и не Гаусса, то указанные математические результаты не могут иметь большого практического значения.

Однако представляется полезным получить доказательства состоятельности оценок изучаемых параметров в возможно более широких, например, непараметрических, постановках. Хотя на основе самого факта сходимости нельзя оценить близость оценок к интересующим исследователя параметрам, но получение доказательства состоятельности - первый шаг при изучении скорости сходимости (см. подраздел 1.4.7).

**Состоятельность оценок.** Наиболее общий подход к установлению асимптотического поведения решений экстремальных статистических задач развит в статистике нечисловых данных для случая пространств произвольной природы (см. главу 2.2). Согласно этому подходу сначала при фиксированном $T$ доказывается сходимость (по вероятности) при $A \to \infty$ значений функционала (показателя разброса) к некоторой предельной функции, а затем проверяются условия, обеспечивающие сходимость Argmin допредельного случайного процесса к Argmin этой детерминированной функции.

Свойства алгоритмов приходится изучать в рамках тех или иных вероятностно-статистических моделей. Моделей может быть много. Достаточно вспомнить историю Центральной Предельной Теоремы (ЦПТ) теории вероятностей. Она на протяжении более 200 лет доказывалась во все более и более широких условиях, вплоть до необходимых и достаточных условий Линдеберга - Феллера (после чего начались обобщения на зависимые слагаемые, на суммы случайных элементов гильбертовых пространств и др.). Отметим, что иногда математические модели далеко выходят за пределы, достаточные для обоснования алгоритмов анализа реальных данных. Так, почти всегда распределения реальных величин дискретны и финитны, а потому, в частности, существуют все моменты. Однако условия финитности и дискретности в вероятностно-статистических моделях часто необоснованно ослабляются. В результате возникают проблемы, не имеющие отношения к реальным данным, например, связанные с измеримостью относительно тех или иных сигма-алгебр. Поэтому ограничимся здесь наиболее простыми моделями из адекватных реальным постановкам. Считаем, что читатель знаком с основными определениями, относящимися к теории случайных процессов.

*Теорема 1.* Пусть случайный процесс $e(t)$ имеет нулевое математическое ожидание, является стационарным и эргодическим (т.е. выполнена теорема Биркгофа-Хинчина) с непрерывными траекториями. Тогда при фиксированном $T$ и $A \to \infty$ имеем

$$\sup\{|E_{cp}(q)|,\ 0 \le q \le 1\} \to 0$$

(сходимость по вероятности), где $E_{cp}(q) = Y_{cp}(q) - X_{cp}(q)$, т.е. $E_{cp}(q)$ - среднее арифметическое погрешностей $e(qT)$, $e(qT+T)$, $e(qT+2T)$,...

Доказательство теоремы 1 проводится стандартными методами теории стационарных временных рядов (с шагом $T$) с использованием известного условия достаточно быстрого

убывания элементов матрицы Лорана по мере удаления от ее главной диагонали (т.е. условия, необходимого и достаточного для справедливости теоремы Биркгофа-Хинчина). С помощью теоремы 1 можно найти асимптотику введенных выше показателей разброса и размаха.

*Теорема 2.* В предположениях теоремы 1 при фиксированном $T$ и $A \to \infty$ пронормированные показатели разброса $F_i(T;Y)$ для наблюдаемого сигнала $Y$ сближаются по распределению с соответствующими положительными случайными величинами $W_i(T, X, \omega)$, зависящими от $T$, характеристик случайного процесса $e(t)$ и периодической составляющей $X$, т.е. существуют числовые последовательности $s_i(k)$ такие, что

$$s_i(k)F_i(T,Y) \Rightarrow W_i(T,X,\omega), \quad i = 1,2,...,11.$$

Доказательство теоремы 2 проводится с помощью достаточно трудоемких (в частности, из-за числа функционалов), но стандартных рассуждений. Они относятся к теории случайных процессов как части теории вероятностей. Эти рассуждения посвящены максимумам (не супремумам, т.к. траектории функции $x(t)$ и случайного процесса $e(t)$ непрерывны) случайных процессов и интегралам от них, с использованием принципа инвариантности (см., например, учебное пособие [12]) и ряда результатов теории стационарных случайных процессов (см., например, монографию [13]). Таким образом, пронормированные функционалы разброса асимптотически не зависят от числа слагаемых - в этом и состоит основной смысл теоремы 2.

*Теорема 3.* В предположениях теоремы 1 при фиксированном $T$ и $A \to \infty$ показатели размаха для наблюдаемого сигнала $Y$ сближаются с соответствующими показателями для периодической составляющей $X$, т.е.

$$G_j(T,Y) - G_j(T,X) \to 0, \quad j = 1,2,...,6.$$

Для доказательства используются стандартные оценки, основанные на виде конкретных функционалов, задающих показатели размаха. В отличие от теоремы 2 предельные показатели детерминированы.

Аналоги теорем 2 и 3 верны также и при использовании (в качестве показателей разброса и размаха) расстояний в функциональных пространствах $L^p$ при произвольном $p \geq 1$. А для оценивания периодической составляющей - не только среднего арифметического, но и других видов средних - медианы, среднего квадратического, среднего геометрического, обобщенных средних по Колмогорову (см. главу 2.1) и др.

*Теорема 4.* Пусть выполнены условия теоремы 1, периодическая составляющая непрерывна и имеет период $T_0$ . Тогда при фиксированном $T$ и $A \to \infty$ показатели разброса (пронормированные) и размаха стремятся к некоторым детерминированным пределам, зависящим только от $T$ и $T_0$ , т.е.

$$s_i(k)F_i(T,Y) \to F_i(T,T_0), \quad i = 1,2,...,11,$$

$$G_j(T,Y) \to G_j(T,X), \quad j = 1,2,...,6.$$

(сходимость по вероятности), минимум каждой из функций $F_i(T;T_0)$, $i$=1, 2, ..., 11, и максимум каждой из функций $G_j(T;T_0)$ , $j$=1, 2, ..., 6, достигается при $T=T_0$ и при $T$, кратных $T_0$, причем у показателей разброса $F_i(T;T_0)$ возможны и иные минимумы, а у показателей размаха $G_j(T;T_0)$ других максимумов нет.

Доказательство вытекает из теорем 2 и 3 и свойств усреднения периодической составляющей при росте длины интервала наблюдения сигнала, описанных в начале настоящего подраздела. Отметим, что предельные значения функционала разброса $F_i(T;T_0)$, вообще говоря, показывают разброс случайной погрешности, другими словами, не всегда зависят от периодической составляющей, а потому из-за нормировки на единичный отрезок в ряде случаев оказываются константами. Вместе с тем численные эксперименты показывают, что отмеченная сходимость к пределу является сравнительно медленной. И минимизация непосредственно функционалов разброса (без учета показателей размаха) при конкретной длине сигнала позволяет достаточно точно выделить периодическую составляющую из массива реальных данных. Однако описанные выше теоретические результаты заставили отказаться от

первоначальной гипотезы о том, что достаточно использовать только показатели разброса, и привели к необходимости скорректировать алгоритмы, введя деление на показатели размаха.

*Теорема 5.* В предположениях теоремы 4 **о**ценки, являющиеся первыми локальными минимумами при минимизации по $T$ отношений одного из 11 перечисленных выше показателей разброса к одному из 6 показателей размаха, являются состоятельными оценками истинного периода $T_0$, а функция $y_{cp}(t)$ является состоятельной оценкой периодической составляющей $x(t)$ на отрезке $[0;T_0]$.

Согласно теоремам 1-4 установлена сходимость (по вероятности) значений допредельных функционалов к предельным при каждом конкретном $T$. Поэтому для доказательства сходимости минимумов допредельных функционалов к минимумам предельных можно воспользоваться общей теорией асимптотического поведения решения экстремальных статистических задач. Условие асимптотической равномерной разбиваемости, сформулированное в работе [14], выполнено, как можно показать, в силу непрерывности траекторий случайного процесса (непрерывного сглаживания для временного ряда) и его периодической составляющей. Откуда и вытекает заключение теоремы 5, дающей теоретико-статистическое обоснование использованию системы описанных выше эвристических алгоритмов оценивания длины периода и периодической составляющей. При известной или достаточно точно оцененной длине периода сама периодическая составляющая естественным образом оценивается с помощью усреднения перенесенных к началу координат кусков временного ряда, и в силу теоремы 1 эта оценка является состоятельной. Затем для получения оценки математического ожидания сигнала на всей области его определения указанную оценку можно периодически продолжить.

*Замечание.* При практическом использовании рассматриваемых алгоритмов целесообразно учитывать дополнительные особенности реальных временных рядов. В частности, обратим внимание на неустойчивость супремумов по отношению к выбросам (резко выделяющимся наблюдениям) сравнительно с функционалами интегрального типа. Бывают ситуации, когда методики или аппаратура, регистрирующие значения реальных временных рядов, могут допускать сбои в отдельные моменты времени. Например, если происходит валютный кризис типа "черного вторника", когда курс доллара по отношению к рублю, строго говоря, не определен, другими словами, с точки зрения экономических агентов одновременно существует масса сильно отличающихся курсов. Аналогичная ситуация бывает и в целом ряде других случаев. Набор подходящих ассоциаций вызывают решения руководства страны об обмене денежных знаков, особенно с дискриминационными составляющими. Во всех подобных ситуациях временные ряды дают резкие выбросы (всплески), которые затем, как правило, сглаживаются. Поэтому целесообразно в качестве показателей разброса и размаха использовать функционалы интегрального типа. Вопросам оценивания длины периода и периодической составляющей посвящены многие публикации, в том числе работа [15].

### 3.3.3. Метод ЖОК оценки результатов взаимовлияний факторов

**Основные идеи метода компьютерного моделирования ЖОК.** При вероятностно-статистическом моделировании технических, социально-экономических, медицинских и иных явлений и процессов исследователь часто сталкивается с тем, что различные факторы постоянно влияют друг на друга. Как правило, для каждого из рассматриваемых факторов можно выделить "непосредственное окружение", которое оказывает на него влияние на него в конкретный момент. На него же этот фактор оказывает некоторое обратное влияние. Дальше начинается самое интересное - волны влияний, порожденные разными факторами, распространяются по всей совокупности факторов, частично усиливают друг друга, частично погашают, порождая в каждый момент времени новые волны.

Разработан компьютерный метод, называемый далее ЖОК, предназначенный для оценки результатов влияния описывающих ситуацию факторов на итоговые показатели и друг на друга. Метод ЖОК позволяет получать выводы, полезные для управления различными структурами на микро- и макроуровнях, от бригад и предприятий до государства в целом. Этот метод использует модель многомерного временного ряда, в которой коэффициенты

непосредственного влияния факторов друг на друга и начальные условия задаются экспертами. Опишем основные составляющие этого метода.

Сначала экспертным путем определяется список факторов, которые необходимо учитывать при анализе конкретной ситуации. В качестве примера рассмотрим здесь типовое промышленное предприятие. Для него такими факторами могут являться, на наш взгляд, устойчивость развития, уровень рентабельности, оценка состояния основных и оборотных фондов, положение на рынке, кадровый потенциал, финансовое положение, технологический уровень, технический уровень и качество продукции, степень учета экологических требований, уровень сертификации, научно-технический потенциал и степень его использования, положение в социальной сфере, развитость профсоюзного движения, оценка отношений с конкурентами и властями, и т.д. Основная часть перечисленных факторов носит качественный характер.

Далее определяются необходимые для работы модели начальные уровни факторов, соответствующие начальному состоянию изучаемого объекта (проводится оцифровка нечисловых переменных). Они оцениваются экспертами на шкале от (-1) до (+1) с шагом 0,1. В методе ЖОК степень привлечения экспертов может быть различна - от использования одного эксперта, хорошо знающего ситуацию и на основе своих знаний и интуиции указывающего необходимые параметры и связи, до подключения к работе комиссии экспертов, коллективно оценивающих указанные параметры и связи, с использованием той или иной схемы сбора и анализа экспертных мнений.

Затем экспертами составляется блок-схема непосредственных влияний факторов друг на друга и оценивается степень непосредственных влияний с помощью такой же шкалы от (-1) до (+1) с шагом 0,1. Получается модель в виде взвешенного ориентированного графа с начальными данными в вершинах. Она несколько напоминает хорошо известную экономистам схему межотраслевого баланса В.Леонтьева, но в отличие от нее использует не только количественные, но - в основном - качественные факторы. Затем просчитываются итерации (опосредованные влияния второго, третьего и т.д. уровней, соответствующие второму, третьему и т.д. моментам времени) вплоть до получения стабильного состояния. Результат работы модели - конечные уровни факторов.

Модель позволяет просчитать развитие экономической структуры при различных сценариях. Обычно одновременно используют три типа сценариев - “Прогноз”, “Поиск” и “Оптимизация”.

Сценарий “Прогноз” показывает результат при отсутствии управляющих воздействий. Он демонстрирует, как будет развиваться ситуация, если в нее не вмешиваться. Исходные данные для сценария “Прогноз” - начальные значения факторов и матрица непосредственных взаимовлияний факторов.

В сценариях типа “Поиск” вводится новое понятие - управляющие факторы. В сценариях этого типа анализируются результаты изменений при наличии тех или иных конкретных воздействий на управляющие факторы. Обычно специалист, работающий с системой ЖОК, имеет целью увеличение значений тех или иных факторов при “удержании” некоторых иных в заданных пределах. Однако от него не требуется сообщать свои цели компьютерной системе. В сценариях типа “Поиск” осуществляется эвристический процесс оптимизации, а также анализ поведения системы при тех или иных воздействиях на начальные значения факторов.

В сценариях типа “Оптимизация” кроме списка управляющих факторов задаются целевые факторы и условия на них, которых необходимо добиться. Обычно это - условия выхода на определенные уровни, например, рентабельность должна быть не менее 0.5, а социальная напряженность - не более 0.3. С помощью оптимизационных алгоритмов находится наилучшее управление, позволяющее достигнуть цели или максимально к ней приблизиться. Однако найденные компьютером рекомендации могут включать слишком резкие изменения тех или иных начальных параметров, поэтому результаты расчетов скорее указывают на перспективные варианты изменения управляющих параметров, чем непосредственно задают план действий. С помощью сценариев типа “Поиск” можно на основе этих результатов найти практически реализуемые рекомендации.

Система ЖОК позволяет проследить динамику изменения значений факторов вплоть до их стабилизации, которая обычно наступает через 15-25 итераций (интервалов времени). Такая

быстрая сходимость вначале кажется неожиданной. Возможно, сам факт стабилизации является самым важным методологическим выводом из экспериментов с моделью ЖОК. После первоначальных всплесков замкнутая экономическая система стабилизируется, хотя бы и на весьма низком уровне производства и потребления.

При этом с помощью оцененных экспертами коэффициентов важности факторов (с учетом знака) можно отслеживать общую оценку экономической ситуации.

Система ЖОК является человеко-машинной. Для эффективной работы специалиста желательно, чтобы общее число факторов, используемых в конкретной модели, не превышало 20, а число непосредственных взаимосвязей - 40, хотя эти ограничения несущественны для математического обеспечения компьютерной системы ЖОК. Однако они существенны для обеспечения наглядности при построении, обсуждении и совершенствовании модели, для того, чтобы факторы и связи между ними можно было изобразить на листе бумаги или экране компьютера в виде блок-схемы.

Система ЖОК с успехом использовалась для анализа ряда конкретных экономических ситуаций. Так, по заказу Минфина РФ она применялась для анализа взаимовлияний факторов, определяющих динамику налогооблагаемой базы и сбора подоходного налога с физических лиц, налога на имущество, налогов и сборов за пользование природными ресурсами и др. Построенная серия моделей обладала некоторыми общими чертами. Прогноз, исходящий из современного экономического положения, во всех случаях указывал на дальнейшее ухудшение ситуации. Активное вмешательство государства в экономику приводило к значительному улучшению показателей, в то время как управление с помощью чисто экономических (монетаристских) методов не позволяло улучшить исходное положение. Полученные результаты подтверждают известную концепцию пяти нобелевских лауреатов по экономике (К.Эрроу, В.Леонтьев и др.), разрабатываемую совместно с Отделением экономики Российской академии наук (Д.С.Львов, С.Ю.Глазьев и др.), о необходимости активного регулирования государством экономических процессов [16, 17].

Другие примеры применения системы ЖОК касались оптимизации экономической стороны деятельности промышленного предприятия или организации в иной сфере, экономических взаимоотношений отраслей народного хозяйства. А также макроэкономического моделирования, в ходе которого удалось вскрыть две неточности в основной схеме известного учебника [18], а затем исправить их, включив дополнительные блоки в соответствующую модель.

**Пример применения метода ЖОК: система моделей налогообложения.** Подоходный налог на физических лиц - один из основных в системе налогообложения, действующей в Российской Федерации. Построена система моделей для изучения влияния различных факторов на налогооблагаемую базу подоходного налога на физических лиц. Построение модели предполагает выявление факторов, влияющих на налогооблагаемую базу, и взаимосвязей между этими факторами, численную оценку взаимовлияний факторов. После формулировки сценариев развития ситуаций последовало проведение расчетов и анализ полученных результатов. Расчеты проводились с помощью оригинального математического и программного обеспечения, разработанного в Институте высоких статистических технологий и эконометрики МГТУ им. Н.Э.Баумана.

Очевидными факторами, влияющими на налогооблагаемую базу подоходного налога, являются:

1) объем производства, соответственно, численность занятых работников и объем ФОТ (фонда оплаты труда),

2) размеры полной, скрытой, частичной безработицы,

3) объем доходов, не облагаемых налогами: от натурального приусадебного (садово-огородного) и домашнего хозяйства, частных услуг в области мелкого ремонта. репетиторства и др.

4) инфляция, частично компенсируемая повышением заработной платы;

5) уклонение от уплаты налога отдельными физическими и юридическими лицами,

6) бартер, "черный нал" (миновавший кассу) и другие явления теневой экономики;

6) объем неплатежей в целом и невыплаты заработной платы в частности.

В учебном пособии [19, с.245-265] перечислено более 50 видов полного или частичного освобождения граждан от уплаты подоходного налога. Перечень этот не полон: например, лицо, купившее квартиру, освобождается от уплаты определенной части подоходного налога (государство частично компенсирует стоимость квартиры).

Надо иметь в виду, что Госкомстат дает не реальное, а рассчитанное распределение населения по величине получаемого дохода (на основе логарифмически нормального закона). Поэтому по данным Госкомстата нельзя рассчитать влияние изменения ставок подоходного налога на суммарный сбор.

В различных странах сложилась различная доля оплаты труда в стоимости изделия. В США в 1988 г. ВНП составил 4862 млрд. долл., в то время как личный доход - 4063 млрд., а индивидуальные налоги - 590 млрд. (14,5% от личных доходов, 12,13% от ВНП). В России в 1995 г. подоходный налог составил 2770,9 млрд. руб., или 2,17 %: от всех налоговых сборов, или примерно 0,26% от ВВП, т.е. доля этого налога примерно в 50 раз меньше, чем в США.

После предварительного анализа налоговой ситуации в России была построена пробная модель динамики экономики, включающая 13 факторов и 32 взаимосвязи между ними. В настоящем подразделе не рассматривается, поскольку включает в себя единый фактор налогообложения, без отдельного выделения подоходного налога с физических лиц.

Далее был рассмотрен список из 41 фактора, которые влияют на налогооблагаемую базу подоходного налога с физических лиц. После анализа из них были отобраны 30 факторов. На их основе построена модель НФЛ-30 (НФЛ - по первым буквам: «Налог на Физических Лиц»). Она включает эти 30 факторов и 465 взаимосвязей между ними. Расчет по модели НФЛ-30 показал, что современная ситуация в России ведет к уходу экономики "в тень", росту сокрытия налогов и коррупции, падению доверия к власти и готовности платить налоги, падению уровня жизни населения и очень сильному сокращению налогооблагаемой базы.

Однако модель НФЛ-30 является слишком сложной и трудной для непосредственного анализа. В частности, трудно было найти рациональное управление, обеспечивающее рост налогооблагаемой базы подоходного налога. Поэтому модель НФЛ-30 была упрощена в основном за счет исключения слабых и/или дублирующих взаимосвязей с переносом внимания на непосредственные взаимодействия. Кроме того, три фактора были исключены, а один добавлен.

В результате построена модель НФЛ-28 с 28 факторами и 64 взаимосвязями. Расчет по этой модели показал наличие больших возможностей у государства в целом и у государственных налоговых органов (ГНО) в частности по расширению налогооблагаемой базы подоходного налога.

Характер взаимодействий факторов в модели НФЛ-28 описывается с помощью ориентированного графа (орграфа), дугам которого приписаны весовые коэффициенты от (-1) до (+1). От конкретного фактора дуги ведут к тем факторам, на которые этот фактор непосредственно влияет. Влияние может быть как положительным, так и отрицательным, возрастание абсолютной величины означает увеличение степени влияния.

Таким образом, построение модели НФЛ-28 изучения роста налогооблагаемой базы подоходного налога с физических лиц , как и других моделей рассматриваемого типа, состоит из ряда операций, осуществляемых экспертами, а именно:

выделения и обоснования системы факторов, включаемых в модель;

оценки важности факторов (по десятибалльной шкале) с учетом знаков - по отношению к задаче расширения налогооблагаемой базы подоходного налога с физических лиц ;

построения ориентированного графа непосредственного влияния факторов друг на друга (важно избегать связей, соответствующих опосредованным влияниям, а также по возможности сократить число циклов);

приписывания дугам весов из интервала [-1; +1], отражающих степень влияния (матрица весов не должна содержать слишком много ненулевых элементов).

Затем с помощью имеющегося диалогового комплекса ЖОК проводятся расчеты по ряду сценариев. Результаты интерпретации итогов расчетов позволяют проанализировать свойства модели.

При описанию конкретных взаимовлияний факторов в моделях рассматриваемого типа используется следующая шкала соответствия между лингвистической и числовой шкалами:

очень сильно возрастает ( 0,9; 1,0);
значительно возрастает ( 0,7; 0,8);
существенно возрастает ( 0,5; 0,6);
умеренно возрастает ( 0,3; 0,4);
очень слабо возрастает (0,1; 0,2);
очень слабо убывает ( - 0,1; - 0,2);
умеренно убывает (- 0,3; - 0,4);
существенно убывает ( - 0,5; - 0,6);
значительно убывает ( - 0,7; - 0,8);
очень сильно убывает ( - 0,9; - 1,0).

Однако и модель НФЛ-28 достаточно сложна для анализа. Поэтому были построены еще две модели - НФЛ-18 и НФЛ-19, множества факторов которых различны.

Основой для анализа служат две модели, обозначаемые как НФЛ-18 и НФЛ-19, т.е. модели подоходного налога на физических лиц с использованием 18 и 19 факторов соответственно. Модель НФЛ-18 с 18 факторами и 31 взаимосвязью межу ними основана на гипотезе об активном участии государства в экономических процессах (с помощью законотворчества, госзаказов и др.). В то время как модель НФЛ-19 с 19 факторами и 31 взаимодействием предполагает лишь косвенное вмешательство государства путем установления ставок таможенных сборов, борьбы с криминальным миром и др. Эта модель построена в основном на чисто экономических взаимодействиях, без непосредственного регулирующего влияния государства.

Начальные значения факторов - это приращения (изменения, "дифференциалы") включенных в модели экономических величин. Они выбраны исходя из оценки экономического положения России в июне 1999 года. Конечные значения факторов определяются рассматриваемым сценарием. Они также интерпретируются как приращения включенных в модели экономических величин.

В каждой из моделей НФЛ-18 и НФЛ-19 рассмотрено четыре сценария:

- прогнозируется динамика налогооблагаемой базы подоходного налога при отсутствии управляющих воздействий (сценарий обозначается как Пассивный-1);
- прогнозируется динамика налогооблагаемой базы подоходного налога при заданных управляющих воздействиях (сценарий обозначается как Активный-1);
- ищутся управляющие воздействия, позволяющие добиться прироста не менее 0.7 налогооблагаемой базы подоходного налога (сценарий обозначается как Цель-1);
- ищутся управляющие воздействия, позволяющие добиться прироста не менее 0.5 налогооблагаемой базы подоходного налога и прироста не менее 0,3 уровня жизни населения (сценарий обозначается как Цель-2).

Таким образом, в сценариях Цель-1 и Цель-2 выделяются целевые факторы, значения которых требуется максимизировать. В сценарии Цель-1 это один фактор - налогооблагаемая база подоходного налога, в сценарии Цель-2 добавляется еще один - уровень жизни населения. В сценарии Активный-1 выделяется другое подмножество факторов - управляющие, путем воздействия на которые специалист, работающий с моделью, может попытаться изменить неблагоприятные тенденции и улучшить ситуацию. В сценариях Цель-1 и Цель-2 управляющие факторы используются несколько по-иному - выбор их значений осуществляет не специалист, а компьютер в соответствии с алгоритмом оптимизации. Выделяются также наблюдаемые факторы, наиболее важные для анализа начальной и конечной экономической ситуации в каждом из сценариев. Во всех сценариях начальные значения факторов соответствуют современному состоянию экономики России.

Используется также количественная оценка экономической ситуации, представляющая собой взвешенную сумму значений факторов. Для ее расчета каждому используемому фактору приписано число от (-10) до 10 - его важность, при этом желательное направление изменения фактора определяет знак указанного числа: если желателен рост, то ставится плюс, если желательно убывание - минус.

Основная часть подраздела начинается с описания факторов, включенных в модели НФЛ-18 и НФЛ-19, и обоснования необходимости их включения в модели. Имеется 12 факторов, включенных в обе модели, поэтому всего используется 25 факторов. Затем приводятся схемы (графы) взаимного влияния факторов и матрицы взаимного влияния факторов в моделях НФЛ-18 и НФЛ-19.

Результаты проведенной работы показывают, что использованный эконометрический аппарат на основе ориентированных графов позволяет получать качественные выводы, полезные для выбора стратегии управления процессами налогообложения.

**Построение моделей НФЛ-18 и НФЛ-19.** Анализ экономической ситуации, нацеленный на изучение динамики налогооблагаемой базы подоходного налога, приводит к выделению следующих блоков факторов: макроэкономические показатели; показатели доходов и уровня жизни населения; участие государства в экономической жизни; факторы, относящиеся непосредственно к сфере налогов; характеристики теневой экономики и криминального мира. Далее выделяются и обосновываются факторы, используемые в моделях НФЛ-18 и НФЛ-19. Поскольку компьютерные модели нацелена на изучение влияния изменений (приращений, дифференциалов) одних факторов на изменения других, то в названиях факторов имеются термины "рост", "прирост" и др. Для компьютерных моделей, имеющих качественный характер, факторы берутся в обобщенном виде, что отмечается при их описании. Модели, охватывающие столь обширную область экономической жизни, не могут не опираться на факторы в обобщенном виде, иначе они станут необозримыми для восприятия и бесполезными для применения.

Приведем сначала описание факторов, используемых в модели НФЛ-18.

1. **Прирост налогооблагаемой базы подоходного налога** - прирост тех доходов населения, с которых согласно действующему законодательству реально может быть взят подоходный налог в денежной форме.
2. **Прирост ВВП** (валового внутреннего продукта) - прирост совокупного объема продукции и услуг, произведенных на территории России, базовый макроэкономический показатель, характеризующий положение и динамику экономики России.
3. **Увеличение роли государства в экономике** связано с созданием эффективно работающего аппарата государственного управления. В том числе налоговых органов; с ростом уровня налогового законодательства (т.е. с ростом согласованности и применимости законов, их обоснованности, соответствия национальным традициям, доступности для понимания масс налогоплательщиков). Например, в части различных льгот по уплате подоходного налога, роста доли зарплаты в цене товара (услуги); управлением финансовой сферой, борьбой с криминальным миром, коррупцией, теневой экономикой, эффективной организацией производством товаров и услуг в государственном секторе, повышением уровня жизни населения через систему трансфертов, и др.
4. **Рост государственных заказов,** в частности, расходов на покупку услуг населения путем организации, расширения или восстановления государственных предприятий, строительства дорог и проведения тех или иных "общественных работ". Согласно современной макроэкономической теории – необходимая мера в период спада.
5. **Рост кредитования отечественных товаропроизводителей** (за счет средств населения, находящихся в банках) - необходимое условие повышения производства отечественных товаров и услуг с целью удовлетворения спроса на них. Кредиты должны быть предназначены как для оптимизации оборотных средств предприятий, исключения бартера и неплатежей, так и для капиталовложений (инвестиций) с целью перехода на современные технологии.
6. **Повышение спроса на отечественную продукцию,** обеспеченное соответствующим повышением ее выпуска, приводит к увеличению

объема средств отечественных производителей, в частности, к увеличению фонда оплаты труда (ФОТ), а потому и поступлений подоходного налога. Речь идет о реализованном, т.е. оплаченном спросе.

7. **Рост уровня работы государственных налоговых органов (ГНО)** предполагает совершенствование налогового законодательства, рост налоговой грамотности населения, индивидуальную работу с каждым налогоплательщиком, рост реальности и неотвратимости санкций за неуплату налогов, широкую информированность общества о них.
8. **Рост качества работы банковской системы** предполагает, своевременное осуществление платежей, существенное снижение необоснованно завышенных плат за банковские услуги, за перевод средств со счета на счет, за пользование банкоматами и др.
9. **Рост доверия населения к государственной власти** и готовности платить налоги становится реальной силой как непосредственно в экономической жизни, так и в борьбе против криминального мира и теневой экономики - основного внутреннего врага государства.
10. **Рост уровня жизни населения** повышает желание и возможность уплаты подоходного налога, а его снижение действует в обратном направлении. Понятие "уровень жизни" многогранно, помимо дохода за последний промежуток времени включает в себя и использование предыдущих накоплений, и способы распоряжения доходом и личной собственностью [20].
11. **Прирост объема** (начисленных) **выплат из ФОТ** предприятий и организаций - основной источник поступлений подоходного налога. Влияние неплатежей сказывается в уменьшении реальных выплат из ФОТ и соответствующего уменьшения поступлений подоходного налога. Влияние изменения объемов других источников доходов, облагаемых подоходным налогом [19], учитывается с помощью связей с показателями уровня жизни и инфляции.
12. **Рост сбережений (накоплений)** населения, находящихся в банках и могущих быть использованными для кредитования отечественных товаропроизводителей - важный показатель доверия к государству, а потому и готовности платить налоги. Он противопоставляется, с одной стороны, практическому отсутствию сбережений у части населения, с другой стороны, хранению сбережений вне банков ("в чулке").
13. **Прирост уровня занятости** населения на официально зарегистрированных работах - понятие, противоположное понятию прироста безработицы. Однако при обсуждении безработицы не всегда ясно, является ли отказ от работы вынужденным или добровольным (домохозяйки с детьми, лица старшего возраста). Имеются расхождения (в несколько раз) при оценке уровня безработицы, например, по числу зарегистрированных на биржах труда и расчетами по правилам МОТ (Международной Организации Труда).
14. **Инфляция** - прирост общего уровня цен (при падении цен он может быть отрицательным), измеряемый по специальным методикам [7].
15. **Прирост неплатежей** (работникам и организациям) дезорганизует экономическую жизнь в целом и налоговую сферу в частности, приводит к приросту бартера**,** к выдаче заработной платы в натуральной форме, в результате - к сокращению налоговых поступлений (в т.ч. подоходного налога) и поступлений во внебюджетные фонды.
16. **Прирост доходов в домохозяйствах, остающихся вне сферы ГНО** - это прирост доходов в натуральной форме (сельскохозяйственная продукция, выращенная на собственном приусадебном или садово-огородном участке, даче для семейного употребления, домашняя работа (ремонт, изготовление

мебели, одежды, пищи)). А также прирост доходов от мелких услуг типа косметических ремонтных работ в квартирах, индивидуального пошива одежды, ремонта обуви, репетиторства.

17. **Прирост сокрытия доходов и уклонения от уплаты налогов** частично связан с нарушениями нормального хода экономической жизни, в том числе с вынужденными неплатежами и бартером, ведущими к выплате заработной платы натурой и выдаче собственности организаций в пользование работникам, с налоговой неграмотностью населения. Но в основном имеет криминальный характер, в частности, порожденный распространением "черного нала".
18. **Рост криминального мира**, **теневой экономики** и коррупции , т.е. сил, противостоящих государственной власти, приводит к "скрытию" от ГНО части экономики (т.н. "теневой") вместе со всеми налогами, которые обязаны были бы платить "спрятавшиеся" организации и лица.

В модели НФЛ-19 отсутствуют факторы 3, 4, 6, 8, 9, 15, но включены следующие семь факторов:

19 **Усиление борьбы государства с криминалом в экономике,** т.е. лишь одна, но наиболее важная часть из сферы **усиления роли государства в экономике** согласно описанному выше фактору 3.

20. **Улучшение финансового положения предприятий,** т.е. увеличение находящихся в распоряжении предприятий денежных средств, позволяет увеличить выпуск продукции и услуг (ВВП), осуществлять инвестиции, уменьшает неплатежи и бартер, увеличивает уровень жизни населения, его денежные доходы, а потому и сбор подоходного налога.
21. **Улучшение финансового положения бюджетной сферы** увеличивает уровень жизни работников бюджетной сферы, его денежные доходы, а потому и сбор подоходного налога.
22. **Улучшение внешнеэкономической ситуации** - повышение цен на отечественные товары и услуги, расширение иностранных инвестиций в отечественные предприятия улучшают экономическую ситуацию в России, повышает уровень жизни населения и увеличивают сбор подоходного налога.
23. **Рост курса доллара США** приводит к инфляции и уменьшает жизненный уровень населения, но увеличивает долю рынка, принадлежащую отечественным товаропроизводителям, и в итоге улучшает их финансовое положение.
24. **Повышение таможенных сборов на импортную продукцию** увеличивает объем доходов государства, а потому улучшает финансовое положение бюджетной сферы и части населения, увеличивает долю рынка, принадлежащую отечественным товаропроизводителям, и в итоге улучшает их финансовое положение, но одновременно приводит к инфляции и уменьшает жизненный уровень населения
25. **Повышение таможенных сборов на экспортную продукцию** увеличивает объем доходов государства, а потому улучшает финансовое положение бюджетной сферы и части населения, но одновременно ухудшает финансовое положения отечественных товаропроизводителей.

Следующий шаг – составление схем взаимодействия факторов в рассматриваемых моделях. На основе опроса экспертов получены схемы, представленные на рис.1 и рис.2.

**Рис.1. Модель НФЛ-18**

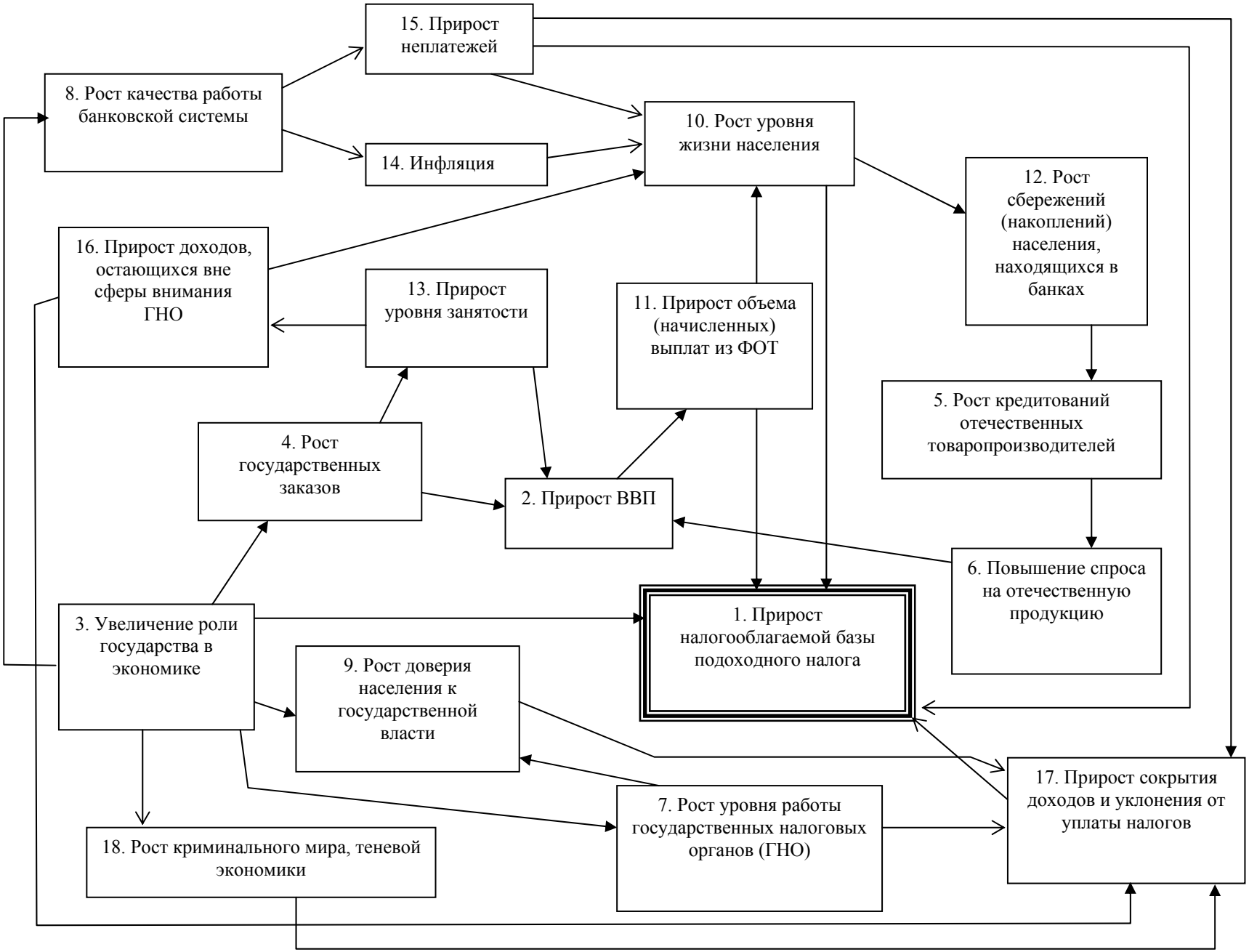
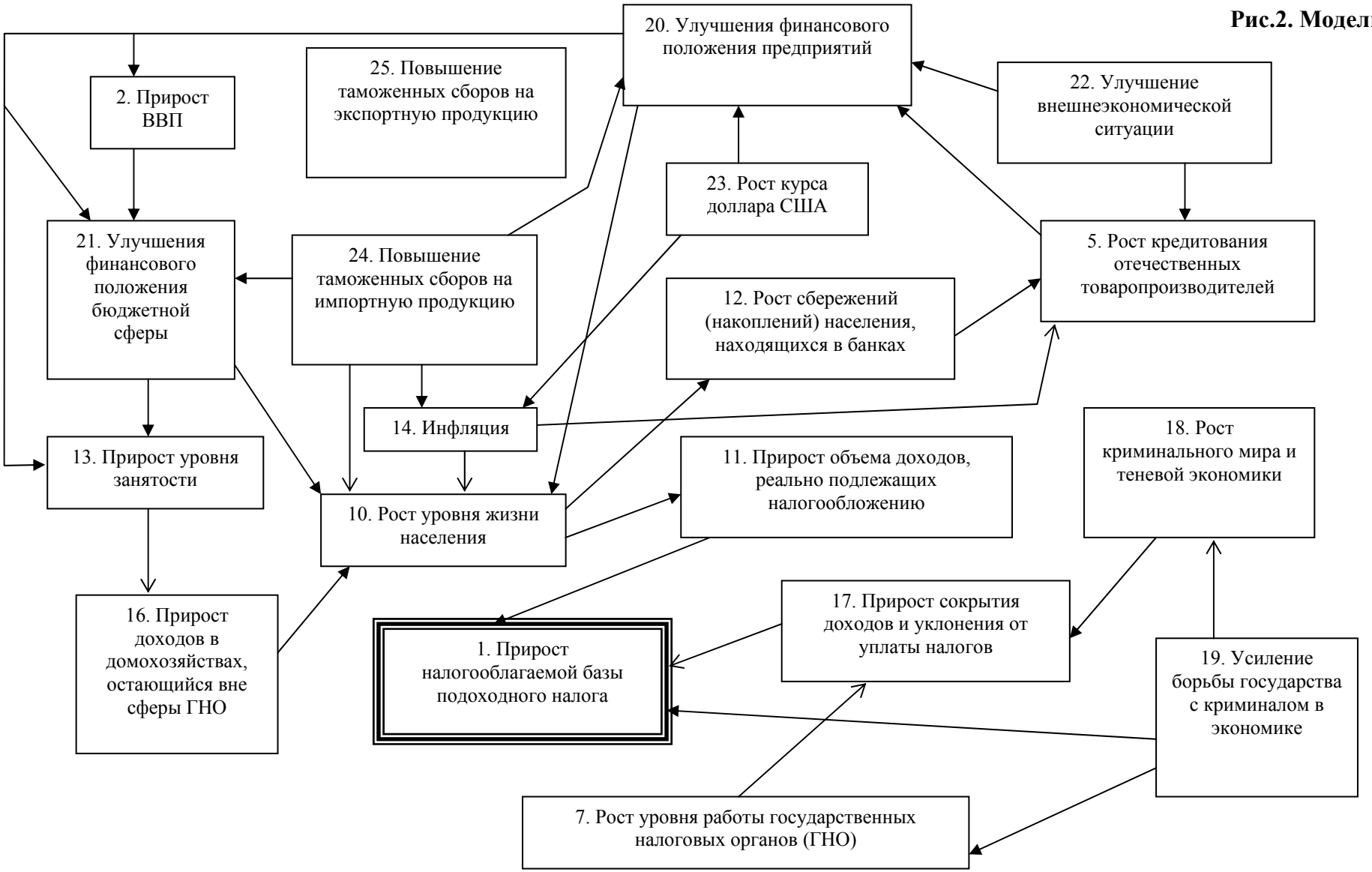


**Рис.2. Модель НФЛ-19**



**Матрицы коэффициентов взаимного влияния факторов в моделях НФЛ-18 и НФЛ-19.** Степень слияния факторов друг на друга можно оценить с помощью элементов матрицы влияния. Перечислим эти элементы в соответствии со схемой влияния факторов. Слева указан влияющий фактор, справа - фактор, на который оказывается влияние. Конкретные значения получены с помощью экспертов.

Начнем с модели НФЛ-18.

2-11 Прирост ВВП оказывает значительное положительное влияние на прирост объема (начисленных) выплат из ФОТ, оцениваемое величиной 0.7.

3-1 Увеличение роли государства в экономике умеренно увеличивает налогооблагаемую базу подоходного налога, в частности, с помощью законодательных мер (установлением льгот и др.), коэффициент влияния принимаем равным 0.3.

3-4 Увеличение роли государства в экономике в значительной мере проявляется в увеличении государственных заказов, влияние оцениваем числом 0,7.

3-7 Увеличение роли государства в экономике очень сильно проявляется в росте уровня работы государственных налоговых органов (ГНС), степень связи оцениваем как 0,9.

3-8 Увеличение роли государства в экономике означает и значительное повышение качества работы банковской системы, что обеспечивается работой законодательных и исполнительных государственных органов, прежде всего Центрального банка и МВД. Степень влияния оценивается числом 0.7.

3-9 Увеличение роли государства в экономике приводит к существенному возрастанию доверия населения к государственной власти, а потому и готовности платить налоги. Возрастает поддержка государственной власти народом, сплочение вокруг нее. Коэффициент влияния принимаем равным 0,6.

3-18 Увеличение роли государства в экономике приводит к значительному уменьшению криминального мира и теневой экономики, как за счет непосредственной борьбы государства с ними, так и за счет "поворота" народных масс от "мафии" к государству. Степень влияния оценивается числом (-0.7).

4-2 Рост государственных заказов приводит к значительному приросту ВВП, и степень влияния оценивается как 0,7.

4-13 Рост государственных заказов ведет к существенному росту уровня занятости. Коэффициент влияния равен 0,5.

5-6 Рост кредитования отечественных товаропроизводителей ведет к значительному увеличению выпуска конкурентоспособной и пользующейся спросом отечественной продукции. Степень влияния оценивается числом 0.7.

6-2 Повышение спроса на отечественную продукцию ведет к очень сильному росту ВВП с коэффициентом влияния 0,9.

7-9 Рост качества работы государственных налоговых органов (ГНО) вызывает существенный рост доверия населения к государству и готовности платить налоги. Коэффициент влияния принимаем равным 0,5.

7-17 Рост качества работы государственных налоговых органов (ГНО) значительно уменьшает сокрытие доходов и уклонение от уплаты налогов. Степень влияния оцениваем числом (-0.7).

8-14 Рост качества работы банковской системы с помощью различных финансовых инструментов позволяет существенно сократить инфляцию. Коэффициент влияния принимаем равным (-0,5).

8-15 Рост качества работы банковской системы позволяет существенно сократить неплатежи. В частности, с помощью системы взаимозачетов

и за счет ускорения и целевого использования выделенных федеральным центром средств. Коэффициент влияния принимаем равным (-0,5).

9-17 Рост доверия к государственной власти и готовности платить налоги ведет к существенному сокращению сокрытия доходов и уклонения от уплаты налогов с коэффициентом влияния (-0,5).

10-1 Рост уровня жизни населения приводит к существенному расширению налогооблагаемой базы подоходного налога Как из-за общего повышения денежных доходов населения, подлежащих налогообложению, так и за счет сокращения льгот для малообеспеченных граждан, доля которых уменьшается при общем росте уровня жизни. Степень влияния оцениваем числом 0.6.

10-12 Рост уровня жизни населения при прочих равных условиях ведет к существенному увеличению сбережений (накоплений) граждан, находящихся в банках, с коэффициентом влияния 0.6.

11-1 Прирост объема (начисленных) выплат из ФОТ ведет к очень сильному приросту налогооблагаемой базы подоходного налога (однако однозначной связи нет из-за неплатежей и льгот) с коэффициентом влияния 0.9.

11-10 Прирост объема (начисленных) выплат из ФОТ к значительному росту уровня жизни населения (однако полного соответствия нет из-за неплатежей, доходов вне ГНО, трансфертов и льгот). Степень влияния оцениваем числом 0.7.

12-5 Рост сбережений (накоплений) населения, находящихся в банках, дает возможность существенного роста кредитования отечественных товаропроизводителей, с коэффициентом влияния 0.6.

13-2 Прирост уровня занятости, т.е. сокращение безработицы, ведет к умеренному возрастанию ВВП с коэффициентом влияния 0.4.

13-16 Прирост уровня занятости, т.е. сокращение безработицы, ведет к умеренному сокращению доходов, остающихся вне сферы влияния ГНС, в частности, доходов от личного натурального хозяйства и доходов от личных услуг между домохозяйствами. Степень влияния оцениваем числом (-0.4).

14-10 Инфляция значительно снижает уровень жизни населения, толкает его в сторону сокрытия доходов, уклонения от уплаты налогов, получения доходов способами, остающимися вне влияния ГНС. Степень влияния оцениваем числом (-0.7).

15-1 Прирост неплатежей существенно сокращает налогооблагаемую базу подоходного налога с коэффициентом влияния (-0.5).

15-10 Прирост неплатежей существенно снижает уровень жизни населения. Степень влияния оценивается числом (-0.6).

15-17 Прирост неплатежей приводит к умеренному росту сокрытия доходов и уклонения от уплаты налогов с коэффициентом влияния 0.4.

16-10. Прирост доходов, остающихся вне сферы внимания ГНС, приводит к умеренному росту уровня жизни населения с коэффициентов влияния 0,3

16-17 Прирост доходов, остающихся вне сферы внимания ГНО, влечет значительный рост сокрытия доходов и уклонения от уплаты налогов с коэффициентом влияния 0.7.

17-1 Прирост сокрытия доходов и уклонения от уплаты налогов значительно снижает налогооблагаемую базу подоходного налога с коэффициентом влияния (-0,7).

18-17 Рост криминального мира и теневой экономики очень сильно влияет на прирост сокрытия доходов и уклонения от уплаты налогов с коэффициентом влияния 0.9.

Перейдем теперь к модели НФЛ-19.

2-21 Прирост ВВП влечет улучшение финансового положения государства, увеличение объема бюджета, а потому существенно улучшает финансовое положение бюджетной сферы с коэффициентом 0,6

5-20 Рост кредитования отечественных товаропроизводителей ведет к значительному увеличению выпуска конкурентоспособной и пользующейся спросом отечественной продукции, к. значительному улучшению финансового положения предприятий. Степень влияния оценивается числом 0.7.

7-17 Рост качества работы государственных налоговых органов (ГНО) значительно уменьшает сокрытие доходов и уклонение от уплаты налогов. Степень влияния оцениваем числом (-0.7).

10-12 Рост уровня жизни населения при прочих равных условиях ведет к существенному увеличению сбережений граждан, находящихся в банках, с коэффициентом влияния 0.6.

11-1 Прирост объема (начисленных) выплат из ФОТ ведет к очень сильному приросту налогооблагаемой базы подоходного налога (однако однозначной связи нет из-за неплатежей и льгот) с коэффициентом влияния 0.9.

11-10 Прирост объема (начисленных) выплат из ФОТ к значительному росту уровня жизни населения (однако полного соответствия нет из-за неплатежей, доходов вне ГНС, трансфертов и льгот). Степень влияния оцениваем числом 0.7.

12-5 Рост сбережений (накоплений) населения, находящихся в банках, дает возможность существенного роста кредитования отечественных товаропроизводителей, с коэффициентом влияния 0.6.

13-16. Прирост уровня занятости, т.е. сокращение безработицы, ведет к умеренному сокращению доходов, остающихся вне сферы влияния ГНС, в частности, доходов от личного натурального хозяйства и доходов от личных услуг между домохозяйствами. Степень влияния оцениваем числом (-0.4).

14-5 Инфляция существенно снижает объем кредитования отечественных товаропроизводителей, поскольку заставляет банки сокращать сроки кредитования и направляет кредитные потоки в сторону торговых предприятий. Степень влияния оцениваем числом (-0.6).

14-10 Инфляция значительно снижает уровень жизни населении. Толкает его в сторону сокрытия доходов, уклонения от уплаты налогов, получения доходов способами, остающимися вне влияния ГНС. Степень влияния оцениваем числом (-0.7).

16-10 Прирост доходов, остающихся вне сферы внимания ГНО, приводит к значительному росту уровня жизни населения с коэффициентов влияния 0,7

17-1 Прирост сокрытия доходов и уклонения от уплаты налогов значительно снижает налогооблагаемую базу подоходного налога с коэффициентом влияния (-0,7).

18-17 Рост криминального мира и теневой экономики очень сильно влияет на прирост сокрытия доходов и уклонения от уплаты налогов с коэффициентом влияния 0.9

19-1 Усиление борьбы государства с криминалом в экономике, в том числе в виде законотворчества, умеренно увеличивает ВВП, переводя часть "теневой" экономики в доступную учету область и защищая официально признанную предпринимательскую деятельность (коэффициент 0,3).

19-7 Усиление борьбы государства с криминалом в экономике очень сильно проявляется в росте уровня работы государственных налоговых органов (ГНС), степень связи оцениваем как 0,9.

19-18 Усиление борьбы государства с криминалом в экономике приводит к значительному уменьшению криминального мира и теневой экономики,

как за счет непосредственной борьбы государства с ними, так и за счет "поворота" народных масс от "мафии" к государству. Степень влияния оценивается числом (-0,7).

20-2 Улучшение финансового положения предприятий приводит к значительному возрастанию ВВП в результате ликвидации неплатежей, роста капиталовложений (инвестиций), что ведет к росту производства товаров и услуг (коэффициент 0,7).

20-11 Улучшение финансового положения предприятий приводит к значительному возрастанию объема выплат из ФОТ (коэффициент 0,8).

20-13 Улучшение финансового положения предприятий приводит к существенному возрастанию занятости. Степень влияния оценивается числом 0,5.

20-21 Улучшение финансового положения предприятий приводит к умеренному улучшению финансового положения бюджетной сферы. В основном за счет увеличения сбора налогов и расширения заказов предприятий организациям бюджетной сферы (коэффициент 0,3).

21-11 Улучшение финансового положения бюджетной сферы приводит к значительному возрастанию объема выплат из ФОТ (коэффициент 0,8).

21-13 Улучшение финансового положения бюджетной сферы приводит к существенному возрастанию занятости. Степень влияния оценивается числом 0,5.

22-5 Улучшение внешнеэкономической ситуации дает возможность умеренного возрастания кредитования отечественных товаропроизводителей, как за счет выручки экспортеров, так и путем прямых иностранных инвестиций (коэффициент 0,3).

22-20 Улучшение внешнеэкономической ситуации приводит к умеренному улучшению финансового положения предприятий, в основном за счет продажи продукции на экспорт (коэффициент 0,4).

23-14 Рост курса доллара США приводит к существенному росту цен (инфляции). Степень влияния оценивается числом 0,5.

23-20 Рост курса доллара США существенно улучшает финансовое положение предприятий. Увеличивая их долю отечественного рынка и облегчая (путем снижения издержек) выход на зарубежный рынок. Степень влияния оценивается числом 0,5.

24-10 Повышение таможенных сборов на импортную продукцию приводит к существенной инфляции (коэффициент 0,5).

24-14. Повышение таможенных сборов на импортную продукцию приводит к повышению ее цены, а потому к существенному снижению уровня жизни населения (коэффициент 0,5).

24-20 Повышение таможенных сборов на импортную продукцию существенно улучшает финансовое положение предприятий, избавляя их от иностранных конкурентов внутри страны (коэффициент 0,5).

24-21 Повышение таможенных сборов на импортную продукцию значительно улучшает финансовое положение бюджетной сферы за счет сборов, поступающих в бюджет государства. Степень влияния оценивается как 0,8.

25-20 Повышение таможенных сборов на экспортную продукцию приводит к умеренному ухудшению финансового положения предприятий за счет фактического изъятия в бюджет части средств за проданную ими продукцию. Степень влияния оценивается числом (-0,3).

25-21 Повышение таможенных сборов на экспортную продукцию значительно улучшает финансовое положение бюджетной сферы за счет сборов, поступающих в бюджет государства. Степень влияния оценивается как 0,8.

**Основные результаты моделирования.** Были рассмотрены сценарии прогноза (Пасс-1), активного воздействия (Акт-1), оптимального управления (Цель-1 и Цель-2). Сценарий Цель-1

состоял в нахождении управления, обеспечивающего значительное увеличение налогооблагаемой базы (не менее чем до 0,7), а сценарий Цель-2 нацелен на решение двухкритериальной задачи – обеспечить существенный (не менее 0,5) прирост налогооблагаемой базы при умеренном (не менее 0,3) приросте уровня жизни.

Основные результаты сведены в сводные таблицы 1 и 2 для НФЛ-18 и НФЛ-19 соответственно, в которых для наблюдаемых (т.е. основных) факторов указаны начальные состояния и приведены итоговые значения для каждого из четырех сценариев. Аналогичные сведения даны для суммарной оценки экономического состояния.

Таблица 1.
Сводная таблица для НФЛ-18

| Наблюдаемые факторы | Начальное состояние | Пасс-1 | Акт-1 | Цель-1 | Цель-2 |
|---|---|---|---|---|---|
| Прирост налогооблагаемой базы подоходного налога | 0 | - 0.76 | 1.71 | 2.45 | 2.45 |
| Прирост ВВП | 0.1 | - 0.2 | 0.71 | 1.62 | 1.62 |
| Рост доверия населения к государственной власти | - 0.5 | - 0.34 | 0.58 | 0.21 | 0.21 |
| Рост уровня жизни населения | - 0.5 | - 0.5 | 0.12 | 0.75 | 0.75 |
| Прирост уровня занятости | - 0.3 | - 0.41 | 0.01 | 0.43 | 0.43 |
| Инфляция | 0.3 | 0.51 | 0.34 | 0.16 | 0.16 |
| Суммарная оценка | - 0.14 | - 0.33 | 0.48 | 0.67 | 0.67 |

Таблица 2.
Сводная таблица для НФЛ-19

| Наблюдаемые факторы | Начальное состояние | Пасс-1 | Акт-1 | Цель-1 | Цель-2 |
|---|---|---|---|---|---|
| Прирост налогооблагаемой базы подоходного налога | 0 | - 2.56 | - 1.4 | - 0.06 | - 2.43 |
| Прирост ВВП | 0.1 | - 0.91 | - 0.72 | -0.82 | - 0.82 |
| Рост уровня жизни населения | - 0.5 | - 1.93 | - 1.81 | - 1.52 | - 1.52 |
| Прирост уровня занятости | - 0.3 | - 1.6 | - 1.21 | - 0.69 | - 0.69 |
| Инфляция | 0.3 | 0.6 | 0.9 | 1.1 | 1.1 |
| Суммарная оценка | - 0.2 | - 0.92 | - 0.58 | - 0.27 | - 0.97 |

Рассмотрим сначала модель НФЛ-18, которая предполагает активное участие государственных органов в регулировании экономических факторов. Естественное (т.е. без вмешательства с помощью управляющих факторов) развитие ситуации описывается сценарием Пассивный-1. Ситуация ухудшается по всем факторам, кроме доверия населения к государственной власти. Налогооблагаемая база подоходного налога значительно убывает (от нулевого начального значения приходим к значению (-0.76)). Рост ВВП (начальное значение 0.1) меняется на спад, хотя и слабый (-0.2). Уровень жизни населения продолжает падать с той же скоростью (-0.5). Падение занятости усиливается (с -0.3 до -0,41). Инфляция растет (с 0,3 до 0.51). Хотя рост доверия населения к государственной власти несколько растет (с -0.5 до -0.34), но коэффициент остается отрицательным, так что более правильно сказать так, скорость нарастания отрицательного отношения населения к государственной власти несколько снизилась. Вполне естественно, что уменьшилась и изначально отрицательная суммарная оценка экономической

ситуации - с (-0.14) до (-0.33). Общий вывод таков: если ничего не делать, то от плохой исходной ситуации страна придет к гораздо худшей.

Необходимы активные действия. Возможность резко изменить ситуацию к лучшему демонстрирует сценарий "Активный-1". В результате предложенных специалистами управляющих воздействий удается существенно улучшить почти все экономические показатели. Налогооблагаемая база подоходного налога очень сильно растет(от нулевого начального значения приходим к значению 1.71)). Валовой внутренний продукт значительно возрастает (от начального значения 0.1 до 0.71). Падение доверия населения к государственной власти (коэффициент -0.5) сменяется заметным ростом (коэффициент 0.58, т.е. в целом коэффициент увеличился на 1.08). Падение уровня жизни населения (коэффициент -0.5) сменилось слабым ростом (коэффициент 0.12). Падение занятости (коэффициент -0.3) прекратилось (коэффициент 0.01). Единственный показатель, по которому ситуация несколько ухудшилась - это инфляция (рост с 0.3 до 0.34), однако это ухудшение весьма незначительно по сравнению с впечатляющим ростом по другим показателям. Вполне естественно, что резко выросла и изначально отрицательная суммарная оценка экономической ситуации - с (-0.14) до 0.48. Итак, сценарий "Активный-1" демонстрирует большие возможности улучшения экономической ситуации вообще и налоговой ситуации в частности с помощью целенаправленных управляющих воздействий государственных органов.

Если в сценарии "Активный-1" система управляющих воздействий формировалась специалистами, оптимальность этой системы оставалась под вопросом, то в сценариях "Цель-1" и "Цель-2" система управляющих воздействий строилась с помощью компьютерной оптимизации в соответствии с заданными значениями целевых факторов. Поэтому вполне естественно, что по целевым факторам в результате оптимизации удалось еще больше улучшить ситуацию, чем в сценарии "Активный-1". При задании "получить коэффициент не менее 0.7" (сценарий "Цель-1) или "не менее 0.5" (сценарий "Цель-2") для налогооблагаемой базы подоходного налога удалось получить значение 2.45, заметно большее, чем в сценарии "Активный-1", т.е. 1.71. При задании "получить коэффициент не менее 0.3" для уровня жизни населения (сценарий "Цель-2") удалось получить его значительный рост с коэффициентом 0.75(по сравнению с 0.12 в сценарии "Активный-1"). Если же сравнить итог с исходным коэффициентом (-0.5), то общий рост уровня жизни - очень сильный, на 1.25. Все другие наблюдаемые факторы, кроме одного, также выросли больше, чем в сценарии "Активный-1". Наблюдаем очень сильный рост валового внутреннего продукта - до 1.62 (впечатляюще по сравнению с начальным значением 0.1 и соответствующим сценарию "Активный-1" значением 0.71). Падение занятости (коэффициент -0.3) сменилось ее умеренным ростом (коэффициент 0.43). Инфляция упала вдвое (с 0.3 до 0.16). Единственный показатель, по которому результаты сценариев "Цель-1" и "Цель-2" уступают результатам сценария "Активный-1" - это рост доверия населения к государственной власти (значения коэффициента 0.21 и 0.58 соответственно при начальном значении (-0.5)). (Уместно по этому поводу вспомнить утверждение о том, что наилучшей властью является та, действия которой незаметны для населения, все совершается как бы само собой.) Вполне естественно, что резко выросла и изначально отрицательная суммарная оценка экономической ситуации - с (-0.14) до 0.67 ( при 0.48 в сценарии "Активный-1"). Примечательно, что оптимальные решения для сценариев "Цель-1" и "Цель-2" совпали. Следовательно, есть оптимальная траектория развития экономики. Поскольку при движении по ней с лихвой выполняются поставленные задания по целевым факторам, то компьютерная оптимизация дает одинаковые решения для двух сценариев.

Подведем итоги по модели НФЛ-18. Прогноз развития экономической ситуации неблагоприятен (сценарий "Пассивный-1"), поэтому необходимы управляющие воздействия. Они позволяют существенно улучшить ситуацию (сценарий "Активный-1"), особенно при оптимизации воздействий (сценарии "Цель-1" и "Цель-2"). Модель НФЛ-18, повторим, демонстрирует большие возможности улучшения экономической ситуации с помощью целенаправленных управляющих воздействий государственных органов.

Перейдем к рассмотрению модели НФЛ-19, основанной на использовании прежде всего экономических взаимодействий. Прогноз экономической ситуации здесь гораздо более неблагоприятен, чем в предыдущей модели. Согласно сценарию "Пассивный-1" налогооблагаемая база подоходного налога очень сильно сокращается (от исходного значения 0 до (-2.56)), ВВП также очень сильно падает (от 0.1 до (-0.91)). Еще более сильно падают уровень жизни населения (от (-0.5) до (-1.93)) и уровень занятости ( от (-0.3) до (-1.6)). Вдвое растет инфляция (от 0.3 до 0.6).

Естественно, что суммарная оценка экономической ситуации также резко падает (от (-0.2) до (-0.92)).

Очевидно, необходимы активные действия, направленные на улучшение ситуации. В сценарии "Активный-1" управляющими воздействиями являются существенное усиление борьбы государства с криминалом в экономике (на 0.5), существенное повышение таможенных сборов на импортную продукцию (на 0.6) и слабое снижение таможенных сборов на экспортную продукцию (на 0.2). В результате удалось добиться некоторого замедления ухудшения ситуации по всем показателям, кроме инфляции. Налогооблагаемая база подоходного налога по-прежнему очень сильно сокращается (от исходного значения 0 до (-1.4), что лучше, чем при отсутствии воздействий (падение до (-2.56))). ВВП снова очень сильно падает (от 0.1 до (-0.72), что все-таки лучше, чем в сценарии "Пассивный-1", в котором падение достигло (-0.91)). Чуть медленнее падают уровень жизни населения (от (-0.5) до (-1.81), а не до (-1.93)) и уровень занятости (от (-0.3) до (-1.21), но не(-1.6)). Однако инфляция растет втрое, а не вдвое(от 0.3 до 0.9, а не до 0.6). Суммарная оценка экономической ситуации, равная (-0.58), показывает ее ухудшение по сравнению с исходным уровнем (-0.2), хотя и не такое резкое, как при отсутствии воздействий (-0.92).

Наилучшие результаты в модели НФЛ-19 получены при использовании сценария "Цель-1". Хотя целевого значения (0.7) для налогооблагаемой базы подоходного налога достичь не удалось, оказалось возможным сохранить ее практически на прежнем уровне (коэффициент (-0.06)). Однако по сравнению с предыдущим сценарием несколько усилилось падение ВВП (до (-0.82) по сравнению с (-0.72)), в то время как несколько улучшилась ситуация с уровнем жизни населения (коэффициент (-0.52) вместо (-1.81)) и уровнем занятости (коэффициент (-0.69), а не (-1.21)). В то же время усилилась инфляция (коэффициент 1.1, а не 0.9). Суммарная оценка экономической ситуации, равная (-0.27), является самой лучшей среди всех сценариев, но при этом показывает ухудшение экономической ситуации по сравнению с исходным уровнем (-0.2).

Интересны результаты, полученные в сценарии "Цель-2". Ни по одному из целевых факторов (налогооблагаемая база подоходного налога и уровень жизни населения) не удалось достичь заданных значений (0.5 и 0.3 соответственно). Однако попытка "одним выстрелом убить двух зайцев" привела к экономическим результатам, которые являются наихудшими среди всех моделей. Налогооблагаемая база подоходного налога очень резко упала (до (-2.43)). Остальные наблюдаемые факторы получили те же стационарные значения, что и в сценарии "Цель-1" (такова же картина по иным факторам - некоторые из них совпадают в этих двух сценариях. некоторые различаются, что видно по материалам раздела 3). .Суммарная оценка экономической ситуации, равная (-0.97), является самой худшей среди всех сценариев, хуже даже, чем при отсутствии каких-либо воздействий ( коэффициент (-0.92) в сценарии "Пассивный-1").

Подведем итоги по модели НФЛ-19. Оказалось, что чисто экономическими методами невозможно добиться поставленных целей, улучшить экономическую ситуацию. Максимум, чего можно достичь - это не позволить ей слишком сильно ухудшиться, удержать почти на исходном уровне налогооблагаемую базу подоходного налога и суммарную оценку экономической ситуации (как в сценарии "Цель-1").

Полученные с помощью моделей НФЛ-18 и НФЛ-19 результаты свидетельствуют о необходимости активного вмешательства государственных органов в экономические процессы, о невозможности улучшения ситуации чисто экономическими средствами. Этот вывод полностью соответствует концепции пяти нобелевских лауреатов по экономике (США) и отечественных академиков РАН о необходимости государственного регулирования экономики [16, 17].

Модели временных рядов НФЛ-18 и НФЛ-19, построенные на основе взвешенных ориентированных графов влияния факторов, допускают разнообразные варианты сценариев. За счет выбора тех или иных начальных значений факторов, наборов управляющих и целевых факторов. А также модификаций самих моделей путем модернизаций наборов факторов, коэффициентов их взаимовлияния, и др.

### 3.3.4. Моделирование и анализ многомерных временных рядов

Рассмотрим методы моделирования и анализа многомерных временных рядов, используемых для изучения реальных процессов взаимовлияния факторов на основе подхода ЖОК, описанного в предыдущем подразделе.

**Основные сведения о системе ЖОК.** Компьютерная система ЖОК – это система поддержки анализа и управления в сложных ситуациях[1], описываемых многомерными временными рядами. Она предназначена для структуризации и анализа сложных, трудно формализуемых, слабо структурированных задач различной природы (экономической, управленческой, прогностической, технической, медицинской, социально-политической, экологической и пр.). Она применяется для построения моделей ситуаций на основе описания влияний факторов. Это делается с помощью ориентированных графов и использования оценок экспертов с последующим определением наиболее эффективных управленческих решений. Компьютерная система ЖОК:

- поддерживает аналитическое обоснование подходов к решению исследуемых проблем;
- позволяет спрогнозировать развитие моделируемой реальной системы; оценить результаты целенаправленного изменения тех или иных факторов;
- дает возможность выработать условия для целенаправленного поведения в исследуемой ситуации;
- обеспечивает возможность решения прямых и обратных задач управления.

Для построения модели изучаемого явления или процесса компьютерная система ЖОК предусматривает выделение основных факторов, описывающих реальную ситуацию, и установление непосредственных взаимосвязей между факторами в виде построения ориентированного взвешенного графа. Опосредованные взаимовлияния и итоговое стационарное состояние рассчитываются по описанным ниже алгоритмам. Система позволяет анализировать три основных типа сценариев:

сценарий “Прогноз”, позволяющий проследить “естественное” развитие моделируемой системы при отсутствии активных воздействий;

сценарий типа “Активный”, при котором работающий с системой специалист изменяет значения тех или иных параметров и анализирует получающуюся динамику и итоговое состояние (например, с целью ручного поиска рационального управления);

сценарий типа “Цель”, когда компьютерная система по заданной цели управления (например, значения определенных параметров должны быть не менее заданных) находит оптимальные воздействия путем решения соответствующей задачи оптимизации. В частности, проводит анализ принципиальной достижимости указанной цели из текущего состояния с использованием выбранных мероприятий (управлений).

Ядром компьютерной системы ЖОК является описанная ниже математическая модель. Преобразование задач анализа реальных явлений и процессов к математической постановке, оценка адекватности реальности и ее модели, процесс выбора управлений, процесс сравнительного анализа различных ситуаций в целом, моделирования и последующей интерпретации результатов математического моделирования относится к области “ручного труда” специалиста в соответствующей области знания и полной автоматизации, как правило, не поддается.

Компьютерная система ЖОК обеспечивает расчет равновесного (стационарного) состояния, к которому будет стремиться система взаимовлияющих факторов, и всех промежуточных состояний на пути от начального состояния к равновесному. В систему включены три варианта расчетов:

- расчет равновесного состояния без управления (учитываются только начальные данные);
- расчет равновесного состояния с управлением импульсного типа (при $t = 0$). (В такой модели система интерпретирует импульсное управление, как поправку к начальным данным.);
- расчет величины управления по заданным значениям величины приращения целевых факторов.

**Математические алгоритмы исследовательской системы ЖОК.** Используются следующие обозначения:

$n$ - количество вершин в ориентированном графе $G$ модели, т.е. число используемых в модели факторов;

$\mathbf{D} = \lfloor d_{i,j} \rfloor_{n \times n}$ - матрица порядка $n$Чn непосредственных влияний факторов (матрица

[1] Использованы разработки В.Н.Жихарева, выполненные в Институте высоких статистических технологий и эконометрики.

смежности графа *G*);

$\mathbf{D}^T = \mathbf{A} = \left[a_{i,j}\right]_{n\times n}$ - матрица, транспонированная к матрице $\mathbf{D}$ (называемая матрицей непосредственных контрвлияний факторов);

*t* – время, принимающее дискретные значения 0, 1, 2, 3, …

вектор $\mathbf{V}(t) = (V_1(t), V_2(t), ..., V_n(t))^T$, *t* = 0, 1, 2, 3, …, - вектор изменений (приращений, дифференциалов) факторов в момент дискретного времени t;

вектор $\mathbf{W}(t) = \Delta\mathbf{V}(t) = \mathbf{V}(t) - \mathbf{V}(t-1)$, *t* = 0, 1, 2, 3, …, является вектором дифференциалов факторов второго порядка в момент дискретного времени *t*;

вектор $\mathbf{V}_{уст} = \mathbf{V}(\infty) = (V_1(\infty), V_2(\infty), ..., V_n(\infty))^T$ обозначает величины предельных стационарных изменений (дифференциалов) факторов при безграничном росте *t*. Очевидно, что если $\mathbf{V}(\infty)$ существует, то

$$\mathbf{W}(\infty) = \Delta\mathbf{V}(\infty) = \lim_{t\to\infty}(\mathbf{V}(t) - \mathbf{V}(t-1)) = 0);$$

вектор $\mathbf{g}(t) = (g_1(t), g_2(t), ..., g_n(t))^T$ обозначает внешние управляющие воздействия, подаваемые на фактор $V_i$.в момент *t;*

вектор $\mathbf{s} = (s_1, s_2, ..., s_n)^T$ обозначает сравнительную важность факторов $V_i$, задаваемую экспертным путем;

вектор $u = (u_1, u_2, ..., u_n)^T$ обозначает отношение составителя модели к направлению изменения величин факторов $V_i$ (+1 – рост значения фактора оценивается положительно, (-1) – отрицательно, 0 – нейтрально);

$\mathbf{E}$ - единичная *n*×*n* матрица (на главной диагонали стоят 1, на остальных позициях – 0);

$\mathbf{C}$ - прореженная единичная *n*×n матрица, в которой единицы стоят на диагонали только на тех позициях, которые соответствуют целевым факторам. Очевидно, что $\mathbf{C}$ является проектором на координатную плоскость целевых факторов, и следовательно $\mathbf{C}^2 = \mathbf{C}$, матрица $\mathbf{C}$ является псевдообратной к матрице $\mathbf{C}$;

$\mathbf{B}$ - прореженная единичная *n*×n матрица, в которой единицы стоят на диагонали только на тех позициях, которые соответствуют управляющим факторам. Очевидно, что $\mathbf{B}$ является проектором на координатную плоскость управляющих факторов, и, следовательно $\mathbf{B}^2 = \mathbf{B}$, матрица $\mathbf{B}$ является псевдообратной к матрице $\mathbf{B}$;

$\mathbf{Q} = (\mathbf{E} - k_{cm}\mathbf{A})^{-1}$ - резольвента, где $k_{cm}$ - множитель-стабилизатор, который используется в целях обеспечения достаточно устойчивой и быстрой сходимости итерационного процесса приближенного вычисления матрицы резольвентного оператора

$$\mathbf{Q} = (\mathbf{E} - k_{cm}\mathbf{A})^{-1} = \sum_{m=0}^{\infty}(k_{cm}\mathbf{A})^m \approx \sum_{m=0}^{p}(k_{cm}\mathbf{A})^m,$$

где p достаточно велико. Полагают $k_{cm} = 1$ в том случае, если собственные числа матрицы $\mathbf{A}$ достаточно малы (обычно принимается, что $k_{cm}\mathbf{A}$ должна иметь собственные числа не только меньше единицы, но и меньше 0.9). Поскольку стабилизатор $k_{cm}$ имеет лишь внутриматематический смысл и не используется при построении модели и интерпретации результатов расчетов, то в дальнейшем его не будем упоминать, предполагая по умолчанию $k_{cm} = 1$.

**Система уравнений в математико-статистической модели.** Для описания динамики факторов в компьютерной системе ЖОК используется математико-статистическая модель в виде

системы линейных конечноразностных рекуррентных уравнений на трехточечном шаблоне {*t*-1, *t*, t+1} следующего вида:

$$V_i(t+1) = V_i(t) + \sum_{j=1}^{n} a_{i,j}\left(V_j(t) - V_j(t-1)\right) + g_i(t) = V_i(t) + \sum_{j=1}^{n} d_{j,i}\left(V_j(t) - V_j(t-1)\right) + g_i(t) \quad (1),$$

с начальными условиями

$$V_i(0) = V_i^0 \quad (2),$$

где $i = 1, 2, \ldots, n$, t = 0, 1, 2, ...

Для рекуррентного уравнения на трехточечном шаблоне необходимо задать начальные условия при $t = 0$ ($V_i(0) = V_i^0$) и $t = 1$ ($V_i(1) = V_i^1$). Следовательно, первым уравнением цепочки рекуррентных уравнений (1) будет уравнение при $t = 1$.

При $t = 1$ уравнение полагается определенным и имеет вид

$$V_i(2) = V_i(1) + \sum_{j=1}^{n} a_{i,j}\left(V_j(1) - V_j(0)\right) + g_i(1)$$

Для $t = 0$ уравнение определяется посредством соотношения

$$V_i(-1) = 0 \quad (3),$$

и тогда недостающие начальные данные $V_i(1) = V_i^1$ вычисляются из уравнения

$$V_i(1) = V_i(0) + \sum_{j=1}^{n} a_{i,j}\left(V_j(0) - V_j(-1)\right) + g_i(0) = V_i(0) + \sum_{j=1}^{n} a_{i,j} V_j(0) + g_i(0) \quad (4)$$

Заметим, что доопределение начальных данных $V_i(-1) = 0$ нулем - всего лишь один из способов. В частности, если положить $V_i(0) = V_i(1) = V_i^0$, то результаты вычислений будут другими.

Из уравнений (1) видно, что используемая модель предполагает, что за один шаг дискретного времени ($\Delta t = 1$) происходит распространение влияния факторов-аргументов только на непосредственно от них зависящие факторы-функции. Времени можно придать содержательный смысл, если за шаг принять реальный интервал времени, необходимый для осуществления непосредственного влияния одного фактора на другой. Этот интервал может быть оценен экспертно, В ряде случаев его можно принять равным кварталу.

Уравнение (1) - (2) в векторной форме имеет вид

$$\mathbf{V}(t+1) = \mathbf{V}(t) + \mathbf{A} \circ \left(\mathbf{V}(t) - \mathbf{V}(t-1)\right) + \mathbf{g}(t) \quad (5)$$

$$\mathbf{V}(-1) = 0, \mathbf{V}(0) = \mathbf{V}_0 \quad (6)$$

где t = 0, 1, 2, ... Решение задачи (5)-(6) определяются формулой

$$\mathbf{V}(t) = \mathbf{V}(0) + \left(\sum_{k=0}^{t} \mathbf{A}^k\right) \circ \mathbf{V}(0) + \sum_{k=0}^{t-1} \sum_{m=0}^{t-1-k} \mathbf{A}^m \circ \mathbf{B} \circ \mathbf{g}(k) \quad (7).$$

**Стационарное состояние и начальные условия.** Стационарное состояние $\mathbf{V}(\infty)$ вычисляется приближенно при $t \to \infty$. Для практических расчетов достаточно принять, что

$t \le \min(n,25)$.

Векторное уравнение (5) может быть представлено в виде уравнения для дифференциалов второго порядка:

$$\mathbf{W}(t+1) = \mathbf{A} \circ \mathbf{W}(t) + \mathbf{g}(t) \qquad (8)$$

$$\mathbf{W}(0) = \Delta\mathbf{V}(0) = \mathbf{V}(0) - \mathbf{V}(-1) = \mathbf{V}(0) \quad , \qquad (9)$$

где $t$ = 0, 1, 2, ... Решение уравнения (8) – (9) имеет вид

$$\mathbf{W}(t) = \left( \sum_{k=0}^{t} \mathbf{A}^k \right) \circ \mathbf{W}(0) + \sum_{k=0}^{t-1} \sum_{m=0}^{t-1-k} \mathbf{A}^m \circ \mathbf{B} \circ \mathbf{g}(k) \qquad (10).$$

Если просуммировать уравнения (8) при $t$ = 0, 1, 2, . . . , то получим (при условии сходимости)

$$\mathbf{V}(\infty) - \mathbf{V}(0) = \mathbf{A} \circ (\mathbf{V}(\infty) - \mathbf{V}(-1)) + (\mathbf{g}(0) + \mathbf{g}(1) + \mathbf{g}(2) + \ldots) \qquad (11),$$

откуда следует

$$\mathbf{V}(\infty) = (\mathbf{E} - \mathbf{A})^{-1} \circ (\mathbf{V}(0) + \mathbf{g}(0) + \mathbf{g}(1) + \mathbf{g}(2) + \ldots) \qquad (12)$$

Если же просуммировать уравнения (8) при $t$ = 1, 2, . . . , то получим (при условии сходимости)

$$\mathbf{V}(\infty) - \mathbf{V}(1) = \mathbf{A} \circ (\mathbf{V}(\infty) - \mathbf{V}(0)) + (\mathbf{g}(1) + \mathbf{g}(2) + \mathbf{g}(3) + \ldots), \qquad (13)$$

и соответственно

$$\mathbf{V}(\infty) = \mathbf{V}(0) + (\mathbf{E} - \mathbf{A})^{-1} \circ ((\mathbf{V}(1) - \mathbf{V}(0)) + \mathbf{g}(1) + \mathbf{g}(2) + \mathbf{g}(3) + \ldots), \qquad (14),$$

откуда видно, что при выборе начальных условий вида $V_i(0) = V_i(1) = V_i^0$ результат (14) отличается от (12).

В частности, при выборе режима прогноза развития ситуации без управления $\mathbf{g}(1) = \mathbf{g}(2) = \mathbf{g}(3) = \ldots = 0$ и выборе начальных условий $V_i(0) = V_i(1) = V_i^0$, которые выражают равенство нулю вторых производных от величин факторов при $t$ = 0, из формулы (14) получим $\mathbf{V}(\infty) = \mathbf{V}(0)$. Это означает, что никакого развития ситуации не происходит. Она продолжает двигаться “равномерно и прямолинейно”, поскольку вторые дифференциалы факторов равны нулю и первые дифференциалы факторов не изменяются во времени.

С другой стороны, формула (12) предполагает, что начальные данные оказывают такое же ударное воздействие в момент t = 0, как и внешнее импульсное при $t$ = 0 управление, играющее роль (и имеющее “размерность”) “механической силы”.

Если предполагается использование только импульсных управляющих воздействий $\mathbf{g}(0) \ne 0$ при $t$ = 0 и в дальнейшем $\mathbf{g}(1) = \mathbf{g}(2) = \mathbf{g}(3) = \ldots = 0$, то задача развития ситуации без управления и с управлением не отличаются друг от друга, поскольку управление в сущности играет роль поправки к начальным данным и, обратно, начальные данные выполняют роль поправки к управлению.

**Режим поиска управления по целевым значениям факторов.** Проекция стационарного решения (12) уравнения (8) - (9) на координатную плоскость целевых факторов может быть представлено в виде

$$\mathbf{Y}_{уст} = \mathbf{Y}(\mathbf{V}(0)) + \mathbf{Y}(\mathbf{g}(0)),$$

где

$$\mathbf{Y}(\mathbf{V}(0)) = \mathbf{C} \circ \mathbf{Q} \circ \mathbf{V}(0), \; \mathbf{Y}(\mathbf{g}(0)) = \mathbf{C} \circ \mathbf{Q} \circ \mathbf{B} \circ \mathbf{g}(0),$$

или иначе

$$\mathbf{Y}_{уст} = \mathbf{C} \circ \mathbf{Q} \circ \mathbf{V}(0) + \mathbf{C} \circ \mathbf{Q} \circ \mathbf{B} \circ \mathbf{g}(0) = \mathbf{C} \circ \mathbf{Q} \circ (\mathbf{V}(0) + \mathbf{B} \circ \mathbf{g}(0)) \quad . \qquad (15).$$

Пусть $\mathbf{Y}^*_{уст}$ - вектор значений дифференциалов целевых факторов, тогда импульсное управление $\mathbf{g}(0)$ определяется по формуле

$$\mathbf{g}(0) = (\mathbf{C} \circ \mathbf{Q} \circ \mathbf{B})^+ \circ (\mathbf{Y}^*_{уст} - \mathbf{C} \circ \mathbf{Q} \circ \mathbf{V}(0)), \quad (16),$$

где "+" обозначает операцию псевдоинверсии, и матрица $(\mathbf{C} \circ \mathbf{Q} \circ \mathbf{B})^+$ является псевдообратной к матрице $\mathbf{C} \circ \mathbf{Q} \circ \mathbf{B}$;

$\mathbf{g}^*(0) = L_1(\mathbf{g}(0))$ является результатом применения к вектору $\mathbf{g}(0)$ операции $L_1$ - ограничения числовых значений компонент вектора $\mathbf{g}(0)$ величинами +1 и -1 , если эти значения выходят за пределы отрезка [-1; +1];

$\mathbf{g}^{**}(0) = Extr_1(\mathbf{g}^*(0))$ получается из $\mathbf{g}^*(0)$ применением операции $Extr_1$ - замены числовых значений $\mathbf{g}^*(0)$ ближайшими к ним экстремальными на отрезке [-1; +1] величинами +1 или -1 соответственно.

Тогда стационарные решения, получаемые с использованием этих управлений, вычисляются по формулам

$$\mathbf{Y}^{**}_{уст} = \mathbf{C} \circ \mathbf{Q} \circ \mathbf{V}(0) + \mathbf{C} \circ \mathbf{Q} \circ \mathbf{B} \circ \mathbf{g}^*(0),$$
$$\mathbf{Y}^{***}_{уст} = \mathbf{C} \circ \mathbf{Q} \circ \mathbf{V}(0) + \mathbf{C} \circ \mathbf{Q} \circ \mathbf{B} \circ \mathbf{g}^{**}(0).$$

**Степени матрицы смежности графа G и опосредованные взаимовлияния факторов.** Пусть вершина *x*1 влияет на вершину *x*2 с силой 0.5, вершина *x*2 влияет на *x*4 с силой 0.6, вершина *x*1 влияет на *x*3 с силой 0.8, вершина *x*3 влияет на *x*4 с силой 0.4. Тогда опосредованное суммарное влияние *x*1 на *x*4 имеет силу

0.5Ч0.6 + 0.8Ч0.4 = 0.62,

что равно сумме весов двух путей *x*1 → *x*2 → *x*4 и *x*1 → *x*3 → x4 из *x*1 в *x*4, веса которых равны соответственно 0.5Ч0.6 = 0.3 и 0.8Ч0.4 = 0.32. Суммарная сила влияния одного фактора на другой равна сумме весов всех маршрутов в ориентированном графе *G*, ведущих из одного фактора в другой. Вес пути (маршрута) определяется как произведение весов дуг составляющих этот путь (маршрут).

Если рассмотреть степени матрицы $\mathbf{D} = \lfloor d_{i,j} \rfloor_{n \times n}$, то их элементам можно придать вполне определенный смысл. Так, например, элемент матрицы $\mathbf{D}^2$ с координатами (1,2) равен сумме весов всех маршрутов из *x*1 в *x*2, содержащих ровно две дуги, а в $\mathbf{D}^3$ сумме весов всех маршрутов из *x*1 в *x*2, содержащих ровно три дуги и т.д. Таким образом, матрица $\sum_{m=0}^{\infty} \mathbf{D}^m$ выражает суммарные опосредованные влияния факторов друг на друга с учетом рефлексивного (при *m* = 0) непосредственного влияния фактора на самое себя с силой +1, а матрица $\sum_{m=1}^{\infty} \mathbf{D}^m$ не учитывает рефлексивного непосредственного влияния.

Матрица $\mathbf{Q} = (\mathbf{E} - \mathbf{A})^{-1} = \sum_{m=0}^{\infty} \mathbf{A}^m$ является матрицей контрвлияний факторов с учетом рефлексивности, а матрица

$$\mathbf{Q} - \mathbf{E} = -\mathbf{E} + \sum_{m=0}^{\infty} \mathbf{A}^m = \sum_{m=1}^{\infty} \mathbf{A}^m = \mathbf{A} \circ \sum_{m=0}^{\infty} \mathbf{A}^m = \mathbf{A} \circ \mathbf{Q} = \mathbf{A} \circ (\mathbf{E} - \mathbf{A})^{-1}$$ - матрицей контрвлияний факторов без учета рефлексивности.

Отдельный интерес представляет собой матрица $\operatorname{sign}\left(\sum_{m=0}^{\infty}\mathbf{D}^m\right)$ знаков элементов матрицы $\sum_{m=0}^{\infty}\mathbf{D}^m$, т.е. матрица направленности интегральных влияний фактора на фактор (или контрвлияний, если рассмотреть матрицу $\operatorname{sign}\left(\sum_{m=0}^{\infty}\mathbf{A}^m\right)$).

### 3.3.5. Балансовые соотношения в многомерных временных рядах

В настоящем подразделе анализируется соотношения между временными рядами основных макроэкономических характеристик на основе балансовых соотношений. За основу взята известная схема Макконнелла и Брю [18] с данными по США за 1988 г. Анализ проведен по методологии ЖОК. Он позволил вскрыть два балансовых нарушения в схеме Макконнелла и Брю для блоков “государство” и “бизнес” и предложить способ восстановления корректности схемы путем введения двух дополнительных блоков. Установлены отличия схемы Макконнелла и Брю от используемых в методологии ЖОК, в частности, наличие двух противоположно направленных финансовых потоков, соединяющих два блока. После устранения (суммирования) подобных потоков на основе количественных данных схемы Макконнелла и Брю построена модель «Расчет ВНП» в соответствии с методологией ЖОК.

Далее кратко рассмотрены возможные перспективные направления математико-компьютерного развития системы ЖОК. В частности, применение аппарата нечетких множеств, осуществление синтеза с количественными эконометрическими моделями, организация автоматизированного изучения устойчивости выводов по отношению к малым отклонениям начальных данных и матрицы взаимовлияний.

**Модели балансового типа в системе ЖОК.** Методология компьютерного моделирования взаимовлияния факторов ЖОК ориентирована на использование экспертных оценок непосредственных влияний факторов. Однако она может быть успешно применена и для построения и анализа моделей балансового типа, в которых непосредственные влияния факторов описываются количественным образом.

В качестве примера возьмем классическую систему макроэкономических характеристик (в соответствии с классической монографией К.Р. Макконнелла и С.Л. Брю [18, с.136-145]). Основная схема на с.144 этой книги не может быть непосредственно использована, поскольку в ней два блока могут быть связаны финансовыми потоками, идущими в противоположных направлениях (например, государство изымает из личных доходов индивидуальные налоги, но при этом направляет в личные доходы трансфертные платежи). Кроме того, часть информации на указанной схеме приведена не при описании макроэкономических характеристик, а при рассмотрении финансовых потоков.

Поэтому классическая схема была переделана в соответствии с методологией ЖОК. Результат представлен на схеме 3 ниже. При этом выявилась необходимость дополнить систему характеристик Макконнелла и Брю двумя новыми, обеспечивающими выполнение балансовых соотношений для блоков “государство” и “бизнес”. Все характеристики и блоки схемы 3 снабжены в качестве примеров численными значениями, соответствующими хозяйственной деятельности США за 1988 г. [18, с.136-145].

Однако в схеме 3 имеется ряд блоков, имеющих один вход и один выход. Естественно упростить схему, “убрав” такие блоки. В результате получена схема 4. Она уже не может удовлетворять балансовым соотношениям (сумма входов равна сумме выходов для каждого блока), поскольку при ее построении было проведено объединение всех финансовых потоков, непосредственно соединяющих те или иные два блока. Схема 4 дополнена относительными величинами, показывающими доли финансовых потоков, направленных из одного блока в другой. Их можно рассматривать как коэффициенты непосредственного влияния в исходной методологии

ЖОК. Однако поскольку они получены из количественных значений, приводим два знака после запятой, а не один, как в исходных моделях типа ЖОК.

**Основные макроэкономические характеристики.** В табл.3 указаны макроэкономические характеристики, используемые в [18] и семах 3 и 4. Содержание основных макроэкономических величин раскрывается ниже через соотношения между ними.

Таблица 3.
Основные макроэкономические характеристики.

| | |
|---|---|
| 1. | Государство (как хозяйствующий субъект) |
| 2. | Валовой национальный продукт |
| 3. | Чистый национальный продукт |
| 4. | Национальный доход |
| 5. | Чистый доход домохозяйств (после уплаты налогов) |
| 6. | Чистый экспорт (экспорт минус импорт) |
| 7. | Бизнес (как объединение хозяйствующих субъектов) |
| 8. | Другие источники финансирования государства (дополнительная характеристика по сравнению со схемой Макконнелла и Брю) |
| 9. | Накопления бизнеса (дополнительная характеристика по сравнению со схемой Макконнелла и Брю) |
| 10. | Государственные закупки товаров и услуг |
| 11. | Трансфертные платежи |
| 12. | Инвестиции бизнеса |
| 13. | Амортизационные отчисления |
| 14. | Нераспределенные доходы корпораций |
| 15. | Косвенные налоги на бизнес |
| 16. | Налоги на прибыли корпораций |
| 17. | Взносы на социальное страхование |
| 18. | Индивидуальные налоги |
| 19. | Личный доход: заработная плата, рента, процент, дивиденды |
| 20. | Личные сбережения |
| 21. | Личные потребительские расходы |

**Соотношения основных макроэкономических величин.** Раскроем содержание и приведем количественные данные для основных макроэкономических характеристик.

В 1988 г. валовой национальный продукт (ВНП) США составлял 4862 миллиарда долларов. В дальнейшем указания на единицу измерения (миллиард долларов) будем опускать. Основные хозяйствующие субъекты - это государство, бизнес и домохозяйства (семьи, конечные потребители).

Будем выписывать соотношения между макроэкономическими характеристиками, а строкой ниже - соответствующие балансовые соотношения (в миллиардах долларов).

Как известно,

ВНП = (чистый национальный продукт) + (амортизационные отчисления),

4862 = 4357 + 505.

Известно, что ВНП отражает не только результат работы страны за год – примерно на 10% он состоит из результатов труда предыдущих лет, перенесенных на продукты и услуги, выпущенные в данном году. Прошлый труд учитывается с помощью амортизационных отчислений.

По определению,

(чистый национальный продукт) = (национальный доход) +
+ (косвенные налоги на бизнес),

в количественном отношении -

4357 = 3964 + 393.

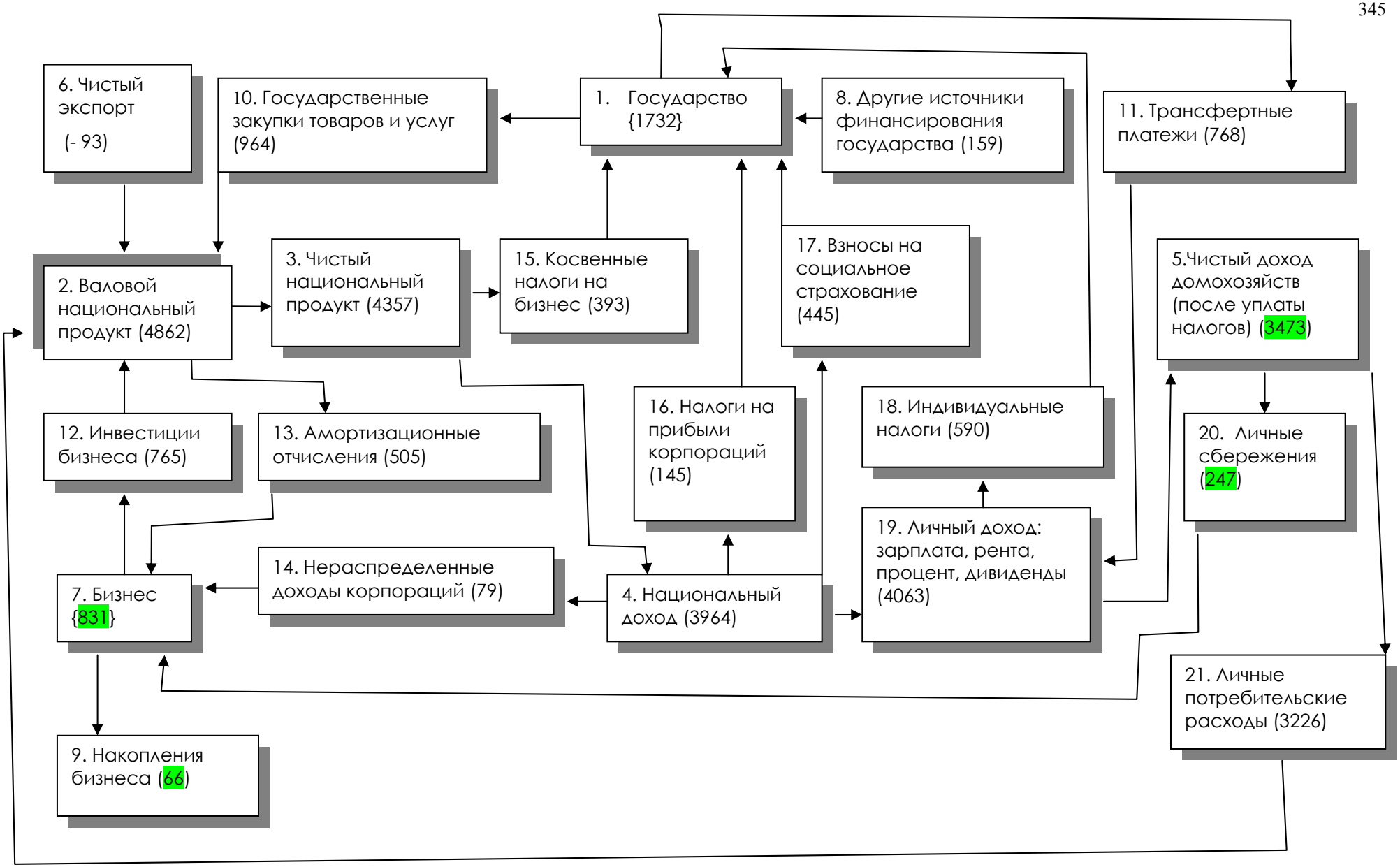



Схема 3. Соотношения основных макроэкономических величин (США, 1988 г., млрд. долл.)

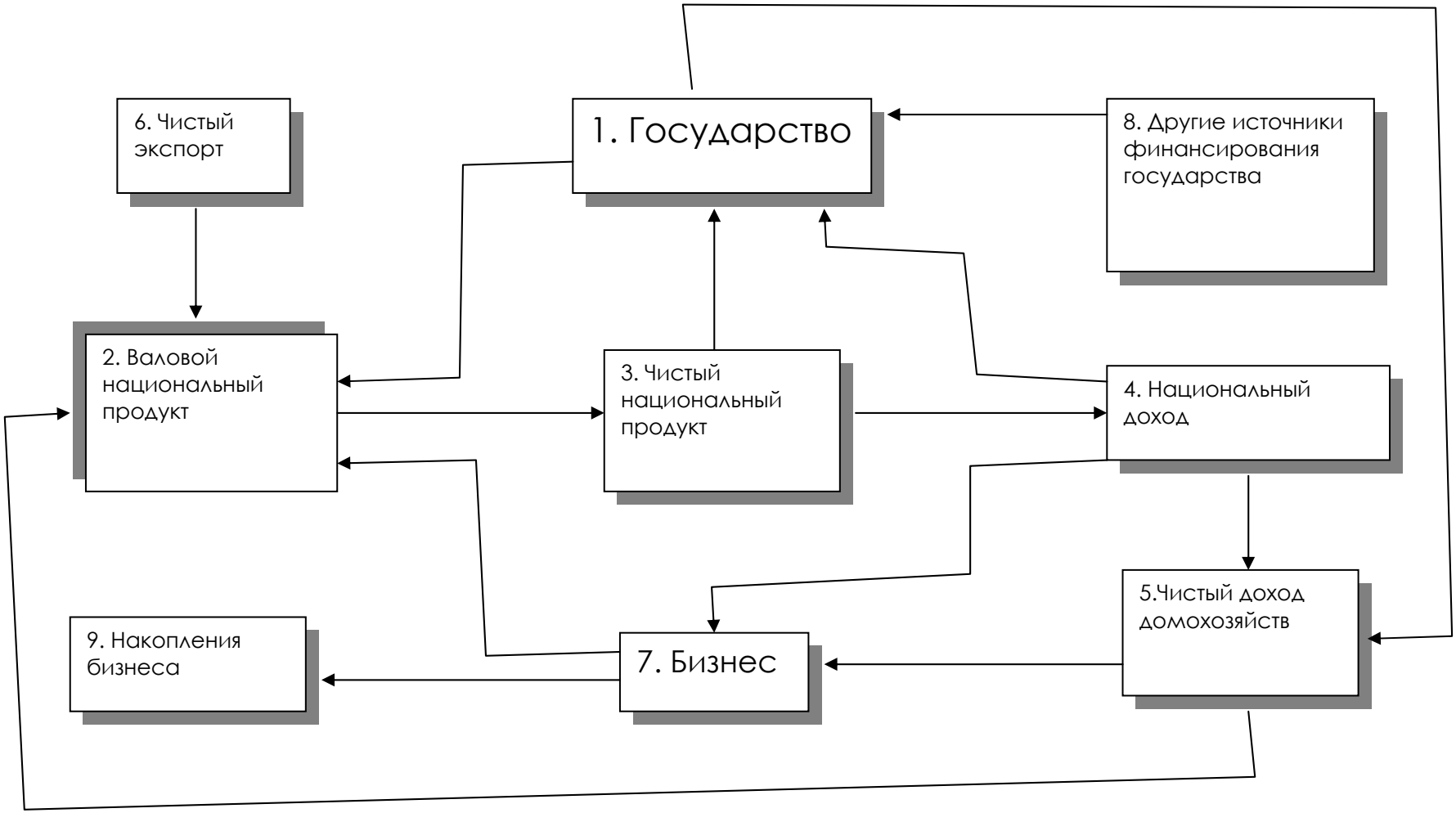



Схема 4. Взаимовлияния основных макроэкономических характеристик

Косвенные налоги на бизнес - это прежде всего акцизы на алкогольную и табачную продукцию, бензин, драгоценности, дорогие автомашины и т.п. Государство собирает косвенные налоги на бизнес, ничего не предоставляя взамен (бремя налогов делят между собой потребители и производители).

Поскольку составляющей (косвенные налоги на бизнес) не соответствует производство реальных товаров и услуг, вполне естественно ее исключить. *Наиболее естественной характеристикой результатов работы страны за год является национальный доход, составляющий лишь 81,5% от ВНП.*

По определению,

(национальный доход) = (личный доход (без трансфертов): зарплата, рента, процент, дивиденды) + (взносы на социальное страхование) + (налоги на прибыли корпораций) + (нераспределенные доходы корпораций),

3964 = 3295 + 445 + 145 + 79.

К личному доходу добавляются трансфертные платежи (пенсии, пособия и др.) со стороны государства в размере 768 миллиардов долларов. В рассматриваемом месте схемы наблюдаем круговое замыкание финансовых потоков - часть национального дохода идет на налоги, поступающие государству, а часть - в личный доход граждан, куда также поступают трансферты от государства:

(личный доход: зарплата, рента, процент, дивиденды) = (личный доход (без трансфертов): зарплата, рента, процент, дивиденды) + (трансфертные платежи),

4063 = 3295 + 768.

С точки зрения сбалансированности бюджета государства настораживает тот факт, что национальный доход (3064) оказывается меньше личного дохода (4063), поскольку часть национального дохода - это нераспределенные доходы корпораций (79), которые никак нельзя отнести к личному доходу.

*Примечание.* К трансфертным платежам относятся

(1) выплаты по страхованию по старости и от несчастных случаев, пособия по безработице, основанные на социальных программах;

(2) выплаты по вспомоществованию;

(3) разнообразные выплаты ветеранам, например, субсидии на образование и пособия по нетрудоспособности;

(4) выплаты частных пенсий и пособий по безработице и вспомоществованию;

(5) процентные платежи, выплачиваемые правительством и потребителями [18, с.143].

Ясно, что

(личный доход: зарплата, рента, процент, дивиденды) = (индивидуальные налоги) + (чистый доход домохозяйств),

4063 = 590 + 3437. 3473

Следовательно, средняя ставка индивидуальных налогов составляет 14,5%.

Очевидно,

(чистый доход домохозяйств) = (личные сбережения) + (личные потребительские расходы),

3473 3437 = 247 211 + 3226.

В кейнсианской макроэкономической теории большое значение имеет склонность к сбережению. Для США 1988 г. этот коэффициент равен 7,1 6,1%.

Основное соотношение для ВНП имеет вид:

ВНП = (личные потребительские расходы) + (государственные закупки товаров и услуг) + (инвестиции бизнеса) + (чистый экспорт),

4862 = 3226 + 964 + 765 - 93.

Заслуживает анализа хозяйственная деятельность государства (второе слагаемое в последней формуле для ВНП) и бизнеса (третье слагаемое в той же формуле).

**Хозяйственная деятельность государства и бизнеса.** Как видно из схемы 3, доходы государства складываются из следующих источников:

(доходы государства) = (косвенные налоги на бизнес) +

+ (налоги на прибыли корпораций) + (взносы на социальное страхование) + (индивидуальные налоги),

1573 = 393 + 145 + 445 + 590.

Согласно той же схеме, расходы государства таковы:

(расходы государства) = (государственные закупки товаров и услуг) + (трансфертные платежи),

1732= 964 + 768.

Следовательно, расходы государства превышают его доходы на 159 миллиардов долларов. Этот факт никак не разъясняется в [18].

Одно из возможных разъяснений может состоять в том, что отдельные разделы монографии Макконнелла и Брю не соответствуют друг другу, Так, проведенные выше рассуждения исходили из схемы на с.144, на которой финансовый поток (трансфертные платежи) ведет от блока (государство) к блоку (личный доход). Однако на с.143 той же книги подробно разъясняется понятие “трансфертные платежи” (это разъяснение процитировано выше), и из сказанного там ясно, что трансфертные платежи осуществляют не только государство.

Естественно выбрать другой путь снятия противоречия - ввести специальный блок (другие источники финансирования государства) с наполнением в 159 миллиардов долларов. К “другим источникам” могут относиться доходы от государственных предприятий, от внешнеэкономической деятельности государства, от эмиссии наличности и ценных бумаг и т.д.

Перейдем к рассмотрению блока (бизнес). Согласно [18]

(доходы бизнеса) = (амортизационные отчисления)

+ (нераспределенные доходы корпораций) + (личные сбережения),

831 795 = 505 + 79 + 211247,

в то время как

(расходы бизнеса) = (инвестиции бизнеса, включаемые в ВНП) = 765.

Итак, доходы бизнеса превышают его расходы на 66 30 миллиардов долларов. Этот факт никак не разъясняется в [18]. Одно из объяснений может состоять в некоторой условности отнесения личных сбережений к доходам бизнеса, с помощью которых он осуществляет инвестиции. Обычно считается, что личные сбережения хранятся в банках, а банки из этих средств выдают кредиты бизнесу. Однако очевидно, что некоторая часть поступивших в банк средств должна использоваться для обеспечения бесперебойной работы банка (с помощью государственной резервной системы, т.е. храниться в качестве резерва в центральном банке, а не выдаваться бизнесу в кредит). Кроме того, часть кредитов может быть не связана с инвестициями (хотя согласно [18] строительство частных домов учитывается в блоке (инвестиции бизнеса), но кредиты частным лицам не сводятся к кредитованию строительства).

Пойдем другим путем снятия противоречия - введем специальный блок (накопления бизнеса) с наполнением в 66 30 миллиардов долларов. Можно считать, что эти накопления выражены в денежной форме и в форме ценных бумаг, но не в виде материальных инвестиций, как в блоке (инвестиции бизнеса).

*Выводы* по результатам анализа данных, приведенных в книге Макконнелла и Брю [18], состоят в следующем.

1. Методология подхода, кратко именуемого ЖОК, позволяет адекватно анализировать схемы типа приведенной на с.144 известной монографии Макконнелла и Брю [18].

2. Методология ЖОК позволяет выявлять “дыры” (нарушения балансовых соотношений) в подобных схемах.

3. Дальнейшие усовершенствования схем типа рассмотренной приводят к достаточно сложным построениям т.н. “национального счетоводства” [21]. Не со всеми построениями “национального счетоводства” можно согласиться. Так, вряд ли можно научно обосновать утверждение о том, что банки создают такую большую долю ВНП, как 13%.

4. Схемы типа приведенной на с.144 известной монографии Макконнелла и Брю [18], основанные на количественных соотношениях, внешне представляются более обоснованными, чем схемы метода ЖОК. Однако при реальном использовании они не являются более полезными для решений стоящих в тематике ЖОК задач, чем схемы, основанные на экспертной оценке непосредственных влияний факторов. Связано это с достаточно большим произволом при формулировках определений численных значений.

5. На основе данных схемы, приведенной на с.144 известной монографии Макконнелла и Брю [18], может быть построена модель «Расчет ВНП» типа ЖОК (см. ниже).

**ВВП и ВНП.** ВНП отличается от валового внутреннего (отечественного) продукта (ВВП) на величину доходов от экономической деятельности, полученных из-за границы, за вычетом аналогичных доходов, переданных другим странам [22, с.55]. Другими словами, ВВП рассчитывается по всем предприятиям, действующим внутри рассматриваемой страны, независимо от доли иностранной собственности в тех или иных предприятиях. В отличие от ВВП, ВНП разделяет предприятия по стране принадлежности. Конкретно, для расчета ВНП надо к ВВП добавить доходы отечественных предприятий, полученные за рубежом, и вычесть доходы зарубежных структур, полученные на территории нашей страны. Понятны сложности, связанные с деятельностью транснациональных корпораций, с акционерными обществами (совместными предприятиями), объединяющими капитал из различных стран. Очевидно, расчет ВВП более прост и обоснован. Однако в учебниках по экономической теории укрепился рассказ о ВНП.

В странах, в которых подавляющая часть экономики принадлежит отечественным хозяйствующим субъектам, различие между ВНП и ВВП незначительно. Полагаем, что Россия относится к таким странам.

В российской экономической практике используется ВВП, а не ВНП. Например, согласно действующему законодательству финансирование Вооруженных Сил, образования, науки и др. установлено в процентах от ВВП.

**Модель ВНП.** Модель ВНП построена на основе методологии ЖОК с помощью приведенной выше схемы 3. Она содержит 9 факторов (табл.4), указанных в схеме 3 под номерами 1-9.

Таблица 4.
Факторы модели «Расчет ВНП».

| | |
|---|---|
| 1. | Государство (как хозяйствующий субъект) |
| 2. | Валовой национальный продукт |
| 3. | Чистый национальный продукт |
| 4. | Национальный доход |
| 5. | Чистый доход домохозяйств (после уплаты налогов) |
| 6. | Чистый экспорт (экспорт минус импорт) |
| 7. | Бизнес (как объединение хозяйствующих субъектов) |
| 8. | Другие источники финансирования государства |
| 9. | Накопления бизнеса |

Взаимовлияния основных макроэкономических характеристик приведены на схеме 4.

Степень влияния одного фактора на другой оценивалась по количественным данным, приведенным выше. Например, если из ВНП (4862 миллиарда долл.) в чистый национальный продукт переходит 4357 миллиардов долл., то влияние оценивается величиной 4357/4862 = 0,90. Другим примером является взаимоотношение блоков (государство) и (чистый доход домохозяйств). С одной стороны, государство берет 590 миллиардов индивидуальных налогов, с другой - выплачивает 768 миллиардов трансфертных платежей. Эти финансовые потоки объединяются, и итог - 178 миллиардов в пользу домохозяйств.

Модель ВНП содержит 14 связей. Они приведены в табл.5. В каждой паре указана сила воздействия первого элемента на второй.

Таблица 5.
Связи между факторами модели «Расчет ВНП»

| | |
|---|---|
| 3-1 | Величина чистого национального продукта влияет на государство (как хозяйствующий субъект) с коэффициентом влияния 0,09. |
| 4-1 | Национальный доход влияет на государство (как хозяйствующий субъект) с коэффициентом влияния 0,15. |
| 8-1 | Другие источники финансирования государства влияют на государство (как хозяйствующий субъект) с коэффициентом влияния |

| | |
|---|---|
| | 0,09. |
| 1-2 | Государство (как хозяйствующий субъект) влияет на валовой национальный продукт с коэффициентом влияния 0,56. |
| 5-2 | Чистый доход домохозяйств (после уплаты налогов) влияет на валовой национальный продукт с коэффициентом влияния 0,94. |
| 6-2 | Чистый экспорт (экспорт минус импорт) влияет на валовой национальный продукт с коэффициентом влияния 0,02. |
| 7-2 | Бизнес (как объединение хозяйствующих субъектов) влияет на валовой национальный продукт с коэффициентом влияния 0,33. |
| 2-3 | Валовой национальный продукт влияет на чистый национальный продукт с коэффициентом влияния 0,90. |
| 3-4 | Чистый национальный продукт влияет на национальный доход с коэффициентом влияния 0,91. |
| 1-5 | Государство (как хозяйствующий субъект) влияет на чистый доход домохозяйств (после уплаты налогов) с коэффициентом влияния 0,10. |
| 4-5 | Национальный доход влияет на чистый доход домохозяйств (после уплаты налогов) с коэффициентом влияния 0,87. |
| 4-7 | Национальный доход влияет на бизнес (как объединение хозяйствующих субъектов) с коэффициентом влияния 0,02. |
| 5-7 | Чистый доход домохозяйств (после уплаты налогов) влияет на бизнес (как объединение хозяйствующих субъектов) с коэффициентом влияния 0,06. |
| 7-9 | Бизнес (как объединение хозяйствующих субъектов) влияет на накопления бизнеса с коэффициентом влияния 0,04. |

**Перспективные направления математико-компьютерного развития системы ЖОК.** Технология ЖОК и ее программное обеспечение могут быть развиты в различных направлениях с целью расширения их познавательных и прикладных возможностей.

*Применение аппарата нечетких множеств.* В процессе распространения влияний факторов переход от одного фактора к другому может происходить многими путями. Возникает два вопроса:

- как рассчитать итоговое влияние при движении по фиксированному пути, если заданы коэффициенты непосредственного влияния между соседними факторами на этом пути?
- как свести вместе результаты влияний по различным путям?

В описанном выше варианте системы ЖОК в ответ на первый вопрос заданные коэффициенты непосредственного влияния между соседними факторами перемножают, а результаты влияний по различным путям складывают.

Представляется целесообразным использовать альтернативный вариант на основе концепций теории нечетких множеств. Например, итоговое влияние при движении по фиксированному пути рассчитывается как минимум заданных коэффициентов непосредственного влияния между соседними факторами на этом пути, а для получения итогового результата влияния находится максимум из результатов влияний по различным путям. Имеется ряд аргументов против используемой ныне процедуры и за предлагаемую процедуру.

*Синтез с количественными эконометрическими моделями* **.** Соединение основанной на не вполне количественных экспертных оценках технологии ЖОК с достаточно развитой технологией количественных эконометрических уравнений, в частности, использование лагов, т.е. не непосредственных влияний, а влияний с задержкой, представляет несомненный теоретический и практический интерес. Можно сравнить методологию ЖОК с эконометрической моделью, в которой исходные данные и правила перехода задаются экспертами в шкале порядка. При движении в обсуждаемом направлении необходимо дистанцироваться от попыток использовать нормальные распределения, поскольку подобных распределений в реальной экономике не может быть в принципе. Другие параметрические модели, в частности, логарифмически нормальные распределения, могут иметь ограниченную область полезности, но наиболее обоснованным является непараметрический подход.

*Автоматизированное изучение устойчивости выводов по отношению к малым отклонениям начальных данных и матрицы взаимовлияний.* Поскольку точность всех исходных экспертных оценок очевидным образом мала, необходимо изучить выводы на устойчивость. Это может быть сделано “малой кровью”, путем добавления в компьютерную систему ЖОК блока, который бы моделировал описанные выше малые отклонения и выдавал исследователи распределение (разброс) получаемых при этом выводов. С научной точки зрения этот подход - один из вариантов метода размножения выборок, кратко - бутстрепа [7].

Система ЖОК получила название по первым буквам основных разработчиков (В.Н.Жихарев, А.И.Орлов, В.Г Кольцов). Опыт практического применения этой системы описан в работах [23, 24]. Система ЖОК развивает идеи когнитивного подхода при решении слабоструктурированных задач, разработанного в Институте проблем управления РАН [25, 26], но на основе иного математического обеспечения.

Метод ЖОК может найти широкое применение для анализа экономического состояния и перспектив развития промышленных предприятий, банков, различных государственных и коммерческих структур. Представленные в настоящем подразделе материалы и результаты свидетельствуют о целесообразности дальнейшего развития рассматриваемой тематики с целью получения новых теоретических и практических результатов.

Подведем итоги главы. Рассмотрены методы анализа и моделирования временных рядов. Они используются прежде всего для прогнозирования технических, социально-экономических, медицинских и иных процессов. Надо отметить, что как самим временным рядам, так и вопросам их прогнозирования посвящена огромная литература. Наряду с вероятностно-статистическими методами при прогнозировании активно применяют экспертные методы [7].

В настоящей главе рассмотрены лишь основы и отдельные вопросы статистики временных рядов. Эта область - одна из наиболее обширных и сложных (с математической точки зрения) в прикладной статистике. Читатель, желающий глубже познакомиться с такой специфической ветвью прикладной статистики, как статистика временных рядов, должен обратиться к литературе, в частности, указанной в конце главы.

## Литература

1. Елисеева И.И., Юзбашев М.М. Общая теория статистики. - М.: Финансы и статистика., 1998. - 368 с.
2. Общая теория статистики. Статистическая методология в изучении коммерческой деятельности. / Под ред.А.А. Спирина, О.Э.Башиной. - М,: Финансы и статистика, 1994. - 296 с.
3. Доугерти К. Введение в эконометрику. - М.: МГУ, 1999. - 402 с.
4. Нейлор Т. Машинные имитационные эксперименты с моделями экономических систем. - М.: Мир, 1975. – 500 с.
5. Кендалл М.Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. - М.: Наука, 1976. – 736 с.
6. Кендэл М. Временные ряды. - М.: Финансы и статистика, 1981. – 199 с.
7. Орлов А.И. Эконометрика. Учебник для вузов. Изд. 2-е, исправленное и дополненное. - М.: Изд-во "Экзамен", 2003. – 576 с.
8. Петров В.М., Мажуль Л.А. Цикличность социокультурной сферы и проблемы среднесрочного прогнозирования ее развития. // Математическое и компьютерное моделирование в науках о человеке и обществе. Тезисы докладов Всероссийской конференции. - М.: Государственный ун-т управления, 1999. - С.63-66.
9. Николаев А.В. Структура исторического цикла. // Математическое и компьютерное моделирование в науках о человеке и обществе. Тезисы докладов Всероссийской конференции. - М.: Государственный ун-т управления, 1999. - С.54-54.
10. Носовский Г.В., Фоменко А.Т. Введение в новую хронологию. (Какой сейчас век?). - М.: КРАФТ+ЛЕАН, 1999. -768 с.
11. Каган А.М., Линник Ю.В., Рао С.Р. Характеризационные задачи математической статистики. - М.: Наука, 1972. – 656 с.
12. Биллингсли П. Сходимость вероятностных мер. - М.: Наука, 1977. – 352 с.
13. Крамер Г., Лидбеттер М. Стационарные случайные процессы. - М.: Мир, 1969.

14. Орлов А.И. Асимптотика решений экстремальных статистических задач // Анализ нечисловых данных в системных исследованиях. Сб. трудов. Вып.10. - М.: ВНИИСИ, 1982. - С.4-12.
15. Орлов А.И. Метод оценивания длины периода и периодической составляющей сигнала. - В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. - Пермь: Изд-во Пермского государственного университета, 1999. С.38-49.
16. Интрилигейтор М., Макинтайр Р., Тейлор Л., Эмсден А. Стратегия эффективного перехода и шоковые методы реформирования российской экономики. - В сб.: Шансы российской экономики / Под ред. Ю.М.Осипова, Е.С.Зотовой. - М.: Изд-во ТЕИС, 1997. - С.168-195.
17. Орлов А.А., Орлов А.И. Нобелевские лауреаты - за государственное регулирование экономики. - Журнал "Обозреватель - Observer", 1998, №1, с.44-46. Перепечатано в кн.: Современная политическая история России (1985-1998), т.1. Хроника и аналитика. - М.: "Духовное наследие"-РАУ-Корпорация, 1999. - С.909-911.
18. Макконнелл Кэмпбелл Р., Брю Стэнли Л. Экономикс: Принципы, проблемы и политика. В 2-х т.: Т.1. Пер. с англ. 11-го изд. - М.: Республика, 1995. - 400 с.
19. Налоги / Под ред. Д.Г.Черника. - М.: Финансы и статистика, 1998. - 688 с.
20. Социальная статистика / Под ред. чл.-корр. РАН И.И.Елисеевой. - М.: Финансы и статистика, 1997. - 416 с.
21. Национальное счетоводство / Под ред. Г.Д.Кулагиной. - М.: Финансы и статистика, 1997. - 448 с.
22. Статистический словарь. - М.: Финансы и статистика, 1989. - 623 с.
23. Жихарев В.Н., Кольцов В.Г., Орлов А.И. Эконометрический метод оценки результатов влияния / Тезисы конференции “Организация производства на предприятиях в современных условиях”. - М.: Изд-во МГТУ им. Н.Э.Баумана, 1999. - С.113-114.
24. Орлов А.И., Жихарев В.Н., Кольцов В.Г. Новый эконометрический метод "ЖОК" оценки результатов взаимовлияний факторов в инженерном менеджменте. - В сб.: Проблемы технологии, управления и экономики. / Под общей редакцией к. э. н. Панкова В.А. Ч.1. Краматорск: Донбасская государственная машиностроительная академия, 1999. - С.87-89.
25. Максимов В.И., Корноушенко Е.К. Аналитические основы применения когнитивного подхода при решении слабоструктурированных задач // Труды Института проблем управления РАН, 1998. №2.
26. Корноушенко Е.К., Максимов В.И. Управление процессами в слабоформализированных средах при стабилизации графовых моделей среды // Труды Института проблем управления РАН, 1998. №2.

## Контрольные вопросы

1. Чем эндогенные переменные отличаются от экзогенных переменных?
2. Какую роль играют показатели разброса и размаха при оценивании длины периода и периодической составляющей?
3. Сколько факторов используют при построении моделей многомерных временных рядов с помощью системы ЖОК?
4. Сколько взаимосвязей между факторами используют при построении моделей с помощью системы ЖОК?
5. В чем состоит один шаг учета взаимовлияний факторов в системе ЖОК?
6. Какая эконометрическая модель лежит в основе системы ЖОК?
7. На чем основана сходимость к соответствующим пределам оценок факторов в системе ЖОК?
8. Чем сценарии типа «Активный» отличаются от сценариев типа «Цель»?

## Темы докладов, рефератов, исследовательских работ

1. Методы сглаживания временных рядов.
2. Авторегрессионные модели.
3. Различные варианты метода наименьших квадратов при решении систем эконометрических уравнений.
4. Спектральная теория временных рядов.

5. Методы оценивания длины периода и периодической составляющей.
6. Различные виды экспертных оценок при построении моделей многомерных временных рядов с помощью системы ЖОК.
7. Роль оцифровки качественных оценок при использовании системы ЖОК.
8. Существование равновесных состояний при различных режимах использования системы ЖОК.
9. Система ЖОК как способ получения нового экономического знания (на примере двух типов моделей динамики налогооблагаемой базы подоходного налога НФЛ-18 и НФЛ-19).
10. Сравнение вариантов практического применения системы ЖОК в случае использования качественных и количественных признаков.

## 3.4. Статистика нечисловых данных

### 3.4.1. Структура статистики нечисловых данных

Статистика нечисловых данных или, как ее еще называют, статистика объектов нечисловой природы как самостоятельное научное направление была выделена в нашей стране. Как уже отмечалось, термин "статистика объектов нечисловой природы" впервые появился в 1979 г. в монографии [1]. В том же году в работе [2] была сформулирована программа развития этого нового направления прикладной статистики.

Со второй половины 80-х годов существенно возрос интерес к этой тематике и у зарубежных исследователей. Это нашло отражение, в частности, на Первом Всемирном Конгрессе Общества математической статистики и теории вероятностей им. Бернулли, состоявшемся в сентябре 1986 г. в Ташкенте. Статистика объектов нечисловой природы используется в нормативно-технической и методической документации, ее применение позволяет получить существенный технико-экономический эффект (сводка дана, например, в статье [3]).

Цель настоящей главы - дать введение в статистику нечисловых данных, выделить ее структуру, указать основные идеи и результаты.

Напомним, что объектами нечисловой природы называют элементы пространств, не являющихся линейными. Примерами являются бинарные отношения (ранжировки, разбиения, толерантности), множества, последовательности символов (тексты). Объекты нечисловой природы нельзя складывать и умножать на числа, не теряя при этом содержательного смысла. Этим они отличаются от издавна используемых в прикладной статистике (в качестве элементов выборок) чисел, векторов и функций.

Прикладную статистику по виду статистических данных принято делить на следующие направления:

статистика случайных величин (одномерная статистика);

многомерный статистический анализ;

статистика временных рядов и случайных процессов;

статистика нечисловых данных (ее важная часть – статистика интервальных данных).

При создании теории вероятностей и математической статистики исторически первыми были рассмотрены объекты нечисловой природы - белые и черные шары в урне. На основе соответствующих вероятностных моделей были введены биномиальное, гипергеометрическое и другие дискретные распределения. Получены теоремы Муавра-Лапласа, Пуассона и др. Современное развитие этой тематики привело, в частности, к созданию теории статистического контроля качества продукции по альтернативному признаку (годен - не годен) в работах А.Н.Колмогорова, Б.В. Гнеденко, Ю.К. Беляева, Я.П. Лумельского и многих других (см., например, классические монографии [4,5]).

В семидесятых годах XX в. в связи с запросами практики весьма усилился интерес к статистическому анализу нечисловых данных. Московская группа, организованная Ю.Н. Тюриным и другими специалистами вокруг созданного в 1973 г. научного семинара "Экспертные оценки и нечисловая статистика", развивала в основном вероятностную статистику нечисловых данных. Были установлены разнообразные связи между различными видами объектов нечисловой природы и изучены свойства этих объектов. Московской группой выпущены десятки сборников и обзоров, перечень которых приведен в итоговой работе [6]. Хотя в названиях многих из этих изданий стоят слова "экспертные оценки", анализ содержания сборников показывает, что подавляющая часть статей посвящена математико-статистическим вопросам, а не проблемам проведения экспертиз. Частое употребление указанных слов отражает лишь один из импульсов, стимулирующих развитие статистики объектов нечисловой природы и идущих от запросов практики. При этом необходимо подчеркнуть, что полученные результаты могут и должны активно использоваться в теории и практике экспертных оценок.

Новосибирская группа (Г.С. Лбов, Б.Г. Миркин и др.), как правило, не использовала вероятностные модели, т.е. вела исследования в рамках анализа данных. В московской группе в рамках анализа данных также велись работы, в частности, Б.Г.Литваком. Исследования по статистике объектов нечисловой природы выполнялись также в Ленинграде, Ереване, Киеве,

Таллинне, Тарту, Красноярске, Минске, Днепропетровске, Владивостоке, Калинине и других отечественных научных центрах.

**Внутреннее деление статистики объектов нечисловой природы.** Внутри рассматриваемого направления прикладной статистики выделяют следующие области.

1. Статистика конкретных видов объектов нечисловой природы.

2. Статистика в пространствах общей (произвольной) природы.

3. Применение идей, подходов и результатов статистики объектов нечисловой природы в классических областях прикладной статистики.

Единство рассматриваемому направлению придает прежде всего вторая составляющая, позволяющая с единой точки зрения подходить к статистическим задачам описания данных, оценивания, проверки гипотез при рассмотрении выборки, элементы которой имеют ту или иную конкретную природу. Внутри первой составляющей рассматривают:

1.1) теорию измерений;

1.2) статистику бинарных отношений;

1.3) теорию люсианов (бернуллиевских векторов);

1.4) статистику случайных множеств;

1.5) статистику нечетких множеств;

1.6) аксиоматическое введение метрик;

1.7) многомерное шкалирование и кластер-анализ (существенную часть этой тематики относят также к многомерному статистическому анализу).

Перечисленные разделы тесно связаны друг с другом, как продемонстрировано, в частности, в работах [1, 7] и предыдущих главах настоящего учебника. Вне данного перечня остались работы по хорошо развитым классическим областям - статистическому контролю, таблицам сопряженности, а также по анализу текстов и некоторые другие (см. [6, 8, 9]). Таким образом, кратко рассмотрим постановки 1970-2003 гг. вероятностной статистики объектов нечисловой природы, чтобы рассмотреть как единое целое это направление прикладной статистики.

**Статистика в пространствах общей природы.** Пусть $x_1,x_2,\ldots,x_n$ -элементы пространства $X$, не являющегося линейным. Как определить среднее значение для $x_1,x_2,\ldots,x_n$? Поскольку нельзя складывать элементы $X$, сравнивать их по величине, то необходимы подходы, принципиально новые по сравнению с классическими. В статистике объектов нечисловой природы предложено использовать показатель различия $d : X^2 \to [0,+\infty)$ (содержательный смысл показателя различия: чем больше $d(x,y)$, тем больше различаются $x$ и $y$) и определять эмпирическое среднее как решение экстремальной задачи

$$E_n(d) = Arg\min\{\sum_{1\le i\le n} d(x_i, x), x \in X\}. \qquad (1)$$

Таким образом, среднее $E_n(d)$- это совокупность всех тех $x \in X$, для которых функция

$$f_n x) = \frac{1}{n}\sum_{1\le i\le n} d(x_i, x) \qquad (2)$$

достигает минимума на $X$.

Как известно, для классического случая $X = R^1$ при $d(x,y) = (x-y)^2$ имеем $E_n(d) = \bar{x}$. При $X = R^1$, $d(x,y)=|x-y|$ среднее $E_n(d)$ при нечетном объеме выборки совпадает с выборочной медианой. А при четном объеме - $E_n(d)$ является отрезком с концами в двух средних элементах вариационного ряда.

Для ряда конкретных объектов среднее как решение экстремальной задачи вводилось рядом авторов. В 1929 г. итальянские статистики Джини и Гальвани применили такой подход для усреднения точек на плоскости и в пространстве Американский исследователь Джон Кемени решение задачи (1) называл медианой или средним для выборки, состоящей из ранжировок (см. монографию [10]). При моделировании лесных пожаров согласно выражению (1) было введено "среднеуклоняемое множество" для описания средней выгоревшей площади (см. об этом в монографии [1]). Общее определение эмпирических средних вида (1) было впервые введено в работе [2].

Основной результат, связанный со средними вида (1) - аналог закона больших чисел. Пусть $x_1,x_2,\ldots,x_n$ - независимые одинаково распределенные случайные элементы со значениями в

пространстве общей природы $X$. Теоретическим средним, или математическим ожиданием, в статистике объектов нечисловой природы называют

$$E_n(x_1, d) = Arg\min\{Md(x_1, x), x \in X\}. \qquad (3)$$

Закон больших чисел состоит в сходимости $E_n(d)$ к $E_n(x_1,d)$ при $n \to \infty$. Поскольку и эмпирическое, и теоретическое средние - множества, то понятие сходимости требует уточнения.

Одно из возможных уточнений, впервые введенное в работе [2], таково. Для функции

$$f(x) = Md(x_1(\omega), x), f : X \to R^1 \qquad (4)$$

введем понятие "$\varepsilon$-пятки" ($\varepsilon$>0)

$$K_\varepsilon(f) = \{x \in X : f(x) < \inf\{f(y), y \in X\} + \varepsilon\}. \qquad (5)$$

Очевидно, $\varepsilon$ - пятка $f$ - это окрестность Argmin($f$) (если он достигается), заданная в терминах минимизируемой функции. Тем самым снимается вопрос о выборе метрики в пространстве $X$. Тогда при некоторых условиях регулярности для любого $\varepsilon$>0 вероятность события

$$\{\omega : E_n(d) \subseteq K_\varepsilon(f)\} \qquad (6)$$

стремится к 1 при $n \to \infty$, т.е. справедлив закон больших чисел. Подробное доказательство приведено в главе 2.1.

Естественное обобщение рассматриваемой задачи позволяет построить общую теорию оптимизационного подхода в статистике. Как известно, большинство задач прикладной статистики может быть представлено в качестве оптимизационных. Как себя ведут решения экстремальных задач? Частные случаи этой постановки: как ведут себя при росте объема выборки оценки максимального правдоподобия и минимального контраста (в том числе робастные в смысле Тьюки - Хьюбера - см. главу 2.2)? Что можно сказать о поведении оценок нагрузок в факторном анализе и методе главных компонент при отсутствии нормальности, об оценках метода наименьших модулей в регрессии и т.д.?

Обычно легко устанавливается, что для некоторых пространств $X$ и последовательности случайных функций $f_n(x)$ при $n \to \infty$ найдется функция $f(x)$ такая, что

$$f_n(x) \to f(x) \qquad (7)$$

для любого $x \in X$ (сходимость по вероятности). Требуется вывести отсюда, что

$$Arg\min f_n(x) \to Arg\min f(x), \qquad (8)$$

т.е. решения экстремальных задач также сходятся. Понятие сходимости в соотношении (8) уточняется, например, с помощью $\varepsilon$-пяток, как это сделано выше для закона больших чисел. Условия регулярности, при которых справедливо предельное соотношение (8), приведены в исследовании [11]. Практически для всех реальных задач эти условия выполняются.

Как оценить распределение случайного элемента в пространстве общей природы? Поскольку понятие функции распределения неприменимо, естественно использовать непараметрические оценки плотности. Что такое плотность распределения вероятностей в пространстве произвольной природы? Это функция $g : X \to [0,+\infty)$ такая, что для любого измеримого множества (т.е. случайного события) $A \subseteq X$ справедливо соотношение

$$P(x_1(\omega) \in A) = \int_A g(x)\mu(dx), \qquad (9)$$

где. $\mu$ - некоторая мера в $X$. Ряд непараметрических оценок плотности был предложен в работе [2]. Например, ядерной оценкой плотности является оценка

$$g_n(x) = \frac{1}{\nu(h_n, x)} \sum_{1 \le i \le n} H\left(\frac{d(x_i, x)}{h_n}\right), \qquad (10)$$

где $d$ - показатель различия; $H$ - ядерная функция; $h_n$ - последовательность положительных чисел; $\nu(h_n, x)$ - нормирующий множитель. Удалось установить, что, что статистики типа (10) обладают такими же свойствами, по крайней мере при фиксированном $x$, что и их классические аналоги при $X = R^1$. В частности, такой же скоростью сходимости. Некоторые изменения необходимы при рассмотрении дискретных $X$, каковыми являются многие пространства конкретных объектов нечисловой природы (см. главу 2.1). С помощью непараметрических оценок плотности можно развивать регрессионный анализ, дискриминантный анализ и другие направления в пространствах общей природы.

Для проверки гипотез согласия, однородности, независимости в пространствах общей природы могут быть использованы статистики интегрального типа

$$\int f_n(x,\omega)dF_n(x,\omega), \qquad (11)$$

где $f_n(x,\omega)$-последовательность случайных функций на $X$; $F_n(x,\omega)$ - последовательность случайных распределений (или зарядов). Обычно $f_n(x,\omega)$ при $n \to \infty$ сходится по распределению к некоторой случайной функции $f(x,\omega)$, а $F_n(x,\omega)$ - к распределению $F(x)$. Тогда распределение статистики интегрального типа (11) сходится к распределению случайного элемента

$$\int f(x,\omega)dF(x). \qquad (12)$$

Условия, при которых это справедливо, даны в главе 2.3 на основе работы [12]. Пример применения - вывод предельного распределения статистики типа омега-квадрат для проверки симметрии распределения (см. главу 3.1).

Перейдем к статистике конкретных видов объектов нечисловой природы.

**Теория измерений.** Цель теории измерений - борьба с субъективизмом исследователя при приписывании численных значений реальным объектам. Так, расстояния можно измерять в верстах, аршинах, саженях, метрах, микронах, милях, парсеках и других единицах измерения. Выбор единиц измерения зависит от исследователя, т.е. субъективен. Статистические выводы могут быть адекватны реальности только тогда, когда они не зависят от того, какую именно единицу измерения предпочтет исследователь, т.е. когда они инвариантны относительно допустимого преобразования шкалы.

Теория измерений известна в нашей стране уже более 30 лет. С начала семидесятых годов активно работают отечественные исследователи. В настоящее время изложение основ теории измерений включают в справочные издания, помещают в научно-популярные журналы и книги для детей. Однако она еще не стала общеизвестной среди специалистов, в частности, среди метрологов. Поэтому опишем одну из задач теории измерений (ср. главу 3.1).

Как известно, шкала задается группой допустимых преобразований (прямой в себя). Номинальная шкала (шкала наименований) задается группой всех взаимно-однозначных преобразований, шкала порядка - группой всех строго возрастающих преобразований. Это - шкалы качественных признаков. Группа линейных возрастающих преобразований $\varphi(x) = ax + b, a > 0,$ задает шкалу интервалов. Группа $\varphi(x) = ax, a > 0,$ определяет шкалу отношений. Наконец, группа, состоящая из одного тождественного преобразования, описывает абсолютную шкалу. Это - шкалы количественных признаков. Используют и некоторые другие шкалы.

Практическую пользу теории измерений обычно демонстрируют на примере задачи сравнения средних значений для двух совокупностей одинакового объема $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_n$. Пусть среднее вычисляется с помощью функции $f : R^n \to R^1$. Если

$$f(x_1, x_2, \ldots, x_n) < f(y_1, y_2, \ldots, y_n), \qquad (13)$$

то необходимо, чтобы

$$f(\varphi(x_1), \varphi(x_2), \ldots, \varphi(x_n)) < f(\varphi(y_1), \varphi(y_2), \ldots, \varphi(y_n)) \qquad (14)$$

для любого допустимого преобразования $\varphi$ из задающей шкалу группы $\Phi$. (В противном случае результат сравнения будет зависеть от того, какое из эквивалентных представлений шкалы выбрал исследователь.)

Требование равносильности неравенств (13) и (14) вместе с некоторыми условиями регулярности приводят к тому, что в порядковой шкале в качестве средних можно использовать только члены вариационного ряда, в частности, медиану, но нельзя использовать среднее геометрическое, среднее арифметическое, и т.д. В количественных шкалах это требование выделяет из всех обобщенных средних по А.Н. Колмогорову в шкале интервалов - только среднее арифметическое, а в шкале отношений - только степенные средние. Кроме средних, аналогичные задачи рассмотрены в статистике нечисловых данных для расстояний, мер связи случайных признаков и других процедур анализа данных.

Приведенные результаты о средних величинах применялись, например, при проектировании системы датчиков в АСУ ТП доменных печей. Велико прикладное значение теории измерений в задачах стандартизации и управления качеством, в частности, в квалиметрии. Так, В.В. Подиновский показал, что любое изменение коэффициентов весомости единичных

показателей качества продукции приводит к изменению упорядочения изделий по средневзвешенному показателю, а Н.В. Хованов развил одну из возможных теорий шкал измерения качества. Теория измерений полезна и в других прикладных областях.

**Статистика бинарных отношений.** Оценивание центра распределения случайного бинарного отношения проводят обычно с помощью медианы Кемени. Состоятельность вытекает из закона больших чисел [1]. Разработаны различные вычислительные процедуры нахождения медианы Кемени.

Методы проверки гипотез развиты отдельно для каждой разновидности бинарных отношений. В области статистики ранжировок, или ранговой корреляции, классической является книга Кендалла [13]. Современные достижения отражены в работах Ю.Н.Тюрина и Д.С. Шмерлинга. Статистика случайных разбиений развита А.В.Маамяги. Статистика случайных толерантностей (рефлексивных симметричных отношений) впервые изложена в работе [1]. Многие ее задачи являются частными случаями задач теории люсианов.

**Теория люсианов (бернуллиевских векторов).** Люсиан (бернуллиевский вектор) - это последовательность испытаний Бернулли с, вообще говоря, различными вероятностями успеха. Реализация люсиана (бернуллиевского вектора) - это последовательность из 0 и 1. Люсианы (бернуллиевские вектора) рассматривались при статистическом анализе случайных множеств с независимыми элементами, а также результатов независимых парных сравнений. Последовательность результатов контроля качества последовательности единиц продукции по альтернативному признаку - также реализация люсиана (бернуллиевского вектора). Случайная толерантность может быть записана в виде люсиана. Поскольку один и тот же математический объект необходим в различных областях, естественно для его наименования применять специально введенный термин "бернуллиевский вектор". Используется также более краткий термин "люсиан".

В рассматриваемой теории изучают методы проверки согласованности (одинаковой распределенности), однородности двух выборок, независимости люсианов. Методы проверки указанных гипотез нацелены на ситуацию, когда число бернуллиевских векторов фиксировано, а их длина растет. При этом число неизвестных параметров возрастает пропорционально объему данных, т.е. теория построена в асимптотике растущего числа параметров. Ранее подобная асимптотика под названием асимптотики А.Н.Колмогорова использовалась в дискриминантном анализе, но там применялись совсем другие методы для решения иных задач прикладной статистики.

Непараметрическая теория парных сравнений (в предположении независимости результатов отдельных сравнений) - часть теории бернуллиевских векторов. Параметрическая теория связана в основном с попытками выразить вероятности того или иного исхода через значения гипотетических или реальных параметров сравниваемых объектов. Известны модели Терстоуна, Бредли-Терри-Льюса и др. В нашей стране построен ряд новых моделей парных сравнений. В частности, имеются модели парных сравнений с тремя исходами (больше, меньше, неразличимо), модели зависимых сравнений, сравнений нескольких объектов (сближающие рассматриваемую область с теорией случайных ранжировок) и т.д.

**Статистика случайных и нечетких множеств.** Давнюю историю имеет статистика случайных геометрических объектов (отрезков, треугольников, кругов и т.д.). Современная теория случайных множеств сложилась при изучении пористых сред и объектов сложной природы в таких областях, как металлография, петрография, биология. Различные направления внутри этой теории рассмотрены в работе [1, гл.4]. Остановимся на двух.

Случайные множества, лежащие в евклидовом пространстве, можно складывать: сумма множеств $A$ и $B$- - это объединение всех векторов $x+y$, где $x \in A, y \in B$. Н.Н. Ляшенко получил аналоги законов больших чисел, центральной предельной теоремы, ряда методов прикладной статистики, систематически используя подобные суммы.

Для статистики объектов нечисловой природы интереснее подмножества пространств, не являющихся линейными. В работе [1] рассмотрены некоторые задачи теории конечных случайных множеств. Позже ряд интересных результатов получил С.А. Ковязин, в частности, он доказал нашу гипотезу о справедливости закона больших чисел при использовании расстояния между множествами

$$d(a,b) = \mu(A \Delta B), \qquad (15)$$

где $\mu$ - некоторая мера; $\Delta$ - знак симметрической разности. Расстояние (15) выведено из некоторой системы аксиом в монографии [1]. Прикладники также делают попытки развивать и применять методы статистики случайных множеств.

С теорией случайных множеств тесно связана теория нечетких множеств, начало которой положено статьей Л.А.Заде 1965 г. Это направление прикладной математики получило бурное развитие - к настоящему времени число публикаций измеряется десятками тысяч, имеются международные журналы, постоянно проводятся конференции, практические приложения дали ощутимый технико-экономический эффект. При изложении теории нечетких множеств обычно не подчеркивается связь с вероятностными моделями. Между тем еще в первой половине 1970-х годов было установлено [1], что теория нечеткости в определенном смысле сводится к теории случайных множеств, хотя эта связь, возможно, имеет лишь теоретическое значение.

С точки зрения статистики нечисловых данных нечеткие множества - лишь один из видов объектов нечисловой природы. Поэтому к ним применима общая теория, развитая для пространств произвольной природы. Имеются работы, в которых совместно используются соображения вероятности и нечеткости.

**Многомерное шкалирование и аксиоматическое введение метрик.** Многомерное шкалирование имеет целью представление объектов точками в пространстве небольшой размерности (1 - 3) с максимально возможным сохранением расстояний между точками.

Из сказанного выше ясно, какое большое место занимают в статистике объектов нечисловой природы метрики (расстояния). Как их выбрать? Предлагают выводить вид метрик из некоторых систем аксиом. Аксиоматически получена метрика в пространстве ранжировок, которая оказалась линейно связанной с коэффициентом ранговой корреляции Кендалла. Метрика (15) в пространстве множеств получена в работе [3] также исходя из некоторой системы аксиом. Г.В. Раушенбахом [14] дана сводка по аксиоматическому подходу к введению метрик в пространствах нечисловой природы. К настоящему времени практически для каждой используемой в прикладных работах метрики удалось подобрать систему аксиом, из которой чисто математическими средствами можно вывести именно эту метрику.

**Применения статистики объектов нечисловой природы.** Идеи, подходы, результаты статистики объектов нечисловой природы оказались полезными и в классических областях прикладной статистики. Статистика в пространствах общей природы позволила с единых позиций рассмотреть всю прикладную статистику, в частности, показать, что регрессионный, дисперсионный и дискриминантный анализы являются частными случаями общей схемы регрессионного анализа в пространстве произвольной природы. Поскольку структура модели - объект нечисловой природы, то ее оценивание, в частности, оценивание степени полинома в регрессии, также относится к статистике нечисловых данных. Если учесть, что результаты измерения всегда имеют погрешность, т.е. являются не числами, а интервалами или нечеткими множествами, то приходим к необходимости пересмотреть некоторые выводы теоретической статистики. Например, отсутствует состоятельность оценок, нецелесообразно увеличивать объем выборок сверх некоторого предела (см. главу 3.5).

Технико-экономическая эффективность от применения методов статистики нечисловых данных достаточно высока. К сожалению, из-за изменения экономической ситуации, в частности, из-за инфляции трудно сопоставить конкретные экономические результаты в разные моменты времени. Кроме того, методы статистики объектов нечисловой природы составляют часть методов прикладной статистики. А те, в свою очередь - часть методов, входящих в систему информационной поддержки принятия решений на предприятии. Какую часть приращения прибыли предприятия надо отнести на эту систему? Мы знаем, как работает система управления фирмой в настоящем виде. Но можем только гадать (а точнее, оценивать, скорее всего, с помощью экспертных оценок), каковы были бы результаты финансово-хозяйственной деятельности предприятия, если бы система управления фирмой была бы иной, например, не содержала методов статистики объектов нечисловой природы.

Статистика объектов нечисловой природы как часть прикладной статистики продолжает бурно развиваться. В частности, увеличивается количество ее практически полезных применений при анализе конкретных экономических данных - в маркетинговых исследованиях, контроллинге, при управлении предприятием и др.

### 3.4.2. Теория случайных толерантностей

В прикладных исследованиях обычно используют три конкретных вида бинарных отношений – ранжировки, разбиения и толерантности. Статистические теории ранжировок [13] и разбиений [15] достаточно сложны с математичеbской точки зрения. Поэтому продвинуться удается не очень далеко. Теория случайных ранжировок, в частности, изучает в основном равномерные распределения на множестве ранжировок. Теория случайных толерантностей позволяет рассмотреть более общие ситуации. Это объясняется, грубо говоря, тем, что для теории толерантностей оказываются полезными суммы некоторых независимых случайных величин, а для теории ранжировок и разбиений аналогичные случайные величины зависимы. Теория случайных толерантностей является частным случаем теории люсианов, рассматриваемой в подразделе 3.4.3. Здесь приводим результаты, специфичные именно для толерантностей.

Пусть $X$ - конечное множество из $k$ элементов. Толерантность $A$ на множестве $X$, как и любое бинарное отношение, однозначно описывается матрицей $||a(i, j)||$, $1 \le i, j \le k$, где $a(i, j) = 1$, если элементы с номерами $i$ и $j$ связаны отношением толерантности, и $a(i, j) = 0$ в противном случае. Поскольку толерантность – это рефлексивное и симметричное бинарное отношение, то достаточно рассматривать часть матрицы, лежащую над главной диагональю: $||a(i, j), 1 \le i<j \le k||$. Между наборами $||a(i, j), 1 \le i<j \le k||$ из 0 и 1 и толерантностями на $X$ имеется взаимнооднозначное соответствие.

Пусть $A = A$(щ) – случайная толерантность, равномерно распределенная на множестве всех толерантностей на $X$. Легко видеть, что в этом случае $a(i, j)$, $1 \le i<j \le k$, - независимые случайные величины, принимающие значения 0 и 1 с вероятностями 0,5. Этот факт, несмотря на свою математическую тривиальность, является решающим для построения теории толерантностей. Для аналогичных постановок в теории ранжировок и разбиений величины $a(i, j)$ оказываются зависимыми.

Следовательно, случайная величина

$$B(A) = \sum_{i=1}^{k} \sum_{j=1}^{k} a(i, j)$$

имеет биномиальное распределение с параметрами $k(k\text{-}1)/2$, S и асимптотически нормальна при $k \to \infty$.

**Проверка гипотез о согласованности.** Рассмотрим $s$ независимых толерантностей $A_1$, $A_2$, …, $A_s$, равномерно распределенных на множестве всех толерантностей на $X$. Рассмотрим вектор

$$\xi_{ks} = \{d(A_p, A_q), 1 \le p < q \le s\} = \sum_{1 \le p<q \le k} \{| a_p(i, j) - a_q(i, j) |, 1 \le p < q \le s\}, \quad (1)$$

где $d(A_p, A_q)$ – расстояние между толерантностями $A_p$ и $A_q$, аксиоматически введенное в главе 1.1. В (1) предполагается, что пары $(p, q)$, $p < q$, располагаются в раз навсегда установленном порядке, для определенности в лексиграфическом (т.е. пары упорядочиваются в соответствии со значением $p$, а при одинаковых $p$ – по значению $q$).

Вектор $o_{ks}$ является суммой $k(k\text{-}1)/2$ независимых одинаково распределенных случайных векторов, а потому асимптотически нормален при $k \to \infty$. Координаты этого вектора независимы, поскольку, как нетрудно видеть, координаты каждого слагаемого независимы (это свойство не сохраняется при отклонении от равномерности распределения). Распределения случайных величин $a_p(i, j)$ и $|a_p(i, j) - a_q(i, j)|$ совпадают, поэтому распределения $B(A)$ и $d(A_p, A_q)$ также совпадают.

В силу многомерной центральной предельной теоремы (глава 1.4) распределение вектора

$$\eta_{ks} = \sqrt{\frac{2}{k(k-1)}} \left( \xi_{rs} - \frac{k(k-1)}{2} \left( \frac{1}{2}, \frac{1}{2}, ..., \frac{1}{2} \right) \right)$$

сходится при $k \to \infty$ к распределению многомерного нормального вектора $z_s$, ковариационная матрица которого совпадает с ковариационной матрицей вектора $z_{ks}$, а математическое ожидание равно 0. Таким образом, координаты случайного вектора $z_s$ независимы и имеют стандартное нормальное распределение с математическим ожиданием 0 и дисперсией 1. В соответствии с теоремами о наследовании сходимости (глава 1.4) распределение $f(z_{ks})$ сходится при $k \to \infty$ к распределению $f(z_s)$ для достаточно широкого класса функций $f$, в частности, для всех непрерывных функций. В качестве примеров рассмотрим статистики

$$W = \sum_{1 \le p < q \le s} d(A_p, A_q), \quad N = \sum_{1 \le p < q \le s} \left( d(A_p, A_q) - \frac{k(k-1)}{4} \right)^2 .$$

При $k \to \infty$ распределения случайных величин

$$\frac{8W - s(s-1)k(k-1)}{2\sqrt{s(s-1)k(k-1)}}, \quad \frac{8N}{k(k-1)}$$

сходятся соответственно к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1 и распределению хи-квадрат с $s(s-1)/2$ степенями свободы. Статистики $W$ и $N$ могут быть использованы для проверки гипотезы о равномерности распределения толерантностей.

Как известно, в теории ранговой корреляции [13], т.е. в теории случайных ранжировок, в качестве единой выборочной меры связи нескольких признаков используется коэффициент согласованности $W(R)$, называемый также коэффициентом конкордации [16, табл.6.10]. Его распределение затабулировано в предположении равномерности распределения на пространстве ранжировок (без связей). Непосредственным аналогом $W(R)$ в случае толерантностей является статистика $W$. Статистики $W$ и $N$ играют ту же роль для толерантностей, что $W(R)$ для ранжировок, однако математико-статистическая теория в случае толерантностей гораздо проще, чем для ранжировок.

Обобщением равномерно распределенных толерантностей являются толерантности с независимыми связями. В этой постановке предполагается, что $a(i, j)$, $1 \le i < j \le k$, - независимые случайные величины, принимающие значения 0 и 1. Обозначим $P(a(i, j) = 1) = p(i,j)$. Тогда $P(a(i, j) = 0) = 1 - p(i,j)$. Таким образом, распределение толерантности с независимыми связями задается нечеткой толерантностью, т.е. вектором

$$P = \{p(i, j), 1 \le i < j \le k\}.$$

Пусть имеется $s$ независимых случайных толерантностей $A_1$, $A_2$, …, $A_s$ с независимыми связями, распределения которых задаются векторами $P_1$, $P_2$, …, $P_s$ соответственно. Рассмотрим проверку гипотезы согласованности

$$H_0: P_1 = P_2 = \ldots = P_s.$$

Она является более слабой, чем гипотеза равномерности

$$H_0^{'}: P_1 = P_2 = \ldots = P_s = (\mathrm{S}, \mathrm{S}, ..., \mathrm{S}),$$

для проверки которой используют статистики $W$ и $N$ (см. выше).

Пусть сначала $s = 2$. Тогда

$$P\{|a_1(i, j) - a_2(i, j)| = 1\} = q(i, j), P\{|a_1(i, j) - a_2(i, j)| = 0\} = 1 - q(i, j),$$

где

$$q(i, j) = p_1(i, j)\,(1 - p_2(i, j)) + p_2(i, j)\,(1 - p_1(i, j)).$$

Следовательно, расстояние $d(A_1, A_2)$ между двумя случайными толерантностями с независимыми связями есть сумма $k(k - 1)/2$ независимых случайных величин, принимающих значения 0 и 1, причем математическое ожидание и дисперсия $d(A_1, A_2)$ таковы:

$$Md(A_1, A_2) = \sum_{1 \le i < j \le k} q(i, j), \quad Dd(A_1, A_2) = \sum_{1 \le i < j \le k} q(i, j)(1 - q(i, j)). \qquad (2)$$

Пусть $k \to \infty$. Если $Dd(A_1, A_2) \to \infty$, то условие Линденберга Центральной Предельной Теоремы теории вероятностей выполнено (см. главу 1.4), и распределение нормированного расстояния

$$\frac{d(A_1, A_2) - Md(A_1, A_2)}{\sqrt{Dd(A_1, A_2)}} \qquad (3)$$

сходится к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1. Если существует число д > 0 такое, что при всех $k$, $i, j$, $1 \le i < j \le k$, вероятности $p_1(i, j)$ и $p_2(i, j)$ лежат внутри интервала (д; 1 – д), то $Dd(A_1, A_2) \to \infty$.

Соотношения (2), (3) и им подобные позволяют рассчитать мощность критериев, основанных на статистиках $W$ и $N$, при $k \to \infty$, подобно тому, как это сделано в [1, глава 4.5]. Поскольку подобные расчеты не требуют новых идей, не будем приводить их здесь.

Обычно $P_1$ и $P_2$ неизвестны. Для проверки гипотезы $P_1 = P_2$ в некоторых случаях можно порекомендовать отвергать гипотезу на уровне значимости б, если $d(A_1, A_2) \ge d_0$, где $d_0$ есть (1-б)-квантиль распределения расстояния между двумя независимыми равномерно распределенными

случайными толерантностями, т.е. квантиль биномиального распределения $B(A)$. Укажем достаточные условия такой рекомендации.

Пусть

$$p = (p_1(i, j) + p_2(i, j))/2, \quad p_1(i, j) = p + Д,$$

тогда

$$p_2(i, j) = p - Д, \quad q = q(i, j) = 2p(1 - p) + 2Д^2. \qquad (4)$$

Если существует число д > 0 такое, что

$$q - S > д > 0 \qquad (5)$$

при всех $k, i, j$, то гипотеза $P_1 = P_2$ будет отвергаться с вероятностью, стремящейся к 1 при $k \to \infty$. Из (4) следует, что при фиксированном $p$ существует Д такое, что выполнено (5), тогда и только тогда, когда $0{,}25 < p < 0{,}75$.

Своеобразие постановки задачи проверки гипотезы состоит в том, что при росте $k$ число неизвестных параметров, т.е. координат векторов $P_i$, растет пропорционально объему данных. Поэтому и столь далекая от оптимальности процедура, как описанная в двух предыдущих абзацах, представляет некоторый практический интерес. Для случая $s \geq 4$ в теории люсианов (глава 3.4.3) разработаны методы проверки гипотезы согласованности $H_0$: $P_1 = P_2 = \ldots = P_s$.

**Нахождение группового мнения.** Пусть $A_1, A_2, \ldots, A_s$ - случайные толерантности, описывающие мнения $s$ экспертов. Для нахождения группового мнения будем использовать медиану Кемени, т.е. эмпирическое среднее относительно расстояния, введенного в главе 1.1. Медианой Кемени является

$$A_{cp} = Arg \min_A \sum_{p=1}^{s} d(A_p, A).$$

Легко видеть, что $A_{cp} = \|a_{cp}(i, j)\|$ удовлетворяет условию: $a_{cp}(i, j) = 1$, если

$$\sum_{p=1}^{s} a_p(i, j) > \frac{s}{2},$$

и $a_{cp}(i, j) = 0$, если

$$\sum_{p=1}^{s} a_p(i, j) < \frac{s}{2}.$$

Следовательно, при нечетном $s$ групповое мнение $A_{cp}$ определяется однозначно. При четном $s$ неоднозначность возникает в случае

$$\sum_{p=1}^{s} a_p(i, j) = \frac{s}{2}.$$

Тогда медиана Кемени $A_{cp}$ - не одна толерантность, а множество толерантностей, минимум суммы расстояний достигается и при $a_{cp}(i,j) = 1$, и при $a_{cp}(i, j) = 0$.

Асимптотическое поведение группового мнения (медианы Кемени для толерантностей) вытекает из общих результатов о законах больших чисел в пространствах произвольной природы (глава 2.1), поэтому рассматривать его здесь нет необходимости.

**Дихотомические (бинарные) признаки в классической асимптотике**. Многое в предыдущем изложении определялось спецификой толерантностей. В частности, особая роль равномерности распределения на множестве всех толерантностей оправдывала специальное рассмотрение статистик $W$ и $N$; аксиоматически введенное расстояние $d$ между толерантностями играло важную роль в приведенных выше результатах. Однако модель толерантностей с независимыми связями уже меньше связана со спецификой толерантностей. В ней толерантности можно рассматривать просто как частный случай люсианов. Широко применяется следующая модель порождения данных.

Пусть $A_1, A_2, \ldots, A_s$ - независимые люсианы. Это значит, что статистические данные имеют вид

$$(A_1, A_2, \ldots, A_s) = \|X_{ij},\ i = 1,2, \ldots, s;\ j = 1, 2, \ldots, k\|, \qquad (6)$$

где $X_{ij}$ - независимые в совокупности испытания Бернулли с вероятностями успеха

$$(P_1, P_2, \ldots, P_s) = \|p_{ij}, ,\ i = 1,2, \ldots, s;\ j = 1, 2, \ldots, k\|, \qquad (7)$$

где $P_i$ - вектор вероятностей, описывающий распределение люсиана $A_i$. Особое значение имеют одинаково распределенные люсианы, для которых $P_1 = P_2 = \ldots = P_s = P$, где символом $P$ обозначен общий вектор вероятностей.

Как обычно в математической статистике, содержательные результаты при изучении модели (6) - (7) можно получить в асимптотических постановках. При этом есть два принципиально разных предельных перехода: $s \to \infty$ и $k \to \infty$. Первый из них - традиционный: число неизвестных параметров постоянно, объем выборки $s$ растет. Во втором число параметров растет, объем выборки остается постоянным, но общий объем данных $ks$ растет пропорционально числу неизвестных параметров. Аналогом является асимптотическое изучение коэффициентов ранговой корреляции Кендалла и Спирмена: число ранжировок, т.е. объем выборки, постоянно (и равно 2), а число ранжируемых объектов растет.

Вторая постановка изучается в следующем подразделе, посвященном люсианам. Некоторые задачи в первой постановке рассмотрим здесь.

Случайные толерантности используются, в частности, для оценки нечетких толерантностей [1]. Для описания результатов опроса группы экспертов о сходстве объектов строят нечеткую толерантность $M = \|м_{ij}\|$, $м_{ij} = l_{ij}/n_{ij}$, где $n_{ij}$ - число ответов о сходстве $i$-го и $j$-го объектов, а $l_{ij}$ - число положительных ответов из них. Если эксперты действуют в соответствии с единым вектором параметров $P$, то $M$ - состоятельная оценка для $P$. Следующий вопрос при таком подходе - верно ли, что две группы экспертов «думают одинаково», т.е. используют совпадающие вектора $P$? Рассмотрим эту постановку на более общем языке люсианов.

Пусть $A_1$, $A_2$, ..., $A_m$ и $B_1$, $B_2$, ..., $B_n$ - независимые в совокупности люсианы, одинаково распределенные в каждой группе с параметрами $P(A)$ и $P(B)$ соответственно. Требуется проверить гипотезу $P(A) = P(B)$. Естественным является переход к пределу при $\min(m, n) \to \infty$.

Пусть гипотеза справедлива. Предположим, что $p_i = p_i(A) = p_i(B) \neq 0$ при всех $i = 1, 2, ..., k$. (Разбор нарушений этого условия очевиден.) Пусть $s_i$ - число единиц на $i$-м месте в первой группе люсианов, а $t_i$ - во второй. Рассмотрим случайные величины

$$\xi_i = \sqrt{\frac{mn}{m+n}}\left(\frac{s_i}{m} - \frac{t_i}{n}\right)\frac{1}{\sqrt{p_i(1-p_i)}}. \qquad (8)$$

Они независимы в совокупности. В соответствии с результатами главы 1.4 распределения $o_i$ при $\min(m, n) \to \infty$ сходятся к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1. Эти свойства сохраняются при замене $p_i$ в (8) на состоятельные оценки, построенные по статистическим данным, соответствующим $i$-му месту. Будем использовать эффективную оценку [17, с.529]

$$p_i^* = \frac{s_i + t_i}{m+n}. \qquad (9)$$

Подставим (9) в (8), получим статистики

$$\xi_i^* = \sqrt{\frac{mn(m+n)}{(s_i+t_i)(m+n-s_i-t_i)}}\left(\frac{s_i}{m} - \frac{t_i}{n}\right).$$

Полученные статистики можно использовать для проверки рассматриваемой гипотезы, например, с помощью критериев, основанных на статистиках

$$W = \frac{1}{\sqrt{k}}\sum_{i=1}^{k} a_i \xi_i^*, \quad T = \sum_{i=1}^{k} (\xi_i^*)^2, \quad \sum_{i=1}^{k} a_i^2 = 1.$$

С помощью результатов главы 1.4 получаем, что $W$ имеет в пределе при $\min(m, n) \to \infty$ стандартное нормальное распределение, а $T$ - распределение хи-квадрат с $k$ степенями свободы.

Рассмотрим распределение статистики $W$ при альтернативных гипотезах. Положим

$$\eta_{1m}^i = \frac{\sqrt{m}\left(\frac{s_i}{m} - p_i(A)\right)}{\sqrt{p_i(A)(1-p_i(A))}}, \quad \eta_{2n}^i = \frac{\sqrt{n}\left(\frac{t_i}{n} - p_i(B)\right)}{\sqrt{p_i(B)(1-p_i(B))}}.$$

Эти случайные величины независимы, распределение каждой из них при $\min(m, n) \to \infty$ сходится к стандартному нормальному распределению. Поскольку

$$\frac{s_i}{m} = \frac{\eta_{1m}^i}{\sqrt{m}}\sqrt{p_i(A)(1-p_i(A))} + p_i(A), \quad \frac{t_i}{n} = \frac{\eta_{2n}^i}{\sqrt{n}}\sqrt{p_i(B)(1-p_i(B))} + p_i(B),$$

то

$$\sqrt{\frac{mn}{m+n}}\left(\frac{s_i}{m}-\frac{t_i}{n}\right)=F+G,$$

где

$$F=\sqrt{\frac{mn}{m+n}}\left(\frac{\eta_{1m}^i}{\sqrt{m}}\sqrt{p_i(A)(1-p_i(A))}-\frac{\eta_{2n}^i}{\sqrt{n}}\sqrt{p_i(B)(1-p_i(B))}\right)$$

и

$$G=\sqrt{\frac{mn}{m+n}}\left(p_i(A)-p_i(B)\right).$$

В силу результатов главы 1.4 распределение $F$ при $\min(m, n) \to \infty$ сближается с нормальным распределением, математическое ожидание которого равно 0, а дисперсия

$$\frac{n}{m+n}p_i(A)(1-p_i(A))+\frac{m}{m+n}p_i(B)(1-p_i(B))\le\frac{1}{4}.$$

Поэтому, чтобы получить собственное (т.е. невырожденное) распределение $W$ при альтернативах, естественно рассмотреть модель

$$p_i(A)=p_i+\frac{\theta_i}{2}\sqrt{\frac{m+n}{mn}}\sqrt{p_i(1-p_i)},\quad p_i(B)=p_i-\frac{\theta_i}{2}\sqrt{\frac{m+n}{mn}}\sqrt{p_i(1-p_i)},\quad i=1,2,...,k,$$

где $u_i$ - некоторые фиксированные числа. Тогда при $\min(m, n) \to \infty$ оценки $p_i^*$ из (9) сходятся к $p_i$ и $\xi_i^*$ являются независимыми асимптотически нормальными случайными величинами с математическими ожиданиями $u_i$ и единичными дисперсиями. Опираясь на результаты главы 1.4, заключаем, что распределение статистики $W$ сходится к нормальному распределению с математическим ожиданием

$$\theta_0=\frac{1}{\sqrt{k}}\sum_{i=1}^{k}a_i\theta_i$$

и единичной дисперсией.

Если в последней формуле $u_0 = 0$, то асимптотическое распределение $W$ таково же, как и в случае справедливости нулевой гипотезы. От указанного недостатка свободна статистика $T$. Тем же путем, как и для $W$, получаем, что при $\min(m, n) \to \infty$ распределение $T$ сходится к нецентральному хи-квадрат распределению с $k$ степенями свободы и параметром нецентральности

$$\Theta=\sum_{i=1}^{k}\theta_i^2\ .$$

Можно рассматривать ряд других задач, например, проверку совпадения параметров для нескольких групп люсианов (аналог дисперсионного анализа), установление зависимости $P(B)$ от $P(A)$ (аналог регрессионного анализа), отнесение вновь поступающего люсиана к одной из групп (задача диагностики - аналог дискриминантного анализа; представляет интерес, например, при применении тестов типа MMPI оценки психического состояния личности) и т.д. Однако принципиальных трудностей на пути развития соответствующих методов не видно, и мы не будем их здесь рассматривать. Создание соответствующих алгоритмов проводится специалистами по прикладной статистике в соответствии с непосредственными заказами пользователей.

### 3.4.3. Теория люсианов

**Асимптотика растущей размерности и проверяемые гипотезы.** Продолжим изучение модели порождения данных (6) - (7) предыдущего подраздела. Будем использовать асимптотику $s$ = const, $k \to \infty$. При этом число неизвестных параметров растет пропорционально объему данных.

В последние десятилетия (с начала 1970-х годов) в прикладной статистике все большее распространение получают постановки, в которых число неизвестных параметров растет вместе с объемом выборки. Результаты, полученные в подобных постановках, называют найденными «в асимптотике растущей размерности» или «в асимптотике А.Н.Колмогорова» [18], перенося терминологию исследований по дискриминантному анализу на общий случай. Как известно, в задаче дискриминации в две совокупности академик АН СССР А.Н. Колмогоров (1903 - 1987) предложил рассматривать асимптотику

$$A \to \infty,\ \ N_i \to \infty,\ \ \frac{A}{N_i} \to \lambda_i > 0,\ \ i = 1,2\ ,$$

где $A$ - размерность пространства (число признаков), $N_i$ - объемы обучающих выборок, л$_i$ - константы, $i$ = 1,2. Эта асимптотика естественна при обработке организационно-экономических, социологических, медицинских данных, поскольку число признаков, определяемых для каждого изучаемого объекта, респондента или пациента, обычно имеет тот же порядок, что и объем выборки.

Пусть $A_1$, $A_2$, ..., $A_s$ - независимые (между собой) люсианы с векторами параметров $P_1$, $P_2$, ..., $P_s$ соответственно. *Гипотезой согласованности* будем называть гипотезу

$$P_1 = P_2 = ...= P_s. \qquad (1)$$

Для ранжировок и разбиений под согласованностью понимают более частную гипотезу, предполагающую отрицание равномерности распределений (т.е. одинаковой вероятности появления каждой возможной ранжировки или разбиения), что соответствует замене проверки гипотезы (1) на проверку гипотезы

$$P_1 = P_2 = ...= P_s = (1/2,\ 1/2,\ ...,\ 1/2). \qquad (2)$$

Как разъяснено в [1,2], гипотеза (1) более адекватна конкретным задачам обработки реальных данных, например, экспертных оценок, чем (2). Поэтому полученные от экспертов данные, содержащие противоречия, целесообразно рассматривать как люсианы и проверять гипотезу (1), а не подбирать ближайшие ранжировки или разбиения , после чего проверять согласованность методами теории случайных ранжировок или разбиений, как иногда рекомендуется.

Пусть $A_1$, $A_2$, ..., $A_m$ и $B_1$, $B_2$, ..., $B_n$ - независимые в совокупности люсианы длины $k$, одинаково распределенные в каждой группе с параметрами $P(A)$ и $P(B)$ соответственно. *Гипотезой однородности* называется гипотеза

$$P(A) = P(B).$$

В асимптотике растущей размерности принимаем, что $m$ и $n$ постоянны, а $k \to \infty$.

Пусть ($A_i$, $B_i$), $i$ = 1. 2, ..., $s$ - последовательность (фиксированной длины) пар люсианов. Пары предполагаются независимыми между собой. Требуется проверить гипотезу независимости $A_i$ и $B_i$, т.е. внутри пар. В ранее введенных обозначениях *гипотеза независимости* - это гипотеза

$$P(X_{ij}(A) = 1, X_{ij}(B) = 1) = P(X_{ij}(A) = 1)P(X_{ij}(B) = 1),$$
$$i = 1, 2, ..., s; j = 1, 2, ..., k,$$

проверяемая в предположении

$$P_1(A) = P_2(A) = ... = P_s(A),\ P_1(B) = P_2(B) = ... = P_s(B).$$

В настоящем подразделе излагается метод проверки гипотез о люсианах в асимптотике растущей размерности на примере гипотезы согласованности. Эти результаты получены в [1, 18, 19]. Дальнейшее изучение проведено нашими учениками Г.В. Рыдановой, Т.Н. Дылько, Г.В. Раушенбахом, О.В. Филипповым, А.М. Никифоровым и др. Гипотеза однородности рассмотрена, например, в [19]. Методы проверки гипотезы однородности люсианов развиты и изучены Г.В. Рыдановой [20] на основе описанного ниже подхода. Она помимо доказательства предельных теорем провела подробное изучение скорости сходимости методом статистических испытаний.

Методы проверки согласованности люсианов нашли практическое применение, в частности, в медицине. Они были использованы в кардиологии при анализе данных кинетотопографии [19, 21, 22]. Эти методы включены в методические рекомендации Академии медицинских наук СССР и Ученого Медицинского Совета Минздрава СССР по управлению научными медицинскими исследованиями [23].

**Метод проверки гипотез о люсианах в асимптотике растущей размерности.** Будем использовать дальнейшее развитие метода, описанного в главе 2.3.4. Почему нельзя использовать иные подходы, имеющиеся в математической статистике? Поскольку число неизвестных параметров растет вместе с объемом выборки и пропорционально ему, эти параметры не являются мешающими. Отметим, что согласно [24] равномерно наиболее мощных критериев не существует, поскольку параметров много. Не останавливаясь на других подходах математической статистики, констатируем необходимость применения метода проверки гипотез по совокупности малых выборок.

Пусть имеются $k$ выборок, независимых между собой. Пусть при справедливости нулевой гипотезы по каждой из выборок можно построить несмещенную оценку $\xi_i \in R^p$ векторного нуля $0 \in R^p$, где $p \geq 1$, $i$ = 1, 2, ..., $k$. Другими словами, пусть распределение $i$-ой выборки описывается

параметром $и_i$, лежащим в произвольном пространстве, а нулевая гипотеза, очевидно, состоит в том, что $и_i \in И_{0i}$, где $И_{0i}$ - собственное подмножество множества $\{и_i\}$. Предполагается, что можно по $i$-ой выборке вычислить статистику $о_i$ такую, что

$$M о_i = 0 \qquad (3)$$

при всех $и_i \in И_{0i}$. Очевидно, $о_i \equiv 0$ удовлетворяют (1). Однако для рассматриваемого метода необходимо, чтобы при всех $и_i \in И_{0i}$ ковариационная матрица вектора $о_i$ была ненулевой:

$$Cov(\xi_i) = M(\xi_i^T \xi_i) \neq 0 . \qquad (4)$$

В теории математической статистики иногда используют понятие полноты параметрического семейства распределений. Если рассматриваемое семейство является полным - а так и есть для люсианов, - то не существует достаточной статистики, удовлетворяющей одновременно условиям (1) и (2) (см., например, [25, §§2.12-2.14]). Поэтому будем использовать статистики, не являющиеся достаточными.

Следующее предположение - ковариационные матрицы статистик $о_i$, т.е. $Cov(о_i)$, также допускают несмещенные оценки $S_i$ по тем же выборкам:

$$M(S_i) = Cov(о_i) \qquad (5)$$

при всех $и_i \in И_{0i}$.

Рассматриваемый метод основан на том, что поскольку случайные вектора $о_i$ определяются по независимым между собой выборкам, то $о_i$ независимы в совокупности, а потому случайный вектор

$$\xi = \sum_{i=1}^{k} \xi_i \qquad (6)$$

является суммой независимых случайных векторов, имеет в силу (3) нулевое математическое ожидание, а его ковариационная матрица равна

$$C_k = \sum_{i=1}^{k} Cov(\xi_i) .$$

При справедливости многомерной центральной предельной теоремы (простейшее условие справедливости этой теоремы для $о_i$ в случае люсианов - отделенность от 0 и 1 всех элементов матриц $P_j$, равномерная по $s$ и $k$) вектор о является асимптотически нормальным, т.е. при $k \to \infty$ распределение о сближается (в смысле, раскрытом в главе 1.4) с многомерным нормальным распределением $N(0; C_k)$.

Однако эту сходимость нельзя непосредственно использовать для проверки исходной гипотезы, поскольку матрица $C_k$ неизвестна статистику. Необходимо оценить эту матрицу по статистическим данным. В силу (5) в качестве оценки $C_k$ естественно использовать

$$C_k^* = \sum_{i=1}^{k} S_i .$$

Простейшая формулировка условий справедливости такой замены - предположение о том, что к последовательности $S_i$ можно применить закон больших чисел. А именно, пусть существует неотрицательно определенная матрица $C$ такая, что при $k \to \infty$

$$\frac{1}{k}(C_k^* - C_k) \to 0, \quad \frac{1}{k} C_k \to C . \qquad (7)$$

В силу результатов главы 1.4 из асимптотической нормальности о и соотношений (7) следует, что распределение статистики

$$\eta = \frac{1}{\sqrt{k}} \xi$$

сходится к нормальному распределению $N(0; C)$. При этом, если некоторый случайный вектор ф имеет распределение $N(0; C)$, то распределение случайной величины $q(з)$ сходится к распределению $q(ф)$ для произвольной интегрируемой по Риману по любому кубу функции $q$: $R^p \to R^1$. Для проверки нулевой гипотезы предлагается пользоваться статистикой $q(з)$ при подходящей функции $q$, а процентные точки брать соответственно распределению $q(ф)$. В этом и состоит рассматриваемый метод проверки гипотез о люсианах в асимптотике растущей размерности. Для реальных расчетов целесообразно использовать линейные или квадратические функции $q$ от координат вектора з.

Отклонения от нулевой гипотезы приводят, как правило, к нарушению равенств (3) и (4). Случайный вектор з при этом обычно остается асимптотически нормальным, но с другими параметрами, что может быть обычным образом использовано для построения оптимального решающего правила, соответствующего заданной альтернативе (например, согласно лемме Неймана-Пирсона). Поведение при альтернативах для некоторых гипотез изучено в [19, 20], здесь его не будем рассматривать, поскольку вычисление мощности не требует новых идей.

**Несмещенные оценки параметров асимптотического распределения вектора попарных расстояний.** Применим описанный выше метод для проверки гипотезы согласованности люсианов. Исходные данные - люсианы

$$A_j = (X_{1j}, X_{2j}, ..., X_{kj}), j = 1, 2, ..., s.$$

В качестве *i*-й выборки возьмем совокупность испытаний Бернулли, стоящих на *i*-м месте в рассматриваемых люсианах:

$$X_{i1}, X_{i2}, ..., X_{is}. \qquad (8)$$

При справедливости нулевой гипотезы в (8) стоят независимые испытания Бернулли с одной и той же вероятностью успеха $p_i$; при нарушении нулевой гипотезы согласованности независимость испытаний Бернулли сохраняется, но вероятности успеха могут различаться.

В качестве вектора о, на основе которого строятся статистики для проверки согласованности, будем использовать вектор попарных расстояний между люсианами

$$\text{о} = \{d(A_p, A_q), 1 \le p < q \le s\}, \qquad (9)$$

в котором пары $(p, q)$ упорядочены лексикографически,

$$d(A_p, A_q) = \sum_{i=1}^{k} \mu_i \mid X_{ip} - X_{iq} \mid, \quad \mu_i > 0. \qquad (10)$$

В главе 1.1 это расстояние выведено из некоторой системы аксиом (напомним, что совокупность векторов из 0 и 1 размерности k находится во взаимнооднозначном соответствии с совокупностью подмножеств множества из *k* элементов; при этом 1 соответствует тому, что элемент входит в подмножество, а 0 - что не входит).

Из вида расстояния в формуле (10) следует, что введенный в (9) вектор о имеет вид (6) с

$$\text{о}_i = \text{м}_i\{|X_{ip} - X_{iq}|, 1 \le p < q \le s\}. \qquad (11)$$

Следовательно, для применения описанного выше метода проверки гипотез о люсианах в асимптотике растущей размерности достаточно построить на основе вектора $\text{о}_i$ из (11) несмещенную оценку 0 и найти несмещенную оценку ковариационной матрицы этой оценки.

Чтобы применить общую схему, необходимо начать с построения статистики в такой, чтобы при всех $p_i$ имело место равенство

$$M(|X_{ip} - X_{iq}| - в) - 0, \ 1 \le p < q \le s.$$

Элементарный расчет дает:

$$M|X_{ip} - X_{iq}| = 2p_i(1 - p_i).$$

Как известно [5, с.56-57], несмещенная оценка многочлена

$$f(p) = \sum_{h=0}^{m} a_h p^h$$

по результатам *m* независимых испытаний Бернулли с вероятностью успеха *p* в каждом имеет вид

$$f^*(p) = \sum_{h=0}^{m} a_h \frac{\gamma^{[h]}}{m^{[h]}}, \qquad (12)$$

где г - общее число успехов в m испытаниях и использовано обозначение

$$n^{[h]} = n(n - 1)...(n - h + 1).$$

Ясно, что многочлены степени $m + 1$ и более высокой невозможно несмещенно оценить по результатам *m* испытаний.

В случае $f(p) = 2p(1 - p)$ в соответствии с (12) получаем несмещенную оценку

$$\beta = \frac{2}{m-1}\left(\gamma - \frac{\gamma^2}{m}\right). \qquad (13)$$

Таким образом, можно применять общий метод проверки гипотез о люсианах в асимптотике растущей размерности с

$$\text{о}_i = \text{м}_i(\{|X_{ip} - X_{iq}|, 1 \le p < q \le s\} - в_i е),$$

где коэффициенты $в_i$ определяются с помощью формулы (13) по $г_i$ - общему числу единиц, стоящих на $i$-м месте в люсианах $A_1$, $A_2$, ..., $A_s$, а $e$ - вектор размерности $s(s-1)/2$ с единичными координатами. Тогда несмещенная оценка 0, о которой идет речь в методе проверки гипотез по совокупности малых выборок, имеет вид

$$\xi = \{d(A_p, A_q),\ 1 \le p < q \le s\} - \sum_{i=1}^{k} \mu_i \beta_i e .$$

Для использования статистики типа з, распределение которой приближается с помощью нормального распределения

$$N\left(0;\ \frac{1}{k}\sum_{i=1}^{k} S_i\right),$$

необходимо уметь несмещенно оценивать ковариационные матрицы $Cov(о_i)$. Для этого достаточно найти математические ожидания элементов матрицы $M(\xi_i^T \xi_i)$ как функции (многочлены) от $p_i$, а затем использовать формулу (12) для получения несмещенных оценок.

Вычисление матрицы $M(\xi_i^T \xi_i)$ хотя и трудоемко, но не содержит каких-либо принципиальных трудностей. В [19] вычислены диагональные элементы рассматриваемой матрицы. Вычисление занимает около 2,5 книжных страниц (с.299-301). Поэтому здесь приведен только окончательный итог.

Обозначим для краткости $p_i = p$. В [19] показано, что

$$D = D(|X_{ip} - X_{iq}| - \beta_i) = \left(2 - \frac{4}{s}\right) p(1-p) - 4\frac{(s-2)(s-3)}{s(s-1)} p^2 (1-p)^2 .$$

Если двухэлементные множества $\{p, q\}$ и $\{r, t\}$ не имеют ни одного общего элемента, то

$$C_1 = M(|X_{ip} - X_{iq}| - \beta_i)(|X_{ir} - X_{it}| - \beta_i) = -\frac{4}{s} p(1-p) + \frac{8(2s-3)}{s(s-1)} p^2 (1-p)^2 ,$$

а если имеют ровно один общий элемент, то

$$C_2 = M(|X_{ip} - X_{iq}| - \beta_i)(|X_{ir} - X_{it}| - \beta_i) = \left(1 - \frac{4}{s}\right) p(1-p) - 4\frac{(s-2)(s-3)}{s(s-1)} p^2 (1-p)^2 .$$

С помощью формулы (12) получаем несмещенные оценки для $D$, $C_1$ и $C_2$ как многочленов от $p$:

$$D^* = \frac{2\gamma_i(s-\gamma_i)}{s^2(s-1)^2}\{(s-2)(s-1) - 2(\gamma_i - 1)(s - \gamma_i - 1)\},$$

$$C_1^* = \frac{4\gamma_i(s-\gamma_i)}{s^2(s-1)}\left\{\frac{2(2s-3)(\gamma_i - 1)(s-\gamma_i - 1)}{(s-1)(s-2)(s-3)} - 1\right\},$$

$$C_2^* = \frac{\gamma_i(s-\gamma_i)}{s^2(s-1)^2}\{(s-4)(s-1) - 4(\gamma_i - 1)(s - \gamma_i - 1)\}.$$

С помощью трех чисел $D^*, C_1^*, C_2^*$ выписывается несмещенная оценка матрицы ковариаций вектора $о_i/м_i$, которую обозначим $B_i$. Тогда асимптотически нормальный вектор о имеет нулевое математическое ожидание и ковариационную матрицу, несмещенно и состоятельно (в смысле соотношений (7)) оцениваемую с помощью

$$Cov(\xi)^* = \sum_{i=1}^{k} \mu_i^2 B_i . \qquad (14)$$

Асимптотическая нормальность доказывается, естественно, в схеме серий. Достаточным условием является существование положительной константы е такой, что

$$\mu_i \ge \varepsilon,\ \frac{1}{\mu_i} \ge \varepsilon,\ p_i \ge \varepsilon,\ 1 - p_i \ge \varepsilon \qquad (15)$$

при всех $k$ и i, $1 \le i \le k$.

Поскольку $D$, $C_1$ и $C_2$ являются многочленами четвертой степени от $p$, то несмещенные оценки для них существуют при $s \ge 4$. Если же $s < 4$, то несмещенных оценок не существует. Поэтому указанным методом проверять согласованность можно лишь при числе люсианов $s \ge 4$.

**Проверка согласованности люсианов.** Пусть б - нормально распределенный случайный вектор размерности $s(s-1)/2$ с нулевым математическим ожиданием и ковариационной матрицей,

определенной формулой (14). Согласно результатам главы 1.4 для любой действительнозначной функции $f$, интегрируемой по Риману по любому гиперкубу, распределения случайных величин $f$(о) и $f$(б) сближаются при $k \to \infty$. Это означает, что вместо распределения $f$(о) для построения критериев проверки гипотез можно использовать распределение $f$(б). Более того, аналогичный результат верен при замене $f$ на $f_n$ (при слабых внутриматематических условиях регулярности, наложенных на последовательность функций $f_n$). Следовательно, для проверки гипотезы согласованности люсианов можно пользоваться любой статистикой $f_n$(о), для которой могут быть вычислены на ЭВМ или заранее табулированы процентные точки распределения $f_n$(б), аппроксимирующего распределение $f_n$(о).

В частности, можно использовать линейные статистики, представляющие собой скалярное произведение случайного вектора о и некоторого заданного детерминированного вектора коэффициентов $a$, т.е.

$$(\xi, a) = \sum_{i=1}^{k} \left( \mu_i \sum_{1 \le j < t \le s} a_{jt} (| X_{ij} - X_{it} | - \beta_i) \right). \qquad (16)$$

Линейные статистики имеют нулевое математическое ожидание и дисперсию, очевидным образом выражающуюся через матрицу коэффициентов $||a_{ij}||$ и числа $D$, $C_1$ и $C_2$, а потому несмещенно и состоятельно оцениваемую с помощью с помощью выписанных выше оценок для $D$, $C_1$ и $C_2$.

Отметим, что (о, $a$) = 0 при $a_{ij} \equiv 1$, $1 \le j < t \le s$. Это следует как из непосредственного вычисления дисперсии (о, $a$), так и из того, что (о, $a$) в рассматриваемом случае выражается через достаточную статистику ($г_1$, $г_2$, ..., $г_k$) и является несмещенной оценкой нуля, а семейство биномиальных распределений полно, т.е. существует только одна несмещенная оценка нуля - тождественный нуль. Таким образом, сумма координат вектора о, т.е. непосредственный аналог коэффициента ранговой конкордации Кендалла-Смита из теории ранговой корреляции, тождественно равна 0.

Распределение статистики (16) при альтернативах изучено в работе [20].

Рассмотрим два частных случая.

*Первый частный случай.* Проверка согласованности двух определенных люсианов (ответов двух экспертов), $j$-го и $t$-го, может осуществляться с помощью статистики (16), в которой отличен от 0 только член с $a_{jt} = 1$. Оценкой дисперсии является $D$*.

*Второй частный случай.* Пусть необходимо проверить согласованность люсианов с одним из них, скажем, с $j$-м (например, люсианы отражают мнения экспертов, а $j$-й из них является наиболее компетентным - по априорной оценке, или «лицом, принимающим решения», или его мнение сильно отличается от мнений остальных). Это можно сделать с помощью статистики (16), в которой

$$a_{jt} = 1,\ t = j+1, j+2, ..., s; \quad a_{tj} = 1,\ t = 1, 2, ..., j-1;$$
$$a_{qt} = 0,\ q \ne j,\ t \ne j,\ 1 \le q < t \le s.$$

Другими словами, она имеет вид

$$W = \sum_{t=1}^{s} d(A_j, A_t) - (s-1) \sum_{i=1}^{k} \mu_i \beta_i , \qquad (17)$$

где расстояние $d$ между люсианами определено в (10), а $в_i$ - в (13) с заменой $m$ на $s$ и г на $г_i$. Используя полученные ранее несмещенные оценки элементов ковариационной матрицы, нетрудно показать, что несмещенная и состоятельная (в смысле формулы (7) выше) оценка дисперсии $W$ имеет вид

$$D^*(W) = \sum_{i=1}^{k} \mu_i^2 \frac{\gamma_i (s - \gamma_i)}{s^2} \{(s-2)^2 - 4(\gamma_i - 1)(s - \gamma_i - 1)\}.$$

Тогда при выполнении некоторых внутриматематических условий регулярности, например, условий (15), распределение статистики

$$\frac{1}{\sqrt{D^*(W)}} W$$

сходится при $k \to \infty$, $s$ = const к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1 (при справедливости гипотезы (1) согласованности люсианов).

Статистика (17) наряду со статистикой, предназначенной для проверки гипотезы однородности люсианов, включена в «Методические рекомендации» АМН СССР и УМС

Минздрава СССР [23]. Последнюю статистику не расписываем здесь, поскольку для этого не требуются новые идеи.

Различные подходы к понятию согласованности. Обсудим условия, при выполнении которых люсианы естественно считать согласованными (а экспертов, чьи мнения отражают люсианы, имеющими единое мнение, искаженное случайными ошибками), т.е. обсудим различные методы проверки гипотезы (1).

*Полное индивидуальное согласие* имеет место, если никакие два эксперта не являются «несогласованными». Уровень значимости определяется описанным выше способом (первый частный случай). Однако наличие одной или нескольких пар экспертов, чьи мнения нельзя считать согласованными, не свидетельствует о необходимости отклонения гипотезы (1), поскольку парных проверок проводится много, а именно, $s(s - 1) \geq 6$, а способы установления уровня значимости при множественных проверках, зависимых между собой, к настоящему времени плохо разработаны (см. главу 2.3.5). Проблема множественных проверок для количественных признаков обсуждается А.А. Любищевым [26, с.36-39], выход дается дисперсионным анализом. Можно брать не все попарные проверки, а только для [$s$/2] пар люсианов, причем разбиение на пары проводить независимо от принятых люсианами значений, как это делает Т.Н. Дылько [27]. Тогда для проверки гипотезы (1) на уровне значимости б надо брать для проверки в каждой паре уровень значимости в, где в рассчитывается понятным образом, приближенно в = б / [$s$/2].

*Полное согласие в целом* означает, что для любого эксперта мнения всех остальных оказываются с ним согласованными при использовании статистики (17) (второй частный случай). Отсутствие подобного согласия для одного или нескольких экспертов не означает отклонения гипотезы согласованности люсианов (1) - по тем же причинам, что и в предыдущем случае.

*Минимальное согласие* имеют мнения экспертов, когда хотя бы для одного из них гипотеза согласованности не отвергается с помощью статистики (17). В этом случае групповое мнение целесообразно строить, выделяя «ядро», о чем подробнее сказано ниже.

Расстояние $d$ между люсианами (см. формулу (10)) введено аксиоматически в главе 1.1.6 (напомним, что реализацию люсиана можно рассматривать как подмножество конечного множества). Там же из иной системы аксиом выведено другое расстояние - $D$-метрика. Рассмотрим проверку согласованности люсианов с использованием $D$-метрики. В этом случае расстояние между люсианами $A_1$ и $A_2$ имеет вид

$$D(A_1, A_2) = \begin{cases} \dfrac{d(A_1, A_2)}{T(A_1, A_2)}, & T(A_1, A_2) \neq 0, \\ 0, & T(A_1, A_2) = 0, \end{cases}$$

где

$$T(A_1, A_2) = \sum_{i=1}^{k} \mu_i \max(X_{i1}, X_{i2}) .$$

Ясно, что теория, основанная на $D$-метрике, существенно сложнее теории, основанной на метрике $d$. Ясно, что описанный выше метод проверки гипотез о люсианах в асимптотике растущей размерности применить не удается. Чтобы продемонстрировать существенное усложнение ситуации, опишем лишь асимптотическое поведение расстояния $D(A_1, A_2)$ между двумя люсианами.

*Теорема* [28]. Пусть $p_{1i}$ и $p_{2i}$ отделены от 0 и 1, а $м_i$ отделены от 0 и $+\infty$. Тогда расстояние $D(A_1, A_2)$ между люсианами $A_1$ и $A_2$ асимптотически нормально при $k \to \infty$ с параметрами

$$t_k = \frac{N_1}{N_2}, \quad q_k = \frac{N_1}{N_2} \sqrt{\frac{N_3}{N_1^2} + \frac{N_4}{N_2^2} - 2\frac{N_5}{N_1 N_2}} ,$$

т.е. для любого числа $x$ справедливо предельное соотношение

$$\lim_{k \to \infty} P\left\{ \frac{D(A_1, A_2) - t_k}{q_k} \leq x \right\} = \Phi(x) ,$$

где Ц($x$) - функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1.

Величины $N_j$, $j$ = 1, 2, 2, 4, 5, выражаются через $м_i$ и величины

$$p_{3i} = p_{1i} + p_{2i} - 2p_{1i}\, p_{2i},\ p_{4i} = p_{1i} + p_{2i} - p_{1i}\, p_{2i}$$

следующим образом:

$$N_1 = \sum_{i=1}^{k} \mu_i p_{3i}, \quad N_2 = \sum_{i=1}^{k} \mu_i p_{4i}, \quad N_3 = \sum_{i=1}^{k} \mu_i^2 p_{3i}(1-p_{3i}),$$

$$N_4 = \sum_{i=1}^{k} \mu_i^2 p_{4i}(1-p_{4i}), \quad N_5 = \sum_{i=1}^{k} \mu_i^2 p_{3i}(1-p_{4i}).$$

*Следствие 1.* Пусть $p_{1i} = p_1$ и $p_{2i} = p_2$ при всех *i, k*, причем $p_1$ и $p_2$ лежат внутри отрезка (0; 1). Пусть $м_i$ отделены от 0 и $+\infty$. Тогда расстояние $D(A_1, A_2)$ между люсианами $A_1$ и $A_2$ асимптотически нормально при $k \to \infty$ с параметрами

$$t_k = \frac{p_3}{p_4}, \quad q_k^2 = \frac{p_1 p_2 p_3}{p_4^3} \frac{\sum_{i=1}^{k} \mu_i^2}{\left( \sum_{i=1}^{k} \mu_i \right)^2},$$

где

$$p_3 = p_1 + p_2 - 2p_1p_2, \quad p_4 = p_1 + p_2 - p_1p_2.$$

*Следствие 2.* Пусть в предположениях следствия 1 $p_1 = p_2 = p$ и $м_i = 1$ при всех *i, k*. Тогда

$$t_k = \frac{2(1-p)}{2-p}, \quad q_k = \frac{2(1-p)}{k(2-p)^3}.$$

*Замечание.* Пусть в следствии 2 $p = 1/2$. Тогда $A_1$ и $A_2$ - люсианы, равномерно распределенные на множестве всех последовательностей из 0 и 1 длины *k*. В частности, эти люсианы могут соответствовать независимым случайным множествам, равномерно распределенным на совокупности всех подмножеств конечного множества из *k* элементов, или независимым толерантностям, равномерно распределенным на множестве всех толерантностей, определенных на множества из *m* элементов, где $m(m - 1)/2 = k$. По следствию 2 расстояние между люсианами $D(A_1, A_2)$ асимптотически нормально с математическим ожиданием 0,667 и дисперсией $0{,}296\ k^{-1}$. Напомним, что распределения коэффициентов ранговой корреляции Кендалла и Спирмена изучены (в основном) лишь при условии равномерности распределения случайных ранжировок на множестве всех возможных ранжировок фиксированного числа объектов. Для теории люсианов случай равномерности распределения - весьма частный, а для теории ранжировок - основной. Как уже говорилось, отказ от равномерности - привлекательная черта теории люсианов.

**Классификация люсианов.** Отсутствие согласованности в одном из перечисленных выше смыслов позволяет сделать заключение о целесообразности разбиения всех люсианов (например, если они выражают мнения экспертов) на группы близких между собой, т.е. о целесообразности классификации люсианов, точнее, их кластер-анализа. Поскольку введена мера близости между люсианами $d(A_1, A_2)$ или $D(A_1, A_2)$, то напрашивается следующий способ действий: провести разбиение на кластеры с помощью одного из алгоритмов, основанных на использовании меры близости, а затем проверить мнения в каждом классе на согласованность. Однако применение того или иного алгоритма кластер-анализа, вообще говоря, может нарушить предпосылки описанных выше способов описанных выше способов проверки согласованности (ср. обсуждение похожей проблемы, связанной с применением регрессионного анализа после кластер-анализа, в главе 2.3.5). Поэтому опишем методы классификации, опирающиеся на результаты проверки согласованности.

Разбиение на кластеры, внутри каждого из которых имеет место «полное индивидуальное согласие», может быть проведено с помощью агломеративного иерархического алгоритма «дальнего соседа», дополненного ограничением сверху на диаметр кластера. Это ограничение строится из статистических соображений, в отличие от методов, описанных в главе 3.2. При этом в качестве меры близости между люсианами используют не расстояния *d* или *D*, а модуль статистики, применяемой для проверки согласованности двух люсианов, т.е. статистики (16), в которой только одно из чисел $a_{ij}$ отлично от 0. Упомянутое ограничение таково: диаметр кластера не должен превосходить процентной точки предельного распределения, соответствующей используемому при анализе рассматриваемых данных уровню значимости (можно порекомендовать 5%-й уровень значимости). В результате работы алгоритма получим кластеры, в которых имеется «полное индивидуальное согласие», причем объединение любых двух кластеров приведет к исчезновению этого свойства у объединения. Поскольку способ выделения итогового

разбиения из иерархического дерева разбиений имеет вероятностно-статистическое обоснование, изложенное выше, то описанный метод классификации люсианов следует считать - в терминологии [29] - не методом анализа данных, а вероятностно-статистическим методом.

Кластеры «с полным согласием в целом» могут быть получены с помощью агломеративного иерархического алгоритма, в котором мерой близости двух кластеров является максимальное значение модуля статистики (17), когда $j$ пробегает номера мнений (люсианов), вошедших в объединение рассматриваемых кластеров, а суммирование в (17) проводится по всем люсианам в этом объединении. Ограничение на сверху на меру близости кластеров определяется процентной точкой предельного распределения статистики $W$, заданной формулой (17).

Кластеры «с минимальным согласием» можно получить, при фиксированном $j$ выделяя совокупность люсианов, согласованных с $A_j$ в смысле статистики $W$ из (17).

На основе двух рассмотренных выше частных случаев линейной статистики (16) можно строить и другие способы классификации. Например, для каждого люсиана $A_m$ можно выделить кластер «типа шара» (см. главу 3.2) из люсианов, попарно согласованных с $A_m$. Все такие способы имеют вероятностно-статистическое обоснование, и потому к ним относится сказанное выше относительно выделения кластеров «с полным индивидуальным согласием».

*Замечание.* Проверка согласованности приведенными выше критериями может привести к отрицательному результату двумя способами - либо значение статистики окажется слишком большим, либо слишком малым. Первое означает, что гипотеза согласованности люсианов (1) неверна, вторая - что неверна вероятностная модель реального явления или процесса, основанная на люсианах. С необходимостью учета второй возможности мы столкнулись при применении теории люсианов для анализа данных топокарт, полученных при проведении кинетокардиографии у больных инфарктом миокарда [21, 22].

**Нахождение среднего.** В результате классификации получаем согласованные (в одном из указанных выше смыслов) группы люсианов. Для каждой из них полезно рассмотреть среднее. В зависимости от конкретных приложений в прикладных исследованиях применяют либо среднее в виде последовательностей 0 и 1, т.е. в виде реализации люсиана, либо среднее в виде последовательности оценок вероятностей ($p_1$, $p_2$, ..., $p_k$). Кроме того, оно может находиться либо с помощью методов, подавляющих «засорения» («выбросы»), либо без учета возможности засорения. Рассмотрим все четыре возможности.

В соответствии с подходом главы 2.1.5 при отсутствии засорения эмпирическое среднее ищется как решение задачи

$$\sum_{j=1}^{m} d(A_j, A) \to \min_{A \in X}, \qquad (18)$$

где $A_1$, $A_2$, ..., $A_m$ - люсианы, входящие в рассматриваемый кластер, $X$ - множество, которому принадлежит среднее.

Если $X$ - совокупность последовательностей из 0 и 1, то правило (18) дает решение по правилу большинства (подробнее см. главу 2.1.5).

Если $X$ - пространство последовательностей вероятностей, то решением задачи (18) является та же последовательность 0 и 1, что и в первом случае. Поэтому в качестве среднего вместо решения задачи (18) целесообразно рассматривать просто последовательность частот.

Асимптотическое поведение средних при $m \to \infty$ вытекает из законов больших чисел (глава 2.1.5), теорем, описывающих асимптотику решений экстремальных статистических задач (глава 2.2.3), и теоремы Муавра-Лапласа соответственно.

В работе [30] при анализе результатов эксперимента показано, что ответы реальных экспертов разбиваются на многочисленное «ядро», расположенное вокруг истинного мнения, и отдельных «диссидентов», разбросанных по периферии. Причем оценка истинного мнения по «ядру» является более точной, чем по все совокупности, поскольку мнения «диссидентов» не отражают истинного мнения. Поэтому для построения группового мнения, в том числе среднего для совокупности люсианов, отражающих мнения экспертов, естественно применять методы, подавляющие мнения «диссидентов», что соответствует методологии робастности.

«Ядро» может быть построено следующим образом. Решается задача (18) с конечным множеством $X$, состоящим из всех исходных люсианов: $X = \{A_1, A_2, ..., A_m\}$, т.е. из результатов наблюдений выбирается тот, что находится «в центре» совокупности результатов наблюдений. Пусть $A_j$ является решением этой задачи. В качестве ядра предлагается рассматривать

совокупность всех люсианов, которые попарно согласованы с $A_j$. Другой вариант: рассматривается кластер с «полным внутренним согласием», куда входит $A_j$. (При этом, очевидно, должно быть изменено (уменьшено) критическое значение критерия по сравнению с процедурой, приведшей к выделению группы, нахождением группового мнения которой мы занимаемся.) Затем групповое мнение ищется лишь для элементов «ядра». Описанная процедура особенно необходима в случае, когда не было предварительного разбиения совокупности люсианов на группы согласованных друг с другом. Новым по сравнению с [30] является придание вероятностного смысла порогу, выделяющему «ядро».

Обобщая идею выделения «ядра», приходим к «взвешенным итеративным методам оценивания среднего» (ВИМОП - оценкам среднего), введенным и изученным в работе [31]. Их применение для люсианов не требует специальных рассмотрений.

Таким образом, в настоящем подразделе представлен ряд методов обработки специального вида объектов нечисловой природы - люсианов. При этом для решения одной и той же задачи, например, задачи классификации, предлагается ряд методов, точно так же, как для решения классической задачи проверки однородности двух независимых выборок имеется большое число методов (см. главу 3.1).

### 3.4.4. Метод парных сравнений

**Пример практического применения метода парных сравнений.** Деятельность предприятия по реализации услуг всегда сопряжена с рядом проблем, от качества решения которых зависит его будущее. Руководителю службы маркетинга необходимо знать факторы, сдерживающие продажи, и оценить степень важности каждого из них. При кажущейся очевидности и простоте решения далеко не вся управленческая команда дает однозначный ответ: какая из проблем на текущий момент является наиболее важной. Необходим экспертный опрос на эту тему.

Целью исследования факторов, влияющих на объемы продаж, является их ранжирование по степени важности. Для этого среди 25 сотрудников отдела сбыта, а также 10 руководителей завода ГАРО (Великий Новгород) А.А. Пивнем был проведен опрос, в котором предлагалось сравнить попарно факторы, определив более важный среди двух. Итог определялся как среднее арифметическое сумм баллов набранных каждым фактором у всех опрашиваемых.

Были проанализированы следующие 15 факторов:

- потребительские свойства изделий (качество, надежность, показатели назначения и т.д.);
- уровень цен;
- срок поставки продукции;
- информация о предлагаемых к продаже изделиях;
- уровень гарантийного и сервисного обслуживания;
- работа дилеров, представительств;
- рекламная деятельность;
- численность персонала;
- мотивация труда;
- инициативность персонала;
- маркетинговая деятельность;
- оснащенность техническими средствами;
- квалификация персонала;
- корпоративная культура;
- репутация Компании.

В результате анализа результатов парных сравнений построена структурная схема, показывающих степень влияния факторов на объемы продаж (рис.1).

Наибольшую значимость на сегодняшний день имеет срок поставки продукции и квалификация персонала. Меняются подходы к продвижению товаров на рынке. Ранее успешно применяемые способы продаж (почтовая рассылка рекламы, участие в специализированных выставках, публикации в газетах и специализированных изданиях, конференции и т.д.) сегодня требуют иного качественного подхода. Срок поставки продукции, как правило, связан с производственно-технологическим циклом изготовления и настройки изделий. Мотивация труда, равно как и уровень гарантийного и сервисного обслуживания, имеют также большое значение.

Разрабатывается и утверждается новая система оплаты труда, которая позволяет устранить возникающие противоречия. Отдел сервисного обслуживания гаражного оборудования должен разработать концепцию развития сервисной сети с целью наиболее полного удовлетворения потребителя, а значит и завоевания преимуществ в конкурентной борьбе.

Среди проблем более низкого уровня значимости необходимо отметить место корпоративной культуры. Понимание и осознание себя, как части сплоченного коллектива - сложный процесс. Достижение синергетического эффекта возможно только в коллективе, в котором отдельный сотрудник понимает и делает свою работу через понимание целей и задач всей Компании. Формированию корпоративной культуры следует уделить особое внимание.

Проведенный анализ дает возможность Компании сосредоточить свои усилия на наиболее важных на данный момент обозначенных проблемах. Выбор пути решения каждой из них определяется возможностями Компании и опытом руководителей.

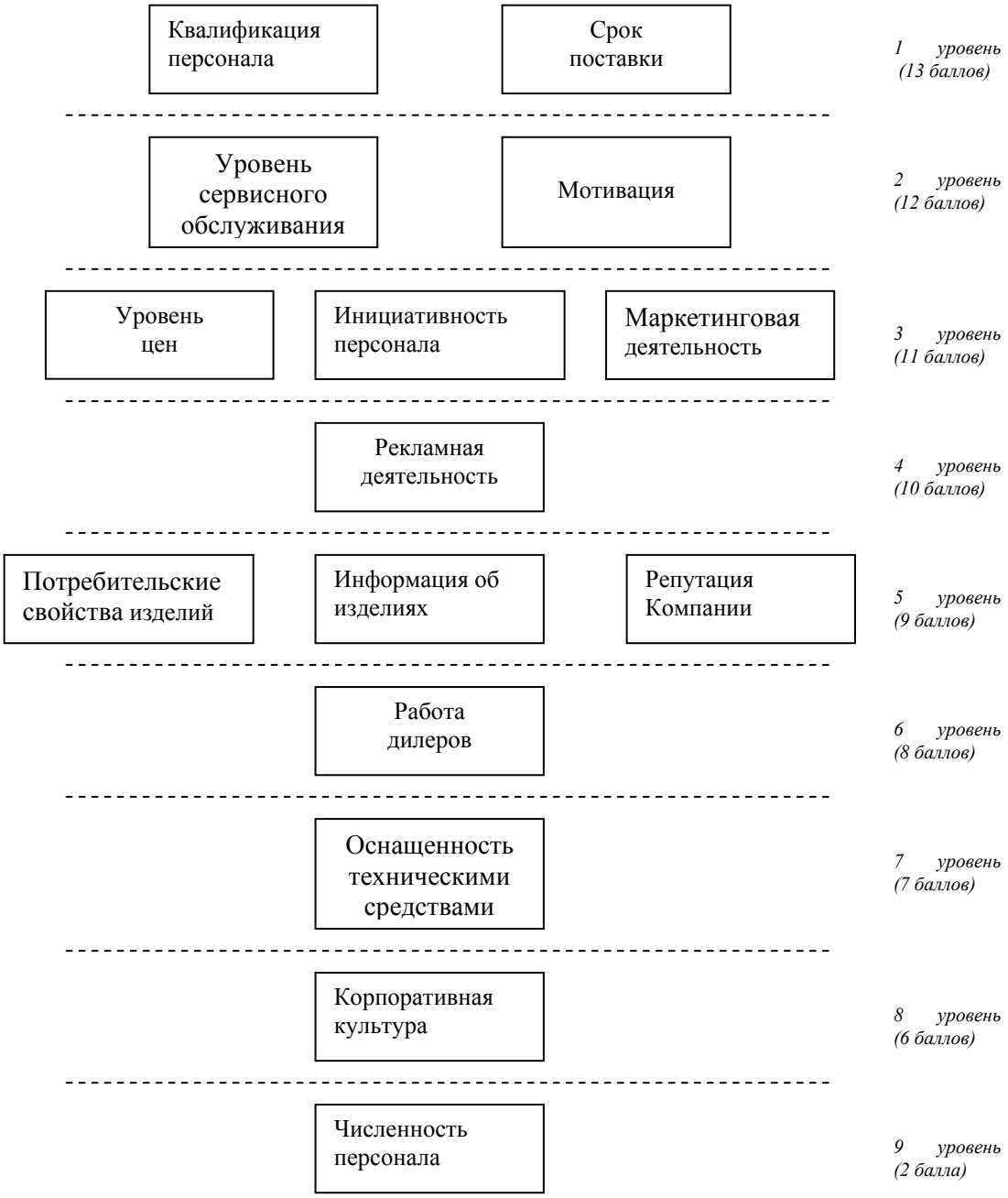



Рис.1. Распределение факторов по их значимости.

**Вероятностное моделирование парных сравнений.** Напомним общую модель парных сравнений, введенную в главе 2.1.4.

Пусть $t$ объектов $A_1, A_2, ..., A_t$ сравниваются попарно каждым из $n$ экспертов. Следовательно, возможных пар для сравнения имеется $s = t(t-1)/2$. Эксперт с номером $\gamma$ делает $r_\gamma$ повторных сравнений для каждой из $s$ возможностей. Пусть $X(i, j, \gamma, \delta)$, $i, j$=1, 2, ..., $t$, $i \neq j$, $\gamma$=1,2,...,$n$; $\delta$=1,2,..., $r_\gamma$, - случайная величина, принимающая значения 1 или 0 в зависимости от того, предпочитает ли эксперт $\gamma$ объект $A_i$ или объект $A_j$ в $\delta$-м сравнении двух объектов. Обычно принимают, что все сравнения проводятся независимо друг от друга, так что случайные величины $X(i, j, \gamma, \delta)$ независимы в совокупности, если не считать того, что $X(i, j, \gamma, \delta) + X(j, i, \gamma, \delta) = 1$. Положим

$$P(X(i, j, \gamma, \delta) = 1) = \pi(i, j, \gamma, \delta).$$

Ясно, что описанная модель парных сравнений представляет собой частный случай люсиана (в другой терминологии -бернуллиевского вектора). В этой модели число наблюдений равно числу неизвестных параметров, поэтому для получения статистических выводов необходимо наложить те или иные априорные условия на $\pi(i, j, \gamma, \delta)$, например:

$\pi(i, j, \gamma, \delta) = \pi(i, j, \gamma)$ (нет эффекта от повторений);

$\pi(i, j, \gamma, \delta) = \pi(i, j)$ (нет эффекта от повторений и от экспертов).

Теорию независимых парных сравнений целесообразно разделить на две части - непараметрическую, в которой статистические задачи ставятся непосредственно в терминах $\pi(i, j, \gamma, \delta)$, и параметрическую, в которой вероятности $\pi(i, j, \gamma, \delta)$ выражаются через меньшее число иных параметров. Ряд результатов непараметрической теории парных сравнений непосредственно вытекает из теории люсианов.

В параметрической теории парных сравнений наиболее популярна линейная модель, в которой предполагается, что каждому объекту $A_i$ можно сопоставить некоторую "ценность" $V_i$ так, что вероятность предпочтения $\pi(i, j)$ (т.е. предполагается дополнительно, что эффект от повторений и от экспертов отсутствует) выражается следующим образом:

$$\pi(i, j) = H(V_i - V_j), \qquad (1)$$

где $H(x)$ - функция распределения, симметричная относительно 0, т.е.

$$H(-x) = 1 - H(x) \qquad (2)$$

при всех $x$.

Широко применяются модели Терстоуна - Мостеллера и Брэдли - Терри, в которых $H(x)$ - соответственно функции нормального и логистического распределений. С прикладной точки зрения эти две модели практически совпадают. Действительно, поскольку функция $\Phi(x)$ стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1 и функция

$$\Psi(x) = e^x (1 + e^x)^{-1}$$

стандартного логистического распределения удовлетворяют соотношению (см. главу 2.1.4)

$$\sup_{x \in R^1} | \Phi(x) - \Psi(1{,}7x) | < 0{,}01,$$

то для обоснованного выбора по статистическим данным между моделями Терстоуна-Мостеллера и Брэдли-Терри необходимо не менее тысячи наблюдений. Ясно, что при реальном проведении экспертного опроса число наблюдений по крайней мере на порядок меньше.

Соотношение (1) вытекает из следующей модели поведения эксперта: он измеряет "ценность" $V_i$ и $V_j$ объектов $A_i$ и $A_j$, но с ошибками $\varepsilon_i$ и $\varepsilon_j$ соответственно, а затем сравнивает свои оценки ценности объектов $y_i = V_i + \varepsilon_i$ и $y_j = V_j + \varepsilon_j$. Если $y_i > y_j$, то он предпочитает $A_i$, в противном случае - $A_j$. Тогда

$$\pi(i, j) = P(\varepsilon_i - \varepsilon_j < V_i - V_j) = H(V_i - V_j). \qquad (3)$$

Обычно предполагают, что субъективные ошибки эксперта $\varepsilon_i$ и $\varepsilon_j$ независимы и имеют одно и то же непрерывное распределение. Тогда функция распределения $H(x)$ из соотношения (3) непрерывна и удовлетворяет функциональному уравнению (2).

*Пример.* При опросе экспертов (август 2001 г.) попарно сравнивались четыре компании ТНК, Лукойл, Юкос, Татнефть, продающие автомобильное топливо. Сравнение проводилось по качеству бензина. При $t = 4$ пар для сравнения имеется $s = t(t-1)/2 = 6$.

Результаты парных сравнений приведены в табл.1. По ним необходимо определить взаимное положение четырех компаний на оси «качество бензина», т.е. найти их «ценности» $V_1$, $V_2$, $V_3$, $V_4$.

Таблица 1.

Сравнение компаний по качеству бензина

| Пары | Частота выбора первого элемента пары | Частота выбора второго элемента пары |
|---|---|---|
| ТНК - Лукойл | p(1,2) = 0,508 | p(2,1) = 0,492 |
| ТНК - Юкос | p(1,3) = 0,331 | p(3,1) = 0,669 |
| ТНК - Татнефть | p(1,4) = 0,990 | p(4,1) = 0,010 |
| Лукойл - Юкос | p(2,3) = 0,338 | p(3,2) = 0,662 |
| Лукойл - Татнефть | p(2,4) = 0,990 | p(4,2) = 0,010 |
| Юкос - Татнефть | p(3,4) = 0,997 | p(4,3) = 0,003 |

Применим модель Терстоуна-Мостеллера, согласно которой погрешности мнений экспертов $\varepsilon_i$ являются независимыми нормально распределенными случайными величинами с нулевым математическим ожиданием и дисперсией $y^2$.

Легко видеть, что «ценности» $V_1$, $V_2$, $V_3$, $V_4$ измерены в шкале интервалов. Начало координат можно выбрать произвольно, поскольку вероятности результатов сравнения зависят только от попарных разностей «ценностей» $V_1$, $V_2$, $V_3$, $V_4$. Например, можно положить $V_4 = 0$. Единицу измерения также можно выбрать произвольно. При изменении единицы измерения меняется $y^2$, точнее, единица измерения однозначно связана с величиной y. Дисперсия разности $\varepsilon_i$ - $\varepsilon_j$ равна $2y^2$. В соответствии с формулой (3) удобно выбрать единицу измерения так, чтобы $2y^2 = 1$, т.е. $\sigma = 1/\sqrt{2}$. Тогда H в формуле (3) - это функция Ф стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1.

В соответствии с (3) имеем систему шести уравнений с тремя неизвестными:

Ц ($V_1$ - $V_2$) = p (1, 2) = 0,508,
Ц ($V_1$ - $V_3$) = p (1, 3) = 0,331,
Ц ($V_1$) = p (1, 4) = 0,990,
Ц ($V_2$ - $V_3$) = p (2, 3) = 0,338,
Ц ($V_2$) = p (2, 4) = 0,990,
Ц ($V_3$) = p (3, 4) = 0,997.

Применяя к каждому их этих уравнений преобразование $\Phi^{-1}$, получаем систему шести линейных уравнений с тремя неизвестными:

$V_1 - V_2 = a_1 = \Phi^{-1}(0{,}508) = 0{,}020054$,
$V_1 - V_3 = a_2 = \Phi^{-1}(0{,}331) = -0{,}437154$,
$V_1 = a_3 = \Phi^{-1}(0{,}990) = 2{,}326348$,
$V_2 - V_3 = a_4 = \Phi^{-1}(0{,}338) = -0{,}417928$,
$V_2 = a_5 = \Phi^{-1}(0{,}990) = 2{,}326348$,
$V_3 = a_6 = \Phi^{-1}(0{,}997) = 2{,}747781$.

(Значения $Ц^{-1}$ взяты из таблицы 1.3 сборника [16].)

В полученной системе число уравнений больше числа неизвестных, т.е. система переопределена. Дальнейшие расчеты могут проводиться разными способами. Простейший из них стоят в том, чтобы выбрать три уравнения, а именно, третье, пятое и шестое, которые и дают искомые значения:

$$V_1 = V_2 = 2{,}326348,\ V_3 = 2{,}747781.$$

Таким образом, качество бензина лучше всего у Юкоса, оно несколько хуже у ТНК и Лукойла, одинаковых по этому показателю, а Татнефть значительно хуже тройки лидеров. Можно показать, что если модель Терстоуна-Мостеллера верна и число экспертов достаточно велико, то отбрасывание «лишних» уравнений является корректным способом обработки экспертных данных, поскольку дает состоятельные оценки «ценностей» $V_1$, $V_2$, ..., $V_n$.

Однако ясно, что при отбрасывании трех уравнений из шести часть информации теряется. Например, первое уравнение показывает, что по мнению экспертов качество бензина у ТНК несколько лучше, у Лукойла. Поэтому целесообразно применить метод наименьших квадратов для оценивания $V_1$, $V_2$, $V_3$, $V_4$. А именно, рассмотрим функцию трех переменных

$$f(V_1, V_2, V_3) = (V_1 - V_2 - a_1)^2 + (V_1 - V_3 - a_2)^2 + (V_1 - a_3)^2 + \\ +(V_2 - V_3 - a_4)^2 +(V_2 - a_5)^2 +(V_3 - a_6)^2 .$$

Оценки по методу наименьших квадратов - это результат минимизации функции $f(V_1, V_2, V_3)$ по совокупности переменных $V_1$, $V_2$, $V_3$. Как и в главе 3.2, для минимизации этой функции достаточно приравнять 0 частные производные этой функции по $V_1$, $V_2$, $V_3$. Имеем:

$$\frac{\partial f}{\partial V_1} = 2(V_1 - V_2 - a_1) + 2(V_1 - V_3 - a_2) + 2(V_1 - a_3),$$

$$\frac{\partial f}{\partial V_2} = -2(V_1 - V_2 - a_1) + 2(V_2 - V_3 - a_4) + 2(V_2 - a_5),$$

$$\frac{\partial f}{\partial V_3} = -2(V_1 - V_3 - a_2) - 2(V_2 - V_3 - a_4) + 2(V_3 - a_6).$$

Приравнивая частные производные 0, деля на 2, раскрывая скобки и перенося свободные члены в правую часть, получаем систему трех линейных уравнений с тремя неизвестными

$$3V_1 - V_2 - V_3 = a_1 + a_2 + a_3,$$
$$-V_1 + 3V_2 - V_3 = -a_1 + a_4 + a_5,$$
$$-V_1 - V_2 + 3V_3 = -a_2 - a_4 + a_6.$$

Решение этой системы не представляет трудностей.

Вообще говоря, не всегда сравниваемые объекты можно представить точками на прямой, т.е. не всегда их можно линейно упорядочить. Возможно, более соответствует данным опроса экспертов представление объектов точками на плоскости или в пространстве большей размерности. В статистике парных сравнений [32] разработаны методы проверки адекватности модели Терстоуна-Мостеллера и других параметрических моделей. Для этого обычно используются статистики типа хи-квадрат.

### 3.4.5. Статистика нечетких множеств

Нечеткие множества – частный вид объектов нечисловой природы. Поэтому при обработке выборки, элементами которой являются нечеткие множества, могут быть использованы различные методы анализа статистических данных произвольной природы - расчет средних, непараметрических оценок плотности, построение диагностических правил и т.д.

**Среднее значение нечеткого множества.** Однако иногда используются методы, учитывающие специфику нечетких множеств. Например, пусть носителем нечеткого множества является конечная совокупность действительных чисел $\{x_1, x_2, ..., x_n\}$. Тогда под средним значением нечеткого множества иногда понимают число. А именно, среднее значение нечеткого множества определяют по формуле:

$$M(A) = \frac{\sum_{i=1}^{n} x_i \mu_A(x_i)}{\sum_{i=1}^{n} \mu_A(x_i)},$$

где $\mu_A(x_i)$ - функция принадлежности нечеткого множества $A$. Если знаменатель равен 1, то эта формула определяет математическое ожидание случайной величины, для которой вероятность попасть в точку $x_i$ равна $\mu_A(x_i)$. Такое определение наиболее естественно, когда нечеткое множество $A$ интерпретируется как нечеткое число.

Очевидно, наряду с $M(A)$ может оказаться полезным использование эмпирических средних, определяемых (согласно статистике в пространствах общей природы) путем решения соответствующих оптимизационных задач. Для конкретных расчетов необходимо ввести то или иное расстояние между нечеткими множествами.

**Расстояния в пространствах нечетких множеств.** Как известно, многие методы статистики нечисловых данных базируются на использовании расстояний (или показателей различия) в соответствующих пространствах нечисловой природы. Расстояние между нечеткими подмножествами $A$ и $B$ множества $X = \{x_1, x_2, \ldots, x_k\}$ можно определить как

$$d(A,B) = \sum_{j=1}^{k} | \mu_A(x_j) - \mu_B(x_j) |,$$

где $м_A(x_j)$ - функция принадлежности нечеткого множества $A$, а $м_B(x_i)$ - функция принадлежности нечеткого множества $B$. Может использоваться и другое расстояние:

$$D(A,B) = \frac{\sum_{j=1}^{k} | \mu_A(x_j) - \mu_B(x_j) |}{\sum_{j=1}^{k} (\mu_A(x_j) + \mu_B(x_j))}.$$

(Примем это расстояние равным 0, если функции принадлежности тождественно равны 0.)

В соответствии с аксиоматическим подходом к выбору расстояний (метрик) в пространствах нечисловой природы разработан обширный набор систем аксиом, из которых выводится тот или иной вид расстояний (метрик) в конкретных пространствах, в том числе в пространствах нечетких множеств (см. главу 1.1). При использовании вероятностных моделей расстояние между случайными нечеткими множествами (т.е. между случайными элементами со значениями в пространстве нечетких множеств) само является случайной величиной, имеющей в ряде постановок асимптотически нормальное распределение [28].

**Проверка гипотез о нечетких множествах.** Пусть ответ эксперта – нечеткое множество. Естественно считать, что его ответ, как показание любого средства измерения, содержит погрешности. Если есть несколько экспертов, то в качестве единой оценки (группового мнения) естественно взять эмпирическое среднее их ответов. Но возникает естественный вопрос: действительно ли все эксперты измеряют одно и то же? Может быть, глядя на реальный объект, они оценивают его с разных сторон? Например, на научную статью можно смотреть как с теоретической точки зрения, как и с прикладной, и соответствующие оценки будут скорее всего различны (если они совпадают, то работа либо никуда не годится, либо является выдающейся).

Итак, возник вопрос: как проверить согласованность мнений экспертов? Надо сначала определить понятие согласованности. Пусть $A$ – нечеткий ответ эксперта. Будем считать, что соответствующая функция принадлежности есть сумма двух слагаемых:

$$\mu_A(u) = \mu_{N(A)}(u) + \xi_A(u),$$

где $N(A)$ – «истинное» нечеткое множество, а $о_A(u)$ – «погрешность» эксперта как прибора. Естественно рассмотреть две постановки.

Мнения экспертов $A(1)$, $A(2)$, …, $A(m)$ будем считать согласованными, если

$$N(A(1)) = N(A(2)) = \ldots, N(A(m)).$$

Рассмотрим две группы экспертов. В первой у всех «истинное» мнение $N(A)$, а во второй у всех - $N(B)$. Две группы будем считать согласованными по мнениям, если

$$N(A) = N(B).$$

Согласованность определена. Как же ее проверить? Если экспертов достаточно много, то эти гипотезы можно проверять отдельно для каждого элемента множества – общего носителя нечетких ответов. Проверка последней гипотезы переходит в проверку однородности двух независимых выборок (глава 3.1). Здесь ограничимся постановками основных гипотез (ср. с аналогичными гипотезами, рассмотренными выше для люсианов).

**Восстановление зависимости между нечеткими переменными.** Рассмотрим две нечеткие переменные $A$ и $B$. Пусть каждый из $n$ испытуемых выдает в ответ на вопрос два нечетких множества $A_i$ и $B_i$, $i = 1, 2, \ldots, n$. Необходимо восстановить зависимость $B$ от $A$, другими словами, наилучшим образом приблизить $B$ с помощью $A$.

Для иллюстрации основной идеи ограничимся парной линейной регрессией нечетких множеств. Нечеткое множество $C$ назовем линейной функцией от нечеткого множества $A$, если для любого $x$ из носителя $A$ функции принадлежности множеств $A$ и $C$ таковы, что $\mu_C(x) = \mu_A(y)$ при $x = бy + в$. Другими словами,

$$\mu_C(x) = \mu_A((x - в)/б)$$

для любого $x$ из носителя $A$. В таком случае естественно писать

$$C = \text{б}A + \text{в}.$$

Однако нечеткие переменные, как и привычные статистикам числовые переменные, обычно несколько отклоняются от линейной связи. Наилучшее линейное приближение нечеткой переменной $B$ с помощью линейной функции от нечеткой переменной $A$ естественно искать, решая задачу минимизации по б, в расстояния от $B$ до $C$. Пусть

$$\text{с}(B, \text{б}_0\text{А} + \text{в}_0) = \min \text{с}(B, \text{б}A + \text{в}),$$

где с – некоторое расстояние между нечеткими множествами, а минимизация проводится по всем возможным значениям б и в. Тогда наилучшей линейной аппроксимацией $B$ является $\text{б}_0\text{А} + \text{в}_0$. Если рассматриваемый минимум равен 0, то имеет место точная линейная зависимость.

Для восстановления зависимости по выборочным парам нечетких переменных естественно воспользоваться подходом, развитым в статистике в пространствах произвольной природы для параметрической регрессии (аппроксимации). В соответствии с рассмотрениями главы 2.2.3 в качестве наилучших оценок параметров линейной зависимости следует рассматривать

$$(\alpha^*, \beta^*) = Arg \min_{\alpha,\beta} \sum_{k=1}^{n} \rho(B_i, \alpha A_i + \beta).$$

Тогда наилучшим линейным приближением $B$ является $C^* = \text{б}^*A + \text{в}^*$.

Вероятностно-статистическая теория регрессионного анализа нечетких переменных строится как частный случай аналогичной теории для переменных произвольной природы (глава 2.2.3). В частности, при обычных предположениях оценки б*, в* являются состоятельными, т.е. $\text{б}^* \to \text{б}_0$ и $\text{в}^* \to \text{в}_0$ при $n \to \infty$.

**Кластер-анализ нечетких переменных**. Строить группы сходных между собой нечетких переменных (кластеры) можно многими способами. Опишем два семейства алгоритмов.

Пусть на пространстве, в котором лежат результаты наблюдений, т.е. на пространстве нечетких множеств, заданы две меры близости с и ф (например, это могут быть введенные выше расстояния $d$ и $D$). Берется один из результатов наблюдений (нечеткое множество) и вокруг него описывается шар радиуса $R$, определяемый мерой близости с. (Напомним, что шаром с центром в $x$ относительно с называется множество всех элементов $y$ рассматриваемого пространства таких, что $\text{с}(x, y) \le R$.) Берутся результаты наблюдений (элементы выборки), попавшие в этот шар, и находится их эмпирическое среднее относительно второй меры близости ф. Оно берется за новый центр, вокруг которого снова описывается шар радиуса $R$ относительно с, и процедура повторяется. (Чтобы алгоритм был полностью определен, необходимо сформулировать правило выбора элемента эмпирического среднего в качестве нового центра, если эмпирическое среднее состоит более чем из одного элемента.)

Когда центр шара зафиксируется (перестанет меняться), попавшие в этот шар элементы объявляются первым кластером и исключаются из дальнейшего рассмотрения. Алгоритм применяется к совокупности оставшихся результатов наблюдений, выделяет из нее второй кластер и т.д.

Всегда ли центр шара остановится? При реальных расчетах в течение многих лет так было всегда. Соответствующая теория была построена в 1977 г. [33]. Было доказано, что описанный выше процесс всегда остановится через конечное число шагов. Причем число шагов до остановки оценивается через максимально возможное число результатов наблюдений в шаре радиуса $R$ относительно с.

Обширное семейство образуют алгоритмы кластер-анализа типа «Дендрограмма», известные также под названием «агломеративные иерархические алгоритмы средней связи». На первом шагу алгоритма из этого семейства каждый результат наблюдения рассматривается как отдельный кластер. Далее на каждом шагу происходит объединение двух самых близких кластеров. Название «Дендрограмма» объясняется тем, что результат работы алгоритма обычно представляется в виде дерева. Каждая его ветвь соответствует кластеру, появляющемуся на каком-либо шагу работы алгоритма. Слияние ветвей соответствует объединению кластеров, а ствол – заключительному шагу, когда все наблюдения оказываются объединенными в один кластер.

Для работы алгоритмов кластер-анализа типа «Дендрограмма» необходимо определить расстояние между кластерами. Естественно использовать ассоциативные средние, которыми, как известно, являются обобщенные средние по Колмогорову всевозможных попарных расстояний

между элементами двух рассматриваемых кластеров. Итак, расстояние между кластерами $K$ и $L$, состоящими из $n_1$ и $n_2$ элементов соответственно, определяется по формуле:

$$\tau(K,L) = F^{-1}\left(\frac{1}{n_1 n_2}\sum_{i\in K}\sum_{j\in L} F(\rho(X_i, X_j))\right),$$

где с – некоторое расстояние между нечеткими множествами, F – строго монотонная функция (строго возрастающая или строго убывающая).

Соображение теории измерений позволяют ограничить круг возможных алгоритмов типа «Дендрограмма». Естественно принять, что единица измерения расстояния выбрана произвольно. Тогда согласно результатам главы 2.1.3 из всех обобщенных средних по Колмогорову годятся только степенные средние, т.е. $F(z) = z^{л}$ при л ≠ 0 или $F(z) = \ln z$. Чтобы получить разбиение на кластеры, надо «разрезать» дерево на определенной высоте, т.е. объединять кластеры лишь до тех пор, пока расстояние между ними меньше заранее выбранной константы. При альтернативном подходе заранее фиксируется число кластеров. Рассматривают и двухкритериальную постановку, когда минимизируют сумму (или максимум) внутрикластерных разбросов и число кластеров. Для решения задачи двухкритериальной минимизации либо один из критериев заменяют на ограничение, либо два критерия «свертывают» в один, либо применяют иные подходы (последовательная оптимизация, построение поверхности Парето и др.).

При классификации нечетких множеств полезны многие подходы, рассмотренные в главе 3.2, а именно, все подходы, основанные только на использовании расстояний.

**Сбор и описание нечетких данных.** Разработано большое количество процедур описания нечеткости. Так, согласно Э.Борелю понятие «Куча» описывается с помощью функции распределения – при каждом конкретном $x$ значение функции принадлежности – это доля людей, считающих совокупность $x$ зерен кучей. Результат подобного опроса может дать и кривую иного вида, например, по поводу понятия «молодой» (слева будут отделены «дети», а справа – «люди зрелого и пожилого возраста»). Нечеткая толерантность может оцениваться с помощью случайных толерантностей (см. выше).

Целесообразно попытаться выделить наиболее практически полезные простые формы функций принадлежности. Видимо, наиболее простой является «ступенька» - внутри некоторого интервала функция принадлежности равна 1, а вне этого интервала равна 0. Это – простейший способ «размывания» числа путем замены его интервалом. Нечеткое множество описывается двумя числами – концами интервала. Оценки этих чисел можно получить с помощью экспертов. Статистическая теория подобных нечетких множеств рассмотрена в главе 3.5.

Тремя числами $a < b < c$ описывается функция принадлежности типа треугольника. При этом левее $a$ и правее $c$ функция принадлежности равна 0. В точке $b$ функция принадлежности принимает значение 1. На отрезке $[a; b]$ функция принадлежности линейно растет от 0 до 1, а на отрезке $[b;c]$ – линейно убывает от 1 до 0. Оценки трех чисел $a < b < c$ получают при опросе экспертов.

Следующий по сложности вид функции принадлежности – типа трапеции – описывается четырьмя числами $a < b < c < d$. Левее $a$ и правее $d$ функция принадлежности равна 0. На отрезке $[a; b]$ она линейно возрастает от 0 до 1, на отрезке $[b; c]$ во всех точках равна 1, а на отрезке $[c; d]$ линейно убывает от 1 до 0. Для оценивания четверки чисел $a < b < c < d$ используют экспертов.

Ряд результатов статистики нечетких данных приведен в первой монографии российского автора по нечетким множествам [34] и во многих дальнейших публикациях.

### 3.4.6. Статистика нечисловых данных в экспертных оценках

Развитие статистики нечисловых данных во многом стимулировалось запросами теории и практики экспертных оценок. Рассмотрим взаимоотношение этих двух областей подробнее.

**Современная теория измерений и экспертные оценки.** Как проводить анализ собранных рабочей группой ответов экспертов? Для более углубленного рассмотрения проблем экспертных оценок понадобятся некоторые понятия *репрезентативной теории измерений*, служащей основой теории экспертных оценок, прежде всего той ее части, которая связана с анализом заключений экспертов, выраженных в качественном (а не в количественном) виде.

Как уже отмечалось, получаемые от экспертов мнения часто выражены в *порядковой шкале*. Другими словами, эксперт может сказать (и обосновать), что один тип продукции будет

более привлекателен для потребителей, чем другой, один показатель качества продукции более важен, чем другой, первый технологический объект более опасен, чем второй, и т.д. Но эксперт не в состоянии обосновать, *во сколько раз* или *на сколько* более важен, соответственно, более опасен. Поэтому экспертов часто просят дать ранжировку (упорядочение) объектов экспертизы, т.е. расположить их в порядке возрастания (или, точнее, неубывания) интенсивности интересующей организаторов экспертизы характеристики.

Рассмотрим в качестве примера применения результатов теории измерений, связанных со средними величинами в порядковой шкале, один сюжет, связанный с ранжировками и рейтингами.

**Методы средних баллов.** В настоящее время распространены экспертные, маркетинговые, квалиметрические, социологические и иные опросы, в которых опрашиваемых просят выставить баллы объектам, изделиям, технологическим процессам, предприятиям, проектам, заявкам на выполнение научно-исследовательских работ, идеям, проблемам, программам, политикам и т.п. Затем рассчитывают средние баллы и рассматривают их *как интегральные (т.е. обобщенные, итоговые) оценки,* выставленные коллективом опрошенных экспертов. Какими формулами пользоваться для вычисления средних величин? Ведь существует очень много разных видов средних величин.

По традиции обычно применяют *среднее арифметическое*. Однако специалисты по теории измерений уже около 30 лет знают, что *такой способ некорректен*, поскольку баллы обычно измерены в *порядковой* шкале (см. главу 2.1.3). Обоснованным является использование медиан в качестве средних баллов. Однако полностью игнорировать средние арифметические нецелесообразно из-за их привычности и распространенности. Поэтому *представляется рациональным использовать одновременно оба метода - и метод средних арифметических рангов (баллов), и методов медианных рангов. Т*акая рекомендация находится в согласии с общенаучной *концепцией устойчивости* (глава 1.4.7), рекомендующей применять различные методы для обработки одних и тех же данных с целью выделить выводы, получаемые одновременно при всех методах. Такие выводы, видимо, соответствуют реальной действительности, в то время как заключения, меняющиеся от метода к методу, зависят от субъективизма исследователя, выбирающего метод обработки исходных экспертных оценок.

**Пример сравнения восьми проектов.** Рассмотрим конкретный пример применения только что сформулированного подхода.

По заданию руководства фирмы анализировались восемь проектов, предлагаемых для включения в план стратегического развития фирмы. Они обозначены следующим образом: Д, Л, М-К, Б, Г-Б, Сол, Стеф, К (по фамилиям менеджеров, предложивших их для рассмотрения). Все проекты были направлены 12 экспертам, включенным в экспертную комиссию, организованную по решению Правления фирмы. В табл.1 приведены ранги восьми проектов, присвоенные им каждым из 12 экспертов в соответствии с представлением экспертов о целесообразности включения проекта в стратегический план фирмы. При этом эксперт присваивает ранг 1 самому лучшему проекту, который обязательно надо реализовать. Ранг 2 получает от эксперта второй по привлекательности проект, ... , наконец, ранг 8 - наиболее сомнительный проект, который реализовывать стоит лишь в последнюю очередь.

Таблица 1. Ранги 8 проектов по степени привлекательности для включения в план стратегического развития фирмы

| № эксперта | Д | Л | М-К | Б | Г-Б | Сол | Стеф | К |
|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 3 | 1 | 2 | 8 | 4 | 6 | 7 |
| 2 | 5 | 4 | 3 | 1 | 8 | 2 | 6 | 7 |
| 3 | 1 | 7 | 5 | 4 | 8 | 2 | 3 | 6 |
| 4 | 6 | 4 | 2,5 | 2,5 | 8 | 1 | 7 | 5 |
| 5 | 8 | 2 | 4 | 6 | 3 | 5 | 1 | 7 |
| 6 | 5 | 6 | 4 | 3 | 2 | 1 | 7 | 8 |
| 7 | 6 | 1 | 2 | 3 | 5 | 4 | 8 | 7 |
| 8 | 5 | 1 | 3 | 2 | 7 | 4 | 6 | 8 |
| 9 | 6 | 1 | 3 | 2 | 5 | 4 | 7 | 8 |
| 10 | 5 | 3 | 2 | 1 | 8 | 4 | 6 | 7 |
| 11 | 7 | 1 | 3 | 2 | 6 | 4 | 5 | 8 |
| 12 | 1 | 6 | 5 | 3 | 8 | 4 | 2 | 7 |

*Примечание.* Эксперт № 4 считает, что проекты М-К и Б равноценны, но уступают лишь одному проекту - проекту Сол. Поэтому проекты М-К и Б должны были бы стоять на втором и третьем местах и получить баллы 2 и 3. Поскольку они равноценны, то получают средний балл (2+3)/ 2 = 5/ 2 = 2,5.

Анализируя результаты работы экспертов (т.е. упомянутую таблицу), члены аналитической подразделения Рабочей группы, анализировавшие ответы экспертов по заданию Правления фирмы, были вынуждены констатировать, что полного согласия между экспертами нет, а потому данные, приведенные в таблице, следует подвергнуть тщательному математическому анализу.

**Метод средних арифметических рангов.** Сначала для получения группового мнения экспертов был применен метод средних арифметических рангов. Прежде всего была подсчитана сумма рангов, присвоенных проектам (см. табл. 1). Затем эта сумма была разделена на число экспертов, в результате рассчитан средний арифметический ранг (именно эта операция дала название методу). По средним рангам строится итоговая ранжировка (в другой терминологии - упорядочение), исходя из принципа - чем меньше средний ранг, чем лучше проект. Наименьший средний ранг, равный 2,625, у проекта Б, - следовательно, в итоговой ранжировке он получает ранг 1. Следующая по величине сумма, равная 3,125, у проекта М-К, - и он получает итоговый ранг 2. Проекты Л и Сол имеют одинаковые суммы (равные 3,25), значит, с точки зрения экспертов они равноценны (при рассматриваемом способе сведения вместе мнений экспертов), а потому они должны бы стоять на 3 и 4 местах и получают средний балл (3+4)/2 = 3,5. Дальнейшие результаты приведены в табл. 2.

Итак, ранжировка по суммам рангов (или, что в данном случае то же самое, по средним арифметическим рангам) имеет вид:

$$Б < М\text{-}К < \{Л, Сол\} < Д < Стеф < Г\text{-}Б < К . \qquad (1)$$

Здесь запись типа "А<Б" означает, что проект А предшествует проекту Б (т.е. проект А лучше проекта Б). Поскольку проекты Л и Сол получили одинаковую сумму баллов, то по рассматриваемому методу они эквивалентны, а потому объединены в группу (в фигурных скобках). В терминологии математической статистики ранжировка (1) имеет одну связь.

**Метод медиан рангов.** Значит, наука сказала свое слово, итог расчетов - ранжировка (1), и на ее основе предстоит принимать решение? Так был поставлен вопрос при обсуждении полученных результатов на заседании Правления фирмы. Но тут наиболее знакомый с современной эконометрикой член Правления вспомнил, что ответы экспертов измерены в порядковой шкале, а потому для них неправомерно проводить усреднение методом средних арифметических. Надо использовать метод медиан.

Что это значит? Надо взять ответы экспертов, соответствующие одному из проектов, например, проекту Д. Это ранги 5, 5, 1, 6, 8, 5, 6, 5, 6, 5, 7, 1. Затем их надо расположить в порядке неубывания (проще было бы сказать – «в порядке возрастания», но поскольку некоторые ответы совпадают, то приходится использовать непривычный термин «неубывание»). Получим последовательность: 1, 1, 5, 5, 5, 5, 5, 6, 6, 6, 7, 8. На центральных местах - шестом и седьмом - стоят 5 и 5. Следовательно, медиана равна 5.

Таблица 2.
Результаты расчетов по методу средних арифметических и методу медиан для данных, приведенных в таблице 1.

| | Д | Л | М-К | Б | Г-Б | Сол | Стеф | К |
|---|---|---|---|---|---|---|---|---|
| Сумма рангов | 60 | 39 | 37,5 | 31.5 | 76 | 39 | 64 | 85 |
| Среднее арифметическое рангов | 5 | 3,25 | 3,125 | 2,625 | 6,333 | 3,25 | 5,333 | 7,083 |
| Итоговый ранг по среднему арифметическому | 5 | 3,5 | 2 | 1 | 7 | 3,5 | 6 | 8 |
| Медианы рангов | 5 | 3 | 3 | 2,25 | 7,5 | 4 | 6 | 7 |
| Итоговый ранг по медианам | 5 | 2,5 | 2,5 | 1 | 8 | 4 | 6 | 7 |

Медианы совокупностей из 12 рангов, соответствующих определенным проектам, приведены в предпоследней строке табл.2. (При этом медианы вычислены по обычным правилам статистики - как среднее арифметическое центральных членов вариационного ряда.) Итоговое

упорядочение комиссии экспертов по методу медиан приведено в последней строке таблицы. Ранжировка (т.е. упорядочение - итоговое мнение комиссии экспертов) по медианам имеет вид:

$$Б < \{М\text{-}К, Л\} < Сол < Д < Стеф < К < Г\text{-}Б \ . \qquad (2)$$

Поскольку проекты Л и М-К имеют одинаковые медианы баллов, то по рассматриваемому методу ранжирования они эквивалентны, а потому объединены в группу (кластер), т.е. с точки зрения математической статистики ранжировка (4) имеет одну связь.

**Сравнение ранжировок по методу средних арифметических и методу медиан.** Сравнение ранжировок (1) и (2) показывает их близость (похожесть). Можно принять, что проекты М-К, Л, Сол упорядочены как М-К < Л < Сол, но из-за погрешностей экспертных оценок в одном методе признаны равноценными проекты Л и Сол (ранжировка (1)), а в другом - проекты М-К и Л (ранжировка (2)). Существенным является только расхождение, касающееся упорядочения проектов К и Г-Б: в ранжировке (3) Г-Б < К, а в ранжировке (4), наоборот, К < Г-Б. Однако эти проекты - наименее привлекательные из восьми рассматриваемых, и при выборе наиболее привлекательных проектов для дальнейшего обсуждения и использования на указанное расхождение можно не обращать внимания.

Рассмотренный пример демонстрирует сходство и различие ранжировок, полученных по методу средних арифметических рангов и по методу медиан, а также пользу от их совместного применения.

**Метод согласования кластеризованных ранжировок.** Проблема состоит в выделении общего нестрогого порядка из набора кластеризованных ранжировок (в другой терминологии - ранжировок со связями). Этот набор может отражать мнения нескольких экспертов или быть получен при обработке мнений экспертов различными методами. Рассмотрим *метод согласования кластеризованных ранжировок, позволяющий «загнать» противоречия внутрь специальным образом построенных кластеров (групп), в то время как упорядочение кластеров соответствует одновременно всем исходным упорядочениям.*

В различных прикладных областях возникает необходимость анализа нескольких кластеризованных ранжировок объектов. К таким областям относятся прежде всего инженерный бизнес, менеджмент, экономика, социология, экология, прогнозирование, научные и технические исследования и т.д., особенно те их разделы, что связаны с экспертными оценками (см., например, [8, 35]). В качестве объектов могут выступать образцы продукции, технологии, математические модели, проекты, кандидаты на должность и др. Кластеризованные ранжировки могут быть получены как с помощью экспертов, так и объективным путем, например, при сопоставлении математических моделей с экспериментальными данными с помощью того или иного критерия качества. Описанный ниже метод был разработан в связи с проблемами химической безопасности биосферы и экологического страхования [35].

В настоящем пункте рассматривается метод построения кластеризованной ранжировки, согласованной (в раскрытом ниже смысле) со всеми рассматриваемыми кластеризованными ранжировками. При этом противоречия между отдельными исходными ранжировками оказываются заключенными внутри кластеров согласованной ранжировки. В результате упорядоченность кластеров отражает общее мнение экспертов, точнее, то общее, что содержится в исходных ранжировках**.**

В кластеры заключены объекты, по поводу которых некоторые из исходных ранжировок *противоречат* друг другу. Для их упорядочения необходимо провести новые исследования. Эти исследования могут быть как формально-математическими (например, вычисление медианы Кемени, упорядочения по средним рангам или по медианам и т.п.), так и требовать привлечения новой информации из соответствующей прикладной области, возможно, проведения дополнительных научных или прикладных работ.

Введем необходимые понятия, затем сформулируем алгоритм согласования кластеризованных ранжировок в общем виде и рассмотрим его свойства.

Пусть имеется конечное число объектов, которые мы для простоты изложения будем изображать натуральными числами 1,2,3,...,$k$ и называть их совокупность «носителем». *Под кластеризованной ранжировкой, определенной на заданном носителе, понимаем следующую математическую конструкцию*. Пусть объекты разбиты на группы, которые будем называть кластерами. В кластере может быть и один элемент. Входящие в один кластер объекты будем заключать в фигурные скобки. Например, объекты 1,2,3,...,10 могут быть разбиты на 7 кластеров:

{1}, {2,3}, {4}, {5,6,7}, {8}, {9}, {10}. В этом разбиении один кластер {5,6,7} содержит три элемента, другой - {2,3} - два, остальные пять - по одному элементу. Кластеры не имеют общих элементов, а объединение их (как множеств) есть все рассматриваемое множество объектов (весь носитель).

Вторая составляющая кластеризованной ранжировки - это строгий линейный порядок между кластерами. Задано, какой из них первый, какой второй, и т.д. Будем изображать упорядоченность с помощью знака < . При этом кластеры, состоящие из одного элемента, будем для простоты изображать без фигурных скобок. Тогда кластеризованную ранжировку на основе введенных выше кластеров можно изобразить так:

$$A = [\ 1 < \{2,3\} < 4 < \{5,6,7\} < 8 < 9 < 10\ ] .$$

Конкретные кластеризованные ранжировки будем заключать в квадратные скобки. Если для простоты речи термин "кластер" применять только к кластеру не менее чем из 2-х элементов, то можно сказать, что в кластеризованную ранжировку $A$ входят два кластера {2,3} и {5,6,7} и 5 отдельных элементов.

Введенная описанным образом кластеризованная ранжировка является бинарным отношением на носителе - множестве {1,2,3,...,10}. Его структура такова. Задано отношение эквивалентности с 7-ю классами эквивалентности, а именно, {2,3}, {5,6,7}, а 5 классов остальные состоят из оставшихся 5 отдельных элементов. Затем введен строгий линейный порядок между классами эквивалентности.

Введенный математический объект известен в литературе как "ранжировка со связями" (М. Холлендер, Д. Вулф), "упорядочение" (Дж. Кемени, Дж. Снелл [10]), "квазисерия" (Б.Г. Миркин), "совершенный квазипорядок" (Ю.А. Шрейдер [36, с.127, 130]). Учитывая разнобой в терминологии, было признано полезным ввести собственный термин *"кластеризованная ранжировка",* поскольку в нем явным образом названы основные элементы изучаемого математического объекта - кластеры, рассматриваемые на этапе согласования ранжировок как классы эквивалентности, и ранжировка - строгий совершенный порядок между ними (в терминологии Ю.А.Шрейдера [36, гл.IV]).

Следующее важное понятие - *противоречивость*. Оно определяется для четверки - две кластеризованные ранжировки на одном и том же носителе и два различных объекта - элементы того же носителя. При этом два элемента из одного кластера будем связывать символом равенства = , как эквивалентные.

Пусть $A$ и $B$ - две кластеризованные ранжировки. *Пару объектов (a,b) назовем «противоречивой» относительно кластеризованных ранжировок A и B, если эти два элемента по-разному упорядочены в A и B, т.е. $a < b$ в A и $a > b$ в B (первый вариант противоречивости) либо $a > b$ в A и $a < b$ в B (второй вариант противоречивости).* Отметим, что в соответствии с этим определением пара объектов ($a$ ,$b$), эквивалентная хотя бы в одной кластеризованной ранжировке, не может быть противоречивой: эквивалентность $a = b$ не образует "противоречия" ни с $a < b$, ни с $a > b$. Это свойство оказывается полезным при выделении противоречивых пар.

В качестве примера рассмотрим, кроме $A$, еще две кластеризованные ранжировки

$$B = [\{1,2\} < \{\ 3,4,\ 5\} < 6 < 7 < 9 < \{8,\ 10\}],$$
$$C = [3 < \{1,\ 4\} < 2 < 6 < \{5,\ 7,\ 8\} < \{9,\ 10\}].$$

*Совокупность противоречивых пар объектов для двух кластеризованных ранжировок A и B назовем «ядром противоречий» и обозначим S(A,B).* Для рассмотренных выше в качестве примеров трех кластеризованных ранжировок $A$, $B$ и $C$, определенных на одном и том же носителе {1, 2, 3,..., 10}, имеем

$$S(A,B) = [(8,\ 9)],\ S(A,\mathrm{C}) = [(1,\ 3),\ (2,4)],$$
$$S(B,C) = [(1,\ 3),\ (2,\ 3),\ (2,\ 4),\ (5,\ 6),\ (8,9)].$$

Как при ручном, так и при программном нахождении ядра можно в поисках противоречивых пар просматривать пары (1,2), (1,3), (1,4), .... , (1,$k$), затем (2,3), (2,4), ..., (2,$k$), потом (3,4), ..., (3, $k$), и т.д., вплоть до последней пары ($k$-1, $k$).

Пользуясь понятиями дискретной математики, «ядро противоречий» можно изобразить *графом* с вершинами в точках носителя. При этом *противоречивые пары задают ребра этого графа.* Граф для $S(A,B)$ имеет только одно ребро (одна связная компонента более чем из одной точки). Граф для $S(A,C)$ - 2 ребра (две связные компоненты более чем из одной точки). Граф для $S(B,C)$ - 5 ребер (три связные компоненты более чем из одной точки, а именно, {1, 2 , 3, 4}, {5, 6} и {8, 9}).

Каждую кластеризованную ранжировку, как и любое бинарное отношение, можно задать матрицей $||x(a,b)||$ из 0 и 1 порядка $k$ x $k$. При этом $x(a,b) = 1$ тогда и только тогда, когда $a < b$ либо $a = b$. В первом случае $x(b,a) = 0$, а во втором $x(b,a) = 1$. При этом хотя бы одно из чисел $x(a,b)$ и $x(b,a)$ равно 1. Из определения противоречивости пары $(a, b)$ вытекает, что для нахождения всех таких пар достаточно поэлементно перемножить две матрицы $||x(a,b)||$ и $||y(a,b)||$, соответствующие двум кластеризованным ранжировкам, и отобрать те и только те пары, для которых $x(a,b)y(a,b) = x(b,a)y(b,a)=0$.

Алгоритм согласования некоторого числа (двух или более) кластеризованных ранжировок состоят из трех этапов. На первом *выделяются противоречивые пары* объектов во всех парах кластеризованных ранжировок. На втором формируются кластеры итоговой кластеризованной ранжировки (т.е. классы эквивалентности - *связные компоненты графов*, соответствующих объединению попарных ядер противоречий). На третьем этапе эти *кластеры* (*классы эквивалентности) упорядочиваются*. Для установления порядка между кластерами произвольно выбирается один объект из первого кластера и второй - из второго, порядок между кластерами устанавливается такой же, какой имеет быть между выбранными объектами в любой из рассматриваемых кластеризованных ранжировок. (Если в одной из исходных кластеризованных ранжировок имеет быть равенство, а в другой – неравенство, то при построении итоговой кластеризованной ранжировки используется неравенство.)

Корректность подобного упорядочивания, т.е. его независимость от выбора той или иной пары объектов, вытекает из соответствующих теорем, доказанных в работе [35].

Два объекта из разных кластеров согласующей кластеризованной ранжировки могут оказаться эквивалентными в одной из исходных кластеризованных ранжировок (т.е. находиться в одном кластере). В таком случае надо рассмотреть упорядоченность этих объектов в какой-либо другой из исходных кластеризованных ранжировок. Если же во всех исходных кластеризованных ранжировках два рассматриваемых объекта находились в одном кластере, то естественно считать (и это является уточнением к этапу 3 алгоритма), что они находятся в одном кластере и в согласующей кластеризованной ранжировке.

Результат согласования кластеризованных ранжировок $A$, $B$, $C$,... обозначим $f(A, B, C,...)$. Тогда

$$f(A, B) = [1<2<3<4<5<6<7<\{8, 9\}<10],$$
$$f(A, C) = [\{1,3\}<\{2, 4\}<6<\{5,7\}<8<9<10],$$
$$f(B, C) = [\{1,2,3,4\}<\{5,6\}<7<\{8,9\}<10],$$
$$f(A, B, C) = f(B, C) = [\{1,2,3,4\} <\{5,6\}<7<\{8, 9\}<10].$$

Итак, в случае $f(A, B)$ дополнительного изучения с целью упорядочения требуют только объекты 8 и 9. В случае $f$(A, $C$) кластер {5,7} появился не потому, что относительно объектов 5 и 7 имеется противоречие, а потому, что в обеих исходных ранжировках эти объекты не различаются. В случае $f(B, C)$ четыре объекта с номерами 1, 2, 3, 4 объединились в один кластер, т.е. кластеризованные ранжировки оказались настолько противоречивыми, что процедура согласования не позволила провести достаточно полную декомпозицию задачи нахождения итогового мнения экспертов.

Обсудим некоторые свойства алгоритмов согласования.

1. Пусть $D = f(A, B, C,...)$. Если $a<b$ в согласующей кластеризованной ранжировке $D$, то $a<b$ или $a=b$ в каждой из исходных ранжировок $A$, $B$, $C$, ..., причем хотя бы в одной из них справедливо строгое неравенство.

2. Построение согласующих кластеризованных ранжировок может осуществляться поэтапно. В частности,

$$f(A,B,C) = f(f(A,B), f(A,C), f(B,C)).$$

Ясно, что ядро противоречий для набора кластеризованных ранжировок является объединением таких ядер для всех пар рассматриваемых ранжировок.

3. Построение согласующих кластеризованных ранжировок нацелено на выделение общего упорядочения в исходных кластеризованных ранжировках. Однако при этом некоторые общие свойства исходных кластеризованных ранжировок могут теряться. Так, при согласовании ранжировок $B$ и $C$, рассмотренных выше, противоречия в упорядочении элементов 1 и 2 не было - в ранжировке $B$ эти объекты входили в один кластер, т.е. 1 = 2, в то время как 1<2 в кластеризованной ранжировке $C$. Значит, при их отдельном рассмотрении можно принять упорядочение 1<2. Однако в $f(B,C)$ они попали в один кластер, т.е. возможность их упорядочения

исчезла. Это связано с поведением объекта 3, который "перескочил" в *C* на первое место и "увлек с собой в противоречие" пару (1, 2), образовав противоречивые пары и с 1, и с 2. Другими словами, связная компонента графа, соответствующего ядру противоречий, сама по себе не всегда является полным графом. Недостающие ребра при этом соответствуют парам типа (1, 2), которые сами по себе не являются противоречивыми, но "увлекаются в противоречие" другими парами.

4. Необходимость согласования кластеризованных ранжировок возникает, в частности, при разработке методики применения экспертных оценок в задачах экологического страхования и химической безопасности биосферы. Как уже говорилось, популярным является метод упорядочения по средним рангам, в котором итоговая ранжировка строится на основе средних арифметических рангов, выставленных отдельными экспертами [8, 37]. Однако из теории измерений известно (см. главу 2.1), что более обоснованным является использование не средних арифметических, а медиан. Вместе с тем метод средних рангов весьма известен и широко применяется, так что просто отбросить его нецелесообразно. Поэтому было принято решение об одновременном применении обеих методов. Реализация этого решения потребовала разработки методики согласования двух указанных кластеризованных ранжировок.

5. Область применения рассматриваемого метода не ограничивается экспертными оценками. Он может быть использован, например, для сравнения качества математических моделей процесса испарения жидкости. Имелись данные экспериментов и результаты расчетов по 8 математическим моделям. Сравнивать модели можно по различным критериям качества. Например, по сумме модулей относительных отклонений расчетных и экспериментальных значений. Можно действовать и по другому. В каждой экспериментальной точке упорядочить модели по качеству, а потом получить единые оценки методами средних рангов и медиан. Использовались и иные методы. Затем применялись методы согласования кластеризованных ранжировок, полученных различными способами. В результате оказалось возможным упорядочить модели по качеству и использовать это упорядочение при разработке банка математических моделей, используемого в задачах химической безопасности биосферы.

6. Рассматриваемый метод согласования кластеризованных ранжировок построен в соответствии с *методологией теории устойчивости*, согласно которой результат обработки данных, инвариантный относительно метода обработки, соответствует реальности, а результат расчетов, зависящий от метода обработки, отражает субъективизм исследователя, а не объективные соотношения.

**Основные математические задачи анализа экспертных оценок.** Ясно, что при анализе мнений экспертов можно применять самые разнообразные статистические методы, описывать их - значит описывать практически всю прикладную статистику. Тем не менее можно выделить основные широко используемые в настоящее время методы математической обработки экспертных оценок - это проверка согласованности мнений экспертов (или классификация экспертов, если нет согласованности) и усреднение мнений экспертов внутри согласованной группы.

Поскольку ответы экспертов во многих процедурах экспертного опроса - не числа, а такие объекты нечисловой природы, как градации качественных признаков, ранжировки, разбиения, результаты парных сравнений, нечеткие предпочтения и т.д., то для их анализа оказываются полезными методы статистики нечисловых данных.

**Почему ответы экспертов часто носят нечисловой характер?** Наиболее общий ответ состоит в том, что люди не мыслят числами. В мышлении человека используются образы, слова, но не числа. Поэтому требовать от эксперта ответ в форме чисел - значит насиловать его разум. Даже в экономике менеджеры и предприниматели, принимая решения, лишь частично опираются на численные расчеты. Это видно из условного (т.е. определяемого произвольно принятыми соглашениями, обычно оформленными в виде нормативных актов и инструкций) характера балансовой прибыли, амортизационных отчислений и других экономических показателей. Поэтому фраза типа «фирма стремится к максимизации прибыли» не может иметь строго определенного смысла. Достаточно спросить: «Максимизация прибыли - за какой период?» И сразу станет ясно, что степень оптимальности принимаемых решений зависит от горизонта планирования (на экономико-математическом уровне этот сюжет рассмотрен в монографии [1]).

Эксперт может сравнить два объекта, сказать, какой из двух лучше (метод парных сравнений), дать им оценки типа "хороший", "приемлемый", "плохой", упорядочить несколько объектов по привлекательности, но обычно не может ответить, во сколько раз или на сколько один

объект лучше другого. Другими словами, ответы эксперта обычно измерены в порядковой шкале, или являются ранжировками, результатами парных сравнений и другими объектами нечисловой природы, но не числами. *Распространенное заблуждение состоит в том, что ответы экспертов стараются рассматривать как числа, занимаются "оцифровкой" их мнений, приписывая этим мнениям численные значения - баллы, которые потом обрабатывают с помощью методов прикладной статистики как результаты обычных физико-технических измерений.* В случае произвольности "оцифровки" выводы, полученные в результате подобной обработки данных, могут не иметь отношения к реальности. В связи с "оцифровкой" уместно вспомнить классическую притчу о человеке, который ищет потерянные ключи под фонарем, хотя потерял их в кустах. На вопрос, почему он так делает, отвечает: "Под фонарем светлее". Это, конечно, верно. Но, к сожалению, весьма малы шансы найти потерянные ключи под фонарем. Так и с "оцифровкой" нечисловых данных. Она дает возможность имитации научной деятельности, но не возможность найти истину.

**Проверка согласованности мнений экспертов и классификация экспертных мнений.** Ясно, что мнения разных экспертов различаются. Важно понять, насколько велико это различие. Если мало - усреднение мнений экспертов позволит выделить то общее, что есть у всех экспертов, отбросив случайные отклонения в ту или иную сторону. Если велико - усреднение является чисто формальной процедурой. Так, если представить себе, что ответы экспертов равномерно покрывают поверхность бублика, то формальное усреднение укажет на центр дырки от бублика, а такого мнения не придерживается ни один эксперт. Из сказанного ясна важность проблемы проверки согласованности мнений экспертов.

Разработан ряд методов такой проверки. Статистические методы проверки согласованности зависят от математической природы ответов экспертов. Соответствующие статистические теории весьма трудны, если эти ответы - ранжировки или разбиения, и достаточно просты, если ответы - результаты независимых парных сравнений. Отсюда вытекает рекомендация по организации экспертного опроса: не старайтесь сразу получить от эксперта ранжировку или разбиение, ему трудно это сделать, да и имеющиеся математические методы не позволяют далеко продвинуться в анализе подобных данных.

Например, рекомендуют проверять согласованность ранжировок с помощью коэффициента ранговой конкордации Кендалла-Смита. Но давайте вспомним, какая статистическая модель при этом используется. Проверяется нулевая гипотеза, согласно которой ранжировки независимы и равномерно распределены на множестве всех ранжировок. Если эта гипотеза принимается, то конечно, ни о какой согласованности мнений экспертов говорить нельзя. А если отклоняется? Тоже нельзя. Например, может быть два (или больше) центра, около которых группируются ответы экспертов. Нулевая гипотеза отклоняется. Но разве можно говорить о согласованности?

Эксперту гораздо легче на каждом шагу сравнивать только два объекта. Пусть он занимается парными сравнениями. *Непараметрическая теория парных сравнений (теория люсианов) позволяет решать более сложные задачи, чем статистика ранжировок или разбиений.* В частности, вместо гипотезы равномерного распределения можно рассматривать гипотезу однородности, т.е. вместо совпадения всех распределений с одним фиксированным (равномерным) можно проверять лишь совпадение распределений мнений экспертов между собой, что естественно трактовать как согласованность их мнений. Таким образом, удается избавиться от неестественного предположения равномерности.

При отсутствии согласованности экспертов естественно разбить их на группы сходных по мнению. Это можно сделать различными методами статистики объектов нечисловой природы, относящимися к кластер-анализу, предварительно введя метрику в пространство мнений экспертов. Идея американского математика Джона Кемени об аксиоматическом введении метрик нашла многочисленных продолжателей. Однако методы кластер-анализа обычно являются эвристическими. В частности, обычно невозможно с позиций статистической теории строго обосновать "законность" объединения двух кластеров в один. Имеется важное исключение - *для независимых парных сравнений (люсианов) разработаны методы, позволяющие проверять возможность объединения кластеров как статистическую гипотезу*. Это - еще один аргумент за то, чтобы рассматривать теорию люсианов как ядро математических методов экспертных оценок.

**Нахождение итогового мнения комиссии экспертов.** Пусть мнения комиссии экспертов или какой-то ее части признаны согласованными. Каково же итоговое (среднее, общее) мнение комиссии? Согласно идее Джона Кемени следует найти среднее мнение как решение

*оптимизационной задачи*. А именно, надо минимизировать суммарное расстояние от кандидата в средние до мнений экспертов. Найденное таким способом среднее мнение называют "медианой Кемени".

Математическая сложность состоит в том, что мнения экспертов лежат в некотором пространстве объектов нечисловой природы. Общая теория подобного усреднения рассмотрена выше (глава 2.1.5). В частности, показано, что в силу закона больших чисел (в пространствах произвольной природы) среднее мнение при увеличении числа экспертов (чьи мнения независимы и одинаково распределены) приближается к некоторому пределу, который, как известно, является *математическим ожиданием* (случайного элемента, имеющего то же распределение, что и ответы экспертов).

В конкретных пространствах нечисловых мнений экспертов вычисление медианы Кемени может быть достаточно сложным делом. Кроме свойств пространства, велика роль конкретных метрик. Так, в пространстве ранжировок при использовании метрики, связанной с коэффициентом ранговой корреляции Кендалла, необходимо проводить достаточно сложные расчеты, в то время как применение показателя различия на основе коэффициента ранговой корреляции Спирмена приводит к упорядочению по средним рангам.

**Бинарные отношения и расстояние Кемени.** Как известно, бинарное отношение $A$ на конечном множестве $Q = \{q_1, q_2,..., q_k\}$ - это подмножество *декартова квадрата* $Q^2 = \{(q_m, q_n), m,n = 1, 2, \ldots, k\}$. При этом пара $(q_m, q_n)$ входит в $A$ тогда и только тогда, когда между $q_m$ и $q_n$ имеется рассматриваемое отношение.

Напомним, что каждую кластеризованную ранжировку, как и любое бинарное отношение, можно задать квадратной матрицей $\|x(a,b)\|$ из 0 и 1 порядка $k$ Ч $k$. При этом $x(a, b) = 1$ тогда и только тогда, когда $a < b$ либо $a = b$. В первом случае $x(b, a) = 0$, а во втором $x(b, a) = 1$. При этом хотя бы одно из чисел $x(a, b)$ и $x(b, a)$ равно 1.

В экспертных методах используют, в частности, такие бинарные отношения, как ранжировки (упорядочения, или разбиения на группы, между которыми имеется строгий порядок), отношения эквивалентности, толерантности (отношения сходства). Как следует из сказанного выше, каждое бинарное отношение $A$ можно описать матрицей $\| a(i ,j) \|$ из 0 и 1, причем $a(i ,j) = 1$ тогда и только тогда, когда $q_i$ и $q_j$ находятся в отношении $A$, и $a(i ,j) = 0$ в противном случае.

***Определение.*** *Расстоянием Кемени между бинарными отношениями A и B, описываемыми матрицами* $\| a(i ,j) \|$ и $\| b(i ,j) \|$ *соответственно, называется число*

$$d(A, B) = \sum \left| a(i ,j) - b(i ,j) \right|,$$

*где суммирование производится по всем i, j от* 1 *до k, т.е. расстояние Кемени между бинарными отношениями равно сумме модулей разностей элементов, стоящих на одних и тех же местах в соответствующих им матрицах.*

Легко видеть, что расстояние Кемени - это число несовпадающих элементов в матрицах $\|a(i,j)\|$ и $\|b(i,j)\|$.

Расстояние Кемени основано на некоторой системе аксиом. Эта система аксиом и вывод из нее формулы для расстояния Кемени между упорядочениями содержится в книге [10]. Она сыграла большую роль в развитии в нашей стране такого научного направления, как анализ нечисловой информации (см. историю вопроса в монографиях [1, 8]). В дальнейшем под влиянием работ Дж. Кемени были предложены различные системы аксиом для получения расстояний в тех или иных нужных для социально-экономических, технических, медицинских и иных исследований пространствах (см. главу 1.1.6).

**Медиана Кемени и законы больших чисел.** С помощью расстояния Кемени находят итоговое мнение комиссии экспертов. Пусть $A_1$ , $A_2$ , $A_3$ ,..., $A_p$ - ответы $p$ экспертов, представленные в виде бинарных отношений. Для их усреднения используют *медиану Кемени*

$$\text{Arg min} \sum d(A_i ,A),$$

где Arg min - то или те значения $A$, при которых достигает минимума указанная сумма расстояний Кемени от ответов экспертов до текущей переменной $A$, по которой и проводится минимизация. Таким образом,

$$\sum d(A_i ,A) = d(A_1 ,A) + d(A_2 ,A) + d(A_3 ,A) + \ldots + d(A_p ,A).$$

Кроме медианы Кемени, используют *среднее по Кемени*, в котором вместо $d(A_i , A)$ стоит $d^2(A_i, A)$.

Медиана Кемени - частный случай определения эмпирического среднего в пространствах нечисловой природы. Для нее справедлив закон больших чисел, т.е. эмпирическое среднее

приближается при росте числа составляющих (т.е. $p$ - числа слагаемых в сумме), к теоретическому среднему:

$$\text{Arg min} \sum d(A_i, A) \rightarrow \text{Arg min } M\, d(A_1, A) .$$

Здесь $M$ - символ математического ожидания. Предполагается, что ответы $p$ экспертов $A_1$, $A_2$, $A_3, \ldots, A_p$ есть основания рассматривать как независимые одинаково распределенные случайные элементы (т.е. как случайную выборку) в соответствующем пространстве произвольной природы, например, в пространстве упорядочений или отношений эквивалентности. Систематически эмпирические и теоретические средние и соответствующие различные варианты законов больших чисел рассмотрены выше в главе 2.1.5.

Законы больших чисел показывают, во-первых, что медиана Кемени обладает *устойчивостью* по отношению к незначительному изменению состава экспертной комиссии; во-вторых, при увеличении числа экспертов она *приближается к некоторому пределу*. Его естественно рассматривать как *истинное мнение* экспертов, от которого каждый из них несколько отклонялся по случайным причинам.

Вычисление медианы Кемени - задача целочисленного программирования. Для ее нахождения используется различные алгоритмы дискретной математики, в частности, основанные на методе ветвей и границ. Применяют также алгоритмы, основанные на идее случайного поиска, поскольку для каждого бинарного отношения нетрудно найти множество его соседей.

Рассмотрим упрощенный пример вычисления медианы Кемени. Пусть дана квадратная матрица (порядка 9) попарных расстояний для множества бинарных отношений из 9 элементов $A_1$, $A_2$, $A_3$ ,..., $A_9$ (см. табл.3). Пусть требуется найти в этом множестве *медиану* для множества из 5 элементов $\{A_2, A_4, A_5, A_8, A_9\}$.

Таблица 3.
Матрица попарных расстояний

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $A_6$ | $A_7$ | $A_8$ | $A_9$ |
|---|---|---|---|---|---|---|---|---|---|
| $A_1$ | 0 | 2 | 13 | 1 | 7 | 4 | 10 | 3 | 11 |
| $A_2$ | 2 | 0 | 5 | 6 | 1 | 3 | 2 | 5 | 1 |
| $A_3$ | 13 | 5 | 0 | 2 | 2 | 7 | 6 | 5 | 7 |
| $A_4$ | 1 | 6 | 2 | 0 | 5 | 4 | 3 | 8 | 8 |
| $A_5$ | 7 | 1 | 2 | 5 | 0 | 10 | 1 | 3 | 7 |
| $A_6$ | 4 | 3 | 7 | 4 | 10 | 0 | 2 | 1 | 5 |
| $A_7$ | 10 | 2 | 6 | 3 | 1 | 2 | 0 | 6 | 3 |
| $A_8$ | 3 | 5 | 5 | 8 | 3 | 1 | 6 | 0 | 9 |
| $A_9$ | 11 | 1 | 7 | 8 | 7 | 5 | 3 | 9 | 0 |

В соответствии с определением медианы Кемени следует ввести в рассмотрение функцию

$$C(A) = \sum d(A_i, A) = d(A_2, A)+d(A_4, A)+d(A_5, A)+d(A_8, A)+d(A_9, A),$$

рассчитать ее значения для всех $A_1$, $A_2$, $A_3$, ..., $A_9$ и выбрать наименьшее. Проведем расчеты:

$$C(A_1) = d(A_2, A_1)+d(A_4, A_1)+d(A_5, A_1)+d(A_8, A_1)+d(A_9, A_1) = \\ = 2 + 1 +7 +3 +11 = 24,$$

$$C(A_2) = d(A_2, A_2)+d(A_4, A_2)+d(A_5, A_2)+d(A_8, A_2)+d(A_9, A_2) = \\ = 0 + 6 + 1 + 5 + 1 = 13,$$

$$C(A_3) = d(A_2, A_3)+d(A_4, A_3)+d(A_5, A_3)+d(A_8, A_3)+d(A_9, A_3) = \\ = 5 + 2 + 2 + 5 +7 = 21,$$

$$C(A_4) = d(A_2, A_4)+d(A_4, A_4)+d(A_5, A_4)+d(A_8, A_4)+d(A_9, A_4) = \\ = 6 + 0 + 5 + 8 + 8 = 27,$$

$$C(A_5) = d(A_2, A_5)+d(A_4, A_5)+d(A_5, A_5)+d(A_8, A_5)+d(A_9, A_5) = \\ = 1 + 5 + 0 +3 + 7 = 16,$$

$$C(A_6) = d(A_2, A_6)+d(A_4, A_6)+d(A_5, A_6)+d(A_8, A_6)+d(A_9, A_6) = \\ = 3 + 4 + 10 + 1 + 5 = 23,$$

$$C(A_7) = d(A_2, A_7)+d(A_4, A_7)+d(A_5, A_7)+d(A_8, A_7)+d(A_9, A_7) = \\ = 2 + 3 +1 + 6 + 3 = 15,$$

$$C(A_8) = d(A_2, A_8)+d(A_4, A_8)+d(A_5, A_8)+d(A_8, A_8)+d(A_9, A_8) = \\ = 5 + 8 + 3 + 0 +9 = 25,$$

$$C(A_9) = d(A_2, A_9)+d(A_4, A_9)+d(A_5, A_9)+d(A_8, A_9)+d(A_9, A_9) =$$
$$= 1 + 8 + 7 + 9 + 0 = 25.$$

Из всех вычисленных сумм наименьшая равна 13, и достигается она при $A=A_2$, следовательно, медиана Кемени - это множество $\{A_2\}$, состоящее из одного элемента $A_2$.

## Литература


1. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука, 1979. - 296 с.
2. Орлов А.И. Статистика объектов нечисловой природы и экспертные оценки. – В сб.: Экспертные оценки / Вопросы кибернетики. Вып.58. - М.: Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1979. - С.17-33.
3. Кривцов В.С., Орлов А.И., Фомин В.Н. Современные статистические методы в стандартизации и управлении качеством продукции. – Журнал «Стандарты и качество». 1988. No.3. С.32-36.
4. Беляев Ю.К. Вероятностные методы выборочного контроля. - М.: Наука, 1975. - 408 с.
5. Лумельский Я.П. Статистические оценки результатов контроля качества. - М.: Изд-во стандартов, 1979. - 200 с.
6. Орлов А.И. Статистика объектов нечисловой природы (Обзор). – Журнал «Заводская лаборатория». 1990. Т.56. No.3. С.76-83.
7. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В. Прохоров. - М.: Большая Российская энциклопедия, 1999. - 910 с.
8. Орлов А.И. Эконометрика. Учебник для вузов. Изд. 2-е, исправленное и дополненное. - М.: Изд-во "Экзамен", 2003. – 576 с.
9. Толстова Ю.Н. Анализ социологических данных. – М.: Научный мир, 2000. – 352 с.
10. Кемени Дж., Снелл Дж. Кибернетическое моделирование: Некоторые приложения. - М.: Советское радио, 1972. - 192 с.
11. Орлов А.И. Асимптотика решений экстремальных статистических задач. – В сб.: Анализ нечисловых данных в системных исследованиях. Сборник трудов. Вып.10. - М.: Всесоюзный научно-исследовательский институт системных исследований, 1982. - С. 4-12.
12. Орлов А.И. Асимптотическое поведение статистик интегрального типа. – В сб.: Вероятностные процессы и их приложения. Межвузовский сборник. - М.: МИЭМ, 1989. С.118-123.
13. Кендэл М. Ранговые корреляции. - М.:Статистика,1975. - 216 с.
14. Раушенбах Г.В. Меры близости и сходства. - В сб.: Анализ нечисловой информации в социологических исследованиях. - М.: Наука, 1985. - С.169-203.
15. Маамяги А.В. Некоторые задачи статистического анализа классификаций. – Таллинн: АН ЭССР, 1982. – 24 с.
16. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1983 (3-е изд.). - 474 с.
17. Крамер Г. Математические методы статистики. - М.: Мир, 1975. - 648 с.
18. Орлов А.И. Парные сравнения в асимптотике Колмогорова. – В сб.: Экспертные оценки в задачах управления. - М.: Изд-во Института проблем управления АН СССР, 1982. - С. 58-66.
19. Орлов А.И. Случайные множества с независимыми элементами (люсианы) и их применения. – В сб.: Алгоритмическое и программное обеспечение прикладного статистического анализа. Ученые записки по статистике, т.36. - М.: Наука, 1980. - С. 287-308.
20. Рыданова Г.В. Некоторые вопросы статистического анализа случайных бинарных векторов. Дисс. ... канд. физ.-мат. наук. - М.: МГУ, ф-т вычислит. математ. и кибернет., 1987. - 139 с.
21. Аксенова Г.А., Кузьмина Е.С., Орлов А.И., Розова Н.К. Кинетотопография в диагностике инфаркта миокарда. – В сб.: Актуальные вопросы клинической и экспериментальной медицины. – М.: 4 Главное Управление при Минздраве СССР, 1979. С.24-26.
22. Попов В.Г., Аксенова Г.А., Орлов А.И., Розова Н.К., Кузьмина Е.С. Кинетокардиография в определении зон асинергии у больных инфарктом миокарда. - Журнал «Клиническая медицина». 1982. Т.LX. No.3. С.25-30.
23. Методические рекомендации по проведению экспертной оценки планируемых и законченных научных работ в области медицины (по проблемам союзного значения) / Составители: Г.В. Раушенбах, О.В. Филиппов. - М.: АМН СССР - Ученый медицинский совет Минздрава СССР, 1982. - 36 с.
24. Леман Э. Проверка статистических гипотез. - М.: Наука, 1979. - 408 с.

25. Боровков А.А. Математическая статистика/Учебное пособие для вузов. - М.: Наука, 1984. - 472 с.
26. Любищев А.А. Дисперсионный анализ в биологии. - М.: Изд-во МГУ, 1986. - 200 с.
27. Дылько Т.Н. Проверка гипотез в экспертном оценивании / Вестник Белорусского государственного университета / Сер. 1: Физика, математика и механика. 1988, N 2. С. 36-40.
28. Орлов А.И., Раушенбах Г.В. Метрика подобия: аксиоматическое введение, асимптотическая нормальность. - В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. - Пермь: Изд-во Пермского государственного университета, 1986, с.148-157.
29. Орлов А.И., Миронова Н.Г., Фомин В.Н., Черчинцев А.Н. Рекомендации. Прикладная статистика. Методы обработки данных. Основные требования и характеристики. - М.: ВНИИСтандартизации, 1987. - 62 с.
30. Тюрин Ю.Н., Василевич А.П. К проблеме обработки рядов ранжировок. - В сб.: Статистические методы анализа экспертных оценок. Ученые записки по статистике, т.29. - М.: Наука, 1977. - С. 96-111.
31. Орлов А.И. Некоторые вероятностные вопросы теории классификации. – В сб.: Прикладная статистика. Ученые записки по статистике, т.45. - М.: Наука, 1983. - С. 166-179.
32. Дэвид Г. Метод парных сравнений. - М.: Статистика, 1978.- 144 с.
33. Орлов А.И. Сходимость эталонных алгоритмов. – В сб.: Прикладной многомерный статистический анализ. Ученые записки по статистике, т.33. - М.: Наука, 1978. С. 361-364.
34. Орлов А.И. Задачи оптимизации и нечеткие переменные. – М.: Знание, 1980. - 64 с.
35. Горский В.Г., Гриценко А.А., Орлов А.И., Метод согласования кластеризованных ранжировок // Автоматика и телемеханика. 2000. №3. С.159-167.
36. Шрейдер Ю.А. Равенство, сходство, порядок. - М.: Наука, 1971. – 256 с.
37. Менеджмент. / Под ред. Ж.В. Прокофьевой. - М.: Знание, 2000. - 288 с.

## Контрольные вопросы и задачи

1. Как случайные толерантности используются в теории нечетких толерантностей?
2. В теории люсианов выведите из общего вида несмещенной оценки многочлена от $p$ по результатам $m$ независимых испытаний Бернулли с вероятностью успеха $p$ в каждом (формула (12)) несмещенную оценку в случае $f(p) = 2p(1 - p)$ (формула (13)).
3. Как можно проводить кластерный анализ совокупности нечетких множеств?
4. Чем метод средних арифметических рангов отличает от метода медиан рангов?
5. Почему необходимо согласование кластеризованных ранжировок и как оно проводится?
6. В чем состоит проблема согласованности ответов экспертов?
7. Как бинарные отношения используются в экспертизах?
8. Как бинарные отношения описываются матрицами из 0 и 1?
9. Что такое расстояние Кемени и медиана Кемени?
10. Чем закон больших чисел для медианы Кемени отличается от "классического" закона больших чисел, известного в статистике?
11. В табл. 4 приведены упорядочения 7 инвестиционных проектов, представленные 7 экспертами.

Таблица 4.
Упорядочения проектов экспертами

| Эксперты | Упорядочения |
|---|---|
| 1 | 1 < {2,3} < 4 < 5 < {6,7} |
| 2 | {1,3} < 4 < 2< 5< 7 < 6 |
| 3 | 1 < 4 < 2 < 3 < 6 < 5 < 7 |
| 4 | 1 < {2, 4} < 3 < 5 < 7 <6 |
| 5 | 2 < 3 < 4 < 5 <1 <6 <7 |
| 6 | 1 < 3 < 2 < 5 < 6 < 7 < 4 |
| 7 | 1 < 5 < 3 < 4 < 2 < 6 < 7 |

Найдите:
а) итоговое упорядочение по средним арифметическим рангам;
б) итоговое упорядочение по медианам рангов;
в) кластеризованную ранжировку, согласующую эти два упорядочения.
12. Выпишите матрицу из 0 и 1, соответствующую бинарному отношению (кластеризованной ранжировке) $5 < \{1, 3\} < 4 < 2 < \{6, 7\}$.
13. Найдите расстояние Кемени между бинарными отношениями - упорядочениями $A = [3< 2 <1< \{4,5\}]$ и $B = [1 < \{2 ,3\} < 4 < 5 ]$.
14. Дана квадратная матрица (порядка 9) попарных расстояний (мер различия) для множества бинарных отношений из 9 элементов $A_1$ , $A_2$ , $A_3$ ,..., $A_9$ (табл.5). Найдите в этом множестве медиану для множества из 5 элементов $\{A_2 , A_3 , A_5 , A_6 , A_9\}$.

Таблица 5.
Попарные расстояния между бинарными отношениями

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0 | 5 | 3 | 6 | 7 | 4 | 10 | 3 | 11 |
| 5 | 0 | 5 | 6 | 10 | 3 | 2 | 5 | 7 |
| 3 | 5 | 0 | 8 | 2 | 7 | 6 | 5 | 7 |
| 6 | 6 | 8 | 0 | 5 | 4 | 3 | 8 | 8 |
| 7 | 10 | 2 | 5 | 0 | 10 | 8 | 3 | 7 |
| 4 | 3 | 7 | 4 | 10 | 0 | 2 | 3 | 5 |
| 10 | 2 | 6 | 3 | 8 | 2 | 0 | 6 | 3 |
| 3 | 5 | 5 | 8 | 3 | 3 | 6 | 0 | 9 |
| 11 | 7 | 7 | 8 | 7 | 5 | 3 | 9 | 0 |

## Темы докладов и рефератов

1. Рассчитайте мощность статистик *W* и *N*, рассматриваемых в теории равномерно распределенных случайных толерантностей.
2. Изучите распределение при альтернативах статистики *T*, используемой для проверки однородности двух групп люсианов (при безграничном росте объемов групп).
3. По данным примера в подразделе 3.4.4 найдите методом наименьших квадратов взаимное положение четырех компаний на оси «качество бензина», т.е. найдите их «ценности» $V_1$, $V_2$, $V_3$, $V_4$.
4. Методы оценивания функции принадлежности нечеткого множества.
5. Описание данных для выборок, элементы которых – нечеткие множества.
6. Регрессионный анализ нечетких переменных.
7. Непараметрические оценки плотности распределения вероятностей в пространстве нечетких множеств.
8. Классификация мнений экспертов и проверка согласованности.
9. Использование люсианов в теории и практике экспертных оценок.
10. Формирование итогового мнения комиссии экспертов.

## 3.5. Статистика интервальных данных

В статистике интервальных данных элементы выборки - не числа, а интервалы. Это приводит к алгоритмам и выводам, принципиально отличающимся от классических. Настоящая глава посвящена основным идеям и подходам асимптотической статистики интервальных данных. Приведены результаты, связанные с основополагающими в рассматриваемой области прикладной математической статистики понятиями нотны и рационального объема выборки. Рассмотрен ряд задач оценивания характеристик и параметров распределения, проверки гипотез, регрессионного, кластерного и дискриминантного анализов.

### 3.5.1. Основные идеи статистики интервальных данных

Перспективная и быстро развивающаяся область статистических исследований последних лет - математическая статистика интервальных данных. Речь идет о развитии методов прикладной математической статистики в ситуации, когда статистические данные - не числа, а интервалы, в частности, порожденные наложением ошибок измерения на значения случайных величин. Полученные результаты отражены, в частности, в выступлениях на проведенной в "Заводской лаборатории" дискуссии [1] и в докладах международной конференции ИНТЕРВАЛ-92 [2]. Приведем основные идеи весьма перспективного для вероятностно-статистических методов и моделей принятия решений асимптотического направления в статистике интервальных данных.

В настоящее время признается необходимым изучение устойчивости (робастности) оценок параметров к малым отклонениям исходных данных и предпосылок модели. Однако популярная среди теоретиков модель засорения (Тьюки-Хьюбера) представляется не вполне адекватной. Эта модель нацелена на изучение влияния больших "выбросов". Поскольку любые реальные измерения лежат в некотором фиксированном диапазоне, а именно, заданном в техническом паспорте средства измерения, то зачастую выбросы не могут быть слишком большими. Поэтому представляются полезными иные, более общие схемы устойчивости, в частности, введенные в [3], в которых, например, учитываются отклонения распределений результатов наблюдений от предположений модели.

В одной из таких схем изучается влияние интервальности исходных данных на статистические выводы. Необходимость такого изучения стала очевидной следующим образом. В государственных стандартах СССР по прикладной статистике в обязательном порядке давалось справочное приложение "Примеры применения правил стандарта". При разработке ГОСТ 11.011-83 [4] были переданы для анализа реальные данные о наработке резцов до предельного состояния (в часах). Оказалось, что все эти данные представляли собой либо целые числа, либо полуцелые (т.е. после умножения на 2 становящиеся целыми). Ясно, что исходная длительность наработок искажена. Необходимо учесть в статистических процедурах наличие такого искажения исходных данных. Как это сделать?

Первое, что приходит в голову - модель группировки данных, согласно которой для истинного значения $X$ проводится замена на ближайшее число из множества $\{0,5n,\ n=1,2,3,...\}$. Однако эту модель целесообразно подвергнуть сомнению, а также рассмотреть иные модели. Так, возможно, что $X$ надо приводить к ближайшему сверху элементу указанного множества - если проверка качества поставленных на испытание резцов проводилась раз в полчаса. Другой вариант: если расстояния от $X$ до двух ближайших элементов множества $\{0,5n,\ n=1,2,3,...\}$ примерно равны, то естественно ввести рандомизацию при выборе заменяющего числа, и т.д.

Целесообразно построить новую математико-статистическую модель, согласно которой **результаты наблюдений - не числа, а интервалы**. Например, если в таблице приведено значение 53,5, то это значит, что реальное значение - какое-то число от 53,0 до 54,0, т.е. какое-то число в интервале [53,5 - 0,5; 53,5 + 0,5], где 0,5 - максимально возможная погрешность. Принимая эту модель, мы попадаем в новую научную область - статистику интервальных данных [5,6]. Статистика интервальных данных идейно связана с интервальной математикой, в которой в роли чисел выступают интервалы (см., например, монографию [7]). Это направление математики является дальнейшим развитием всем известных правил приближенных вычислений,

посвященных выражению погрешностей суммы, разности, произведения, частного через погрешности тех чисел, над которыми осуществляются перечисленные операции.

В интервальной математике сумма двух интервальных чисел $[a,b]$ и $[c,d]$ имеет вид $[a,b] + [c,d] = [a+c, b+d]$, а разность определяется по формуле $[a,b] - [c,d] = [a-d, b-c]$. Для положительных $a, b, c, d$ произведение определяется формулой $[a,b] * [c,d] = [ac, bd]$, а частное имеет вид $[a,b] / [c,d] = [a/d, b/c]$. Эти формулы получены при решении соответствующих оптимизационных задач. Пусть $x$ лежит в отрезке $[a,b]$, а $y$ – в отрезке $[c,d]$. Каково минимальное и максимальное значение для $x+y$? Очевидно, $a+c$ и $b+d$ соответственно. Минимальные и максимальные значения для $x-y$, $xy$, $x/y$ задают нижние и верхние границы для интервальных чисел, задающих результаты арифметических операций. А от арифметических операций можно перейти ко всем остальным математическим алгоритмам. Так строится интервальная математика.

Как видно из сборника трудов Международной конференции [2], к настоящему времени удалось решить, в частности, ряд задач теории интервальных дифференциальных уравнений, в которых коэффициенты, начальные условия и решения описываются с помощью интервалов. По мнению ряда специалистов, статистика интервальных данных является частью интервальной математики [7]. Впрочем, есть точка зрения, согласно которой такое включение нецелесообразно, поскольку статистика интервальных данных использует несколько иные подходы к алгоритмам анализа реальных данных, чем сложившиеся в интервальной математике (подробнее см. ниже).

В настоящей главе развиваем асимптотические методы статистического анализа интервальных данных при больших объемах выборок и малых погрешностях измерений. В отличие от классической математической статистики, сначала устремляется к бесконечности объем выборки и только потом - уменьшаются до нуля погрешности. В частности, еще в начале 1980-х годов с помощью такой асимптотики были сформулированы правила выбора метода оценивания в ГОСТ 11.011-83 [4].

Разработана [8] общая схема исследования, включающая расчет нотны (максимально возможного отклонения статистики, вызванного интервальностью исходных данных) и рационального объема выборки (превышение которого не дает существенного повышения точности оценивания). Она применена к оцениванию математического ожидания и дисперсии [1], медианы и коэффициента вариации [9], параметров гамма-распределения [4, 10] и характеристик аддитивных статистик [8], при проверке гипотез о параметрах нормального распределения, в т.ч. с помощью критерия Стьюдента, а также гипотезы однородности с помощью критерия Смирнова [9]. Изучено асимптотическое поведение оценок метода моментов и оценок максимального правдоподобия (а также более общих - оценок минимального контраста), проведено асимптотическое сравнение этих методов в случае интервальных данных, найдены общие условия, при которых, в отличие от классической математической статистики, метод моментов дает более точные оценки, чем метод максимального правдоподобия [11].

Разработаны подходы к рассмотрению интервальных данных в основных постановках регрессионного, дискриминантного и кластерного анализов [12]. В частности, изучено влияние погрешностей измерений и наблюдений на свойства алгоритмов регрессионного анализа, разработаны способы расчета нотн и рациональных объемов выборок, введены и исследованы новые понятия многомерных и асимптотических нотн, доказаны соответствующие предельные теоремы [12,13]. Начата разработка интервального дискриминантного анализа, в частности, рассмотрено влияние интервальности данных на показатель качества классификации [12,14]. Основные идеи и результаты рассматриваемого направления в статистике интервальных данных приведены в публикациях обзорного характера [5,6].

Как показала, в частности, международная конференция ИНТЕРВАЛ-92, в области асимптотической математической статистики интервальных данных мы имеем мировой приоритет. По нашему мнению, со временем во все виды статистического программного обеспечения должны быть включены алгоритмы интервальной статистики, "параллельные" обычно используемым алгоритмам прикладной математической статистики. Это позволит в явном виде учесть наличие погрешностей у результатов наблюдений, сблизить позиции метрологов и статистиков.

Многие из утверждений статистики интервальных данных весьма отличаются от аналогов из классической математической статистики. В частности, не существует состоятельных оценок;

средний квадрат ошибки оценки, как правило, асимптотически равен сумме дисперсии оценки, рассчитанной согласно классической теории, и некоторого положительного числа (равного квадрату т.н. нотны - максимально возможного отклонения значения статистики из-за погрешностей исходных данных) - в результате метод моментов оказывается иногда точнее метода максимального правдоподобия [11]; нецелесообразно увеличивать объем выборки сверх некоторого предела (называемого рациональным объемом выборки) - вопреки классической теории, согласно которой чем больше объем выборки, тем точнее выводы.

В стандарт [4] был включен раздел 5, посвященный выбору метода оценивания при неизвестных параметрах формы и масштаба и известном параметре сдвига и основанный на концепциях статистики интервальных данных. Теоретическое обоснование этого раздела стандарта опубликовано лишь через 5 лет в статье [10].

Следует отметить, что хотя в 1982 г. при разработке стандарта [4] были сформулированы основные идеи статистики интервальных данных, однако из-за недостатка времени они не были полностью реализованы в ГОСТ 11.011-83, и этот стандарт написан в основном в классической манере. Развитие идей статистики интервальных данных продолжается уже в течение 20 лет, и еще много чего надо сделать! Большое значение статистики интервальных данных для современной прикладной статистики обосновано в [15,16].

Ведущая научная школа в области статистики интервальных данных - это школа проф. А.П. Вощинина, активно работающая с конца 70-х годов. Полученные результаты отражены в ряде монографий (см., в частности, [17,18,19]), статей [1, 20, 21], докладов, в частности, в трудах [2] Международной конференции ИНТЕРВАЛ-92, диссертаций [22,23]. В частности, изучены проблемы регрессионного анализа, планирования эксперимента, сравнения альтернатив и принятия решений в условиях интервальной неопределенности. Рассматриваемое ниже направление отличается нацеленностью на асимптотические результаты, полученные при больших объемах выборок и малых погрешностях измерений, поэтому оно и названо **асимптотической статистикой интервальных данных.**

Сформулируем сначала основные идеи асимптотической математической статистики интервальных данных, а затем рассмотрим реализацию этих идей на перечисленных выше примерах. Следует сразу подчеркнуть, что основные идеи достаточно просты, в то время как их проработка в конкретных ситуациях зачастую оказывается достаточно трудоемкой.

Пусть существо реального явления описывается выборкой $x1\ ,\ x2\ ,\ ...,\ xn$ . В вероятностной теории математической статистики, из которой мы исходим (см. терминологическую статью [24]), выборка - это набор независимых в совокупности одинаково распределенных случайных величин. Однако беспристрастный и тщательный анализ подавляющего большинства реальных задач показывает, что статистику известна отнюдь не выборка $x_1\ ,\ x_2\ ,\ ...,\ x_n$ , а величины

$$yj = xj + \varepsilon j\ ,\ j = 1, 2, ...\ , n\ ,$$

где $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ – некоторые погрешности измерений, наблюдений, анализов, опытов, исследований (например, инструментальные ошибки).

Одна из причин появления погрешностей - запись результатов наблюдений с конечным числом значащих цифр. Дело в том, что для случайных величин с непрерывными функциями распределения событие, состоящее в попадании хотя бы одного элемента выборки в множество рациональных чисел, согласно правилам теории вероятностей имеет вероятность 0, а такими событиями в теории вероятностей принято пренебрегать. Поэтому при рассуждениях о выборках из нормального, логарифмически нормального, экспоненциального, равномерного, гамма - распределений, распределения Вейбулла-Гнеденко и др. приходится принимать, что эти распределения имеют элементы исходной выборки $x_1\ ,\ x_2\ ,\ ...,\ x_n$, в то время как статистической обработке доступны лишь искаженные значения $yj = xj + \varepsilon j$.

Введем обозначения

$$x = (x_1\ ,\ x_2\ ,\ ...,\ x_n\ ),\ y = (y_1\ ,\ y_2\ ,\ ...,\ y_n\ ),\ \varepsilon = \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_n\ .$$

Пусть статистические выводы основываются на статистике $f : R^n \to R^1$, используемой для оценивания параметров и характеристик распределения, проверки гипотез и решения иных

статистических задач. Принципиально важная для статистики интервальных данных идея такова: СТАТИСТИК ЗНАЕТ ТОЛЬКО *f(y),* НО НЕ *f(x).*

Очевидно, в статистических выводах необходимо отразить различие между *f(y)* и *f(x).* Одним из двух основных понятий статистики интервальных данных является понятие нотны.

**Определение.** Величину максимально возможного (по абсолютной величине) отклонения, вызванного погрешностями наблюдений $\varepsilon$, известного статистику значения *f(y)* от истинного значения *f(x*), т.е.

$$Nf(x) = \sup | f(y) - f(x) | ,$$

где супремум берется по множеству возможных значений вектора погрешностей $\varepsilon$ (см. ниже), будем называть НОТНОЙ .

Если функция $f$ имеет частные производные второго порядка, а ограничения на погрешности имеют вид

$$| \varepsilon_i | \le \Delta, \ i = 1,2,...,n, \qquad (1)$$

причем $\Delta$ мало, то приращение функции $f$ с точностью до бесконечно малых более высокого порядка описывается главным линейным членом, т.е.

$$f(y) - f(x) = \sum_{1 \le i \le n} \frac{\partial f(x)}{\partial x_i} \varepsilon_i + O(\Delta^2).$$

Чтобы получить асимптотическое (при $\Delta \to 0$ ) выражение для нотны, достаточно найти максимум и минимум линейной функции (главного линейного члена) на кубе, заданном неравенствами (1). Легко видеть, что максимум достигается, если положить

$$\varepsilon_i = \begin{cases} \Delta, \dfrac{\partial f(x)}{\partial x_i} \ge 0, \\ -\Delta, \dfrac{\partial f(x)}{\partial x_i} < 0, \end{cases}$$

а минимум, отличающийся от максимума только знаком, достигается при $\varepsilon_i^{'} = -\varepsilon_i$. Следовательно, *нотна* с точностью до бесконечно малых более высокого
порядка имеет вид

$$N_f(x) = (\sum_{1 \le i \le n} | \frac{\partial f(x)}{\partial x_i} |)\Delta.$$

Это выражение назовем *асимптотической нотной*.

Условие (1) означает, что исходные данные представляются статистику в виде интервалов $[y_i - \Delta; y_i + \Delta], i = 1,2,...,n$ (отсюда и название этого научного направления). Ограничения на погрешности могут задаваться разными способами - кроме абсолютных ошибок используются относительные или иные показатели различия между *x* и *y*.

Если задана не предельная абсолютная погрешность $\Delta$, а предельная относительная погрешность $\delta$, т.е. ограничения на погрешности вошедших в выборку результатов измерений имеют вид

$$| \varepsilon_i | \le \delta | x_i |, i = 1,2,...,n,$$

то аналогичным образом получаем, что нотна с точностью до бесконечно малых более высокого порядка, т.е. асимптотическая нотна, имеет вид

$$N_f(x) = (\sum_{1 \le i \le n} | x_i \frac{\partial f(x)}{\partial x_i} |)\delta.$$

При практическом использовании рассматриваемой концепции необходимо провести тотальную замену символов *x* на символы *y*. В каждом конкретном случае удается показать, что в силу малости погрешностей разность $N_f(y) - N_f(x)$ является бесконечно малой более высокого порядка сравнительно с $N_f(x)$ или $N_f(y)$.

**Основные результаты в вероятностной модели.** В классической вероятностной модели элементы исходной выборки $x_1$, $x$ , ..., $x_n$ рассматриваются как независимые одинаково распределенные случайные величины. Как правило, существует некоторая константа $C > 0$ такая,

что в смысле сходимости по вероятности

$$\lim_{n\to\infty} N_f(x) = C\Delta. \qquad (2)$$

Соотношение (2) доказывается отдельно для каждой конкретной задачи.

При использовании классических эконометрических методов в большинстве случаев используемая статистика *f(x)* является асимптотически нормальной. Это означает, что существуют константы *a* и $\sigma^2$ такие, что

$$\lim_{n\to\infty} P\left(\sqrt{n}\,\frac{f(x)-a}{\sigma} < x\right) = \Phi(x),$$

где $\Phi(x)-$ функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. При этом обычно оказывается, что

$$\lim_{n\to\infty} \sqrt{n}(Mf(x)-a) = 0$$

и

$$\lim_{n\to\infty} nDf(x) = \sigma^2,$$

а потому в классической эконометрике средний квадрат ошибки статистической оценки равен

$$M(f(x)-a)^2 = (Mf(x)-a)^2 + Df(x) = \frac{\sigma^2}{n}$$

с точностью до членов более высокого порядка.

В статистике интервальных данных ситуация совсем иная - обычно можно доказать, что средний квадрат ошибки равен

$$\max_{\{\varepsilon\}} M(f(y)-a)^2 = \frac{\sigma^2}{n} + N_f^2(y) + o(\Delta^2 + \frac{1}{n}). \qquad (3)$$

Из соотношения (3) можно сделать ряд важных следствий. Прежде всего отметим, что правая часть этого равенства, в отличие от правой части соответствующего классического равенства, не стремится к 0 при безграничном возрастании объема выборки. Она остается больше некоторого положительного числа, а именно, квадрата нотны. Следовательно, статистика *f(x)* не является состоятельной оценкой параметра *a*. Более того, состоятельных оценок вообще не существует.

Пусть доверительным интервалом для параметра *a*, соответствующим заданной доверительной вероятности $\gamma$, в классической математической статистике является интервал $(c_n(\gamma); d_n(\gamma))$. В статистике интервальных данных аналогичный доверительный интервал является более широким. Он имеет вид $(c_n(\gamma) - N_f(y); d_n(\gamma) + N_f(y))$. Таким образом, его длина увеличивается на две нотны. Следовательно, при увеличении объема выборки длина доверительного интервала не может стать меньше, чем $2C\Delta$ (см. формулу (2)).

В статистике интервальных данных методы оценивания параметров имеют другие свойства по сравнению с классической математической статистикой. Так, при больших объемах выборок метод моментов может быть заметно лучше, чем метод максимального правдоподобия (т.е. иметь меньший средний квадрат ошибки - см. формулу (3)), в то время как в классической математической статистике второй из названных методов всегда не хуже первого.

**Рациональный объем выборки.** Анализ формулы (3) показывает, что в отличие от классической математической статистики нецелесообразно безгранично увеличивать объем выборки, поскольку средний квадрат ошибки остается всегда большим квадрата нотны. Поэтому представляется полезным ввести понятие "рационального объема выборки" $n_{rat}$, при достижении которого продолжать наблюдения нецелесообразно.

Как установить "рациональный объем выборки"? Можно воспользоваться идеей "принципа уравнивания погрешностей", выдвинутой в монографии [3]. Речь идет о том, что вклад погрешностей различной природы в общую погрешность должен быть примерно одинаков. Этот принцип дает возможность выбирать необходимую точность оценивания тех или иных характеристик в тех случаях, когда это зависит от исследователя. В статистике интервальных данных в соответствии с "принципом уравнивания погрешностей" предлагается определять рациональный объем выборки $n_{rat}$ из условия равенства двух величин - метрологической

составляющей, связанной с нотной, и статистической составляющей - в среднем квадрате ошибки (3), т.е. из условия

$$\frac{\sigma^2}{n_{rat}} = N_f^2(y), \quad n_{rat} = \frac{\sigma^2}{N_f^2(y)}.$$

Для практического использования выражения для рационального объема выборки неизвестные теоретические характеристики необходимо заменить их оценками. Это делается в каждой конкретной задаче по-своему.

Исследовательскую программу в области статистики интервальных данных можно "в двух словах" сформулировать так: для любого алгоритма анализа данных (алгоритма прикладной статистики) необходимо вычислить нотну и рациональный объем выборки. Или иные величины из того же понятийного ряда, возникающие в многомерном случае, при наличии нескольких выборок и при иных обобщениях описываемой здесь простейшей схемы. Затем проследить влияние погрешностей исходных данных на точность оценивания, доверительные интервалы, значения статистик критериев при проверке гипотез, уровни значимости и другие характеристики статистических выводов. Очевидно, классическая математическая статистика является частью статистики интервальных данных, выделяемой условием $\Delta = 0$.

### 3.5.2. Интервальные данные в задачах оценивания характеристик и параметров распределения

Поясним теоретические концепции статистики интервальных данных на простых примерах.

**Пример 1. Оценивание математического ожидания.** Пусть необходимо оценить математическое ожидание случайной величины с помощью обычной оценки - среднего арифметического результатов наблюдений, т.е.

$$f(x) = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Тогда при справедливости ограничений (1) на абсолютные погрешности имеем $N_f(x) = \Delta$. Таким образом, нотна полностью известна и не зависит от многомерной точки, в которой берется. Вполне естественно: если каждый результат наблюдения известен с точностью до $\Delta$, то и среднее арифметическое известно с той же точностью. Ведь возможна систематическая ошибка - если к каждому результату наблюдению добавить $\Delta$, то и среднее арифметическое увеличится на $\Delta$.

Поскольку

$$D(\bar{x}) = \frac{D(x_1)}{n},$$

то в обозначениях предыдущего пункта

$$\sigma^2 = D(x_1).$$

Следовательно, рациональный объем выборки равен

$$n_{rat} = \frac{D(x_1)}{\Delta^2}.$$

Для практического использования полученной формулы надо оценить дисперсию результатов наблюдений. Можно доказать, что, поскольку $\Delta$ мало, это можно сделать обычным способом, например, с помощью несмещенной выборочной оценки дисперсии

$$s^2(y) = \frac{1}{n-1} \sum_{1 \le i \le n} (y_i - \bar{y})^2.$$

Здесь и далее рассуждения часто идут на двух уровнях. Первый - это уровень "истинных" случайных величин, обозначаемых *"x",* описывающих реальность, но неизвестных специалисту по анализу данных. Второй - уровень известных этому специалисту величин *"y",* отличающихся погрешностями от истинных. Погрешности малы, поэтому функции от *x* отличаются от функций от *y* на некоторые бесконечно малые величины. Эти соображения и позволяют использовать $s^2(y)$ как оценку $D(x_1)$.

Итак, выборочной оценкой рационального объема выборки является

$$n_{sample-rat} = \frac{s^2(y)}{\Delta^2}.$$

Уже на этом первом рассматриваемом примере видим, что рациональный объем выборки находится не где-то вдали, а непосредственно рядом с теми объемами, с которыми имеет дело любой практически работающий статистик. Например, если статистик знает, что $\Delta = \frac{\sigma}{6}$, то $n_{rat}$ = 36. А именно такова погрешность контрольных шаблонов во многих технологических процессах! Поэтому, занимаясь управлением качеством, необходимо обращать внимание на действующую на предприятии систему измерений.

По сравнению с классической математической статистикой доверительный интервал для математического ожидания (для заданной доверительной вероятности $\gamma$) имеет другой вид:

$$(\bar{y} - \Delta - u(\gamma)\frac{s}{\sqrt{n}}; \bar{y} + \Delta + u(\gamma)\frac{s}{\sqrt{n}}), \qquad (4)$$

где $u(\gamma)$ - квантиль порядка (1+ $\gamma$)/2 стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1..

По поводу формулы (4) была довольно жаркая дискуссия среди специалистов. Отмечалось, что она получена на основе Центральной Предельной Теоремы теории вероятностей и может быть использована при любом распределении результатов наблюдений (с конечной дисперсией). Если же имеется дополнительная информация, то, по мнению отдельных специалистов, формула (4) может быть уточнена. Например, если известно, что распределение $x_i$ является нормальным, в качестве $u(\gamma)$ целесообразно использовать квантиль распределения Стьюдента. К этому надо добавить, что по небольшому числу наблюдений нельзя надежно установить нормальность, а при росте объема выборки квантили распределения Стьюдента приближаются к квантилям нормального распределения. Вопрос о том, часто ли результаты наблюдений имеют нормальное распределение, подробно обсуждался среди специалистов. Выяснилось, что распределения встречающихся в практических задачах результатов измерений почти всегда отличны от нормальных [25]. А также и от распределений из иных параметрических семейств, описываемых в учебниках.

Применительно к оцениванию математического ожидания (но не к оцениванию других характеристик или параметров распределения) факт существования границы возможной точности, определяемой точностью исходных данных, не0днократно отмечался в литературе ([26, с.230-234], [31, с.121] и др.).

**Пример 2. Оценивание дисперсии.** Для статистики $f(y) = s^2(y)$, где $s^2(y)$ - выборочная дисперсия (несмещенная оценка теоретической дисперсии), при справедливости ограничений (1) на абсолютные погрешности имеем

$$N_f(y) = \frac{2\Delta}{n-1}\sum_{i=1}^{n} | y_i - \bar{y} | + O(\Delta^2).$$

Можно показать, что нотна $N_f(y)$ сходится к

$$2\Delta M \mid x_1 - M(x_1) \mid$$

по вероятности с точностью до $o(\Delta)$, когда $n$ стремится к бесконечности. Это же предельное соотношение верно и для нотны $N_f(x)$, вычисленной для исходных данных. Таким образом, в данном случае справедлива формула (2) с

$$C = 2M \mid x_1 - M(x_1) \mid .$$

Известно, что случайная величина

$$\frac{s^2 - \sigma^2}{\sqrt{n}}$$

является асимптотически нормальной с математическим ожиданием 0 и дисперсией $D(x_1^2)$.

Из сказанного вытекает, что в статистике интервальных данных асимптотический доверительный интервал для дисперсии $\sigma^2$ (соответствующий доверительной вероятности $\gamma$) имеет вид

$$(s^2(y) - A; \quad s^2 + A),$$

где

$$A = \frac{u(\gamma)}{\sqrt{n(n-1)}} \sqrt{\sum_{i=1}^{n} (y_i^2 - \frac{1}{n}\sum_{j=1}^{n} y_j^2)^2} + \frac{2\Delta}{n-1}\sum_{i=1}^{n} | y_i - \bar{y} |,$$

где $u(\gamma)$ обозначает тот же самый квантиль стандартного нормального распределения, что и выше в случае оценивания математического ожидания.

Рациональный объем выборки при оценивании дисперсии равен

$$n_{rat} = \frac{D(x_1^2)}{4\Delta^2 (M \mid x_1 - M(x_1) \mid)^2},$$

а выборочную оценку рационального объема выборки $n_{sample-rat}$ можно вычислить, заменяя теоретические моменты на соответствующие выборочные и используя доступные статистику результаты наблюдений, содержащие погрешности.

Что можно сказать о численной величине рационального объема выборки? Как и в случае оценивания математического ожидания, она отнюдь не выходит за пределы обычно используемых объемов выборок. Так, если распределение результатов наблюдений $x_i$ является нормальным с математическим ожиданием 0 и дисперсией $\sigma^2$, то в результате вычисления моментов случайных величин в предыдущей формуле получаем, что

$$n_{rat} = \frac{\sigma^2}{\pi\Delta^2},$$

где $\pi$ - отношение длины окружности к диаметру, $\pi = 3{,}141592...$ Например, если $\Delta = \sigma / 6$, то $n_{rat} = 11$. Это меньше, чем при оценивании математического ожидания в предыдущем примере.

**Пример 3. Аддитивные статистики.** Пусть $g : R^1 \to R^1$ - некоторая непрерывная функция. Аддитивные статистики имеют вид

$$f(x) = \frac{1}{n}\sum_{1 \le i \le n} g(x_i).$$

Тогда

$$\sum_{1 \le i \le n} | \frac{\partial f(x)}{\partial x_i} | = \frac{1}{n}\sum_{1 \le i \le n} | \frac{dg(x_i)}{dx_i} | \to M \mid \frac{dg(x_1)}{dx_1} |,$$

$$\sum_{1 \le i \le n} | x_i \frac{\partial f(x)}{\partial x_i} | = \frac{1}{n}\sum_{1 \le i \le n} | x_i \frac{dg(x_i)}{dx_i} | \to M \mid x_1 \frac{dg(x_1)}{dx_1} |$$

по вероятности при $n \to \infty$, если математические ожидания в правых частях двух последних соотношений существуют. Применяя рассмотренные выше общие соображения, получаем, что при малых фиксированных $\Delta$ и $\delta$ и достаточно больших *n* значения *f(y)* могут принимать любые величины из разрешенных (например, записываемых заданным числом значащих цифр) в замкнутом интервале

$$[f(x) - \Delta M \mid \frac{dg(x_1)}{dx_1} |; f(x) + \Delta M \mid \frac{dg(x_1)}{dx_1} |] \quad (5)$$

при ограничениях (1) на абсолютные ошибки и в замкнутом интервале

$$[f(x) - \delta M \mid x_1 \frac{dg(x_1)}{dx_1} |; f(x) + \delta M \mid x_1 \frac{dg(x_1)}{dx_1} |] \ldots (6)$$

при ограничениях на относительные погрешности результатов наблюдений. Обратим внимание, что длины этих интервалов независимы от объема выборки, в частности, не стремятся к 0 при его росте.

К каким последствиям это приводит в задачах статистического оценивания? Поскольку для статистик аддитивного типа

$$f(x) = \frac{1}{n} \sum_{1 \le i \le n} g(x_i) \to Mg(x_1) \quad (7)$$

по вероятности при $n \to \infty$, если математическое ожидание в правой части формулы (7) существует, то аддитивную статистику *f(x)* естественно рассматривать как непараметрическую оценку этого математического ожидания. Термин «непараметрическая» означает, что не делается предположений о принадлежности функции распределения выборки к тому или иному параметрическому семейству распределения. Распределение статистики *f(x)* зависит от распределения результатов наблюдений. Однако для любого распределения результатов наблюдений с конечной дисперсией статистика *f(x)* является состоятельной и асимптотически нормальной оценкой для математического ожидания, указанного в правой части формулы (7).

Как известно, в рамках классической математической статистики в предположении существования ненулевой дисперсии $Dg(x_1)$ в силу асимптотической нормальности аддитивной статистики *f(x)* асимптотический доверительный интервал, соответствующий доверительной вероятности $\gamma$, имеет вид

$$[f(x) - u(\frac{1+\gamma}{2}) \frac{s(g(x))}{\sqrt{n}}; f(x) + u(\frac{1+\gamma}{2}) \frac{s(g(x))}{\sqrt{n}}],$$

где *s(g(x))* – выборочное среднее квадратическое отклонение, построенное по $g(x_1), g(x_2), \ldots, g(x_n)$, а $u(\frac{1+\gamma}{2})$ - квантиль стандартного нормального распределения порядка $\frac{1+\gamma}{2}$.

В рассматриваемой модели порождения интервальных данных вместо *f(x)* необходимо использовать *f(y),* а вместо $g(x_i)$ – соответственно $g(y_i)$, *i*-1,2,…,*n*. При этом доверительный интервал необходимо расширить с учетом формул (5) и (6).

В соответствии с проведенными рассуждениями для аддитивных статистик асимптотическая нотна имеет вид

$$N_f(x) = \Delta M \mid \frac{dg(x_1)}{dx_1} \mid$$

при ограничениях (1) на абсолютную погрешность и

$$N_f(x) = \delta M \mid x_1 \frac{dg(x_1)}{dx_1} \mid$$

при ограничениях на относительную погрешность. В первом случае нотна является обобщением понятия предельной абсолютной систематической ошибки, во втором – предельной относительной систематической ошибки. Отметим, что, как и в примерах 1 и 2, асимптотическая нотна не зависит от точки, в которой вычисляется. Таким образом, она является константой для конкретного метода статистического анализа данных.

Поскольку *n* велико, а $\Delta$ и $\delta$ малы, то можно пренебречь отличием выборочного среднего квадратического отклонения *s(g(y)),* вычисленного по выборке преобразованных значений $g(y_1), g(y_2), \ldots, g(y_n)$, от выборочного среднего квадратического отклонения *s(g(x)),* построенного по выборке $g(x_1), g(x_2), \ldots, g(x_n)$. Разность этих двух величин является бесконечно малой, они приближаются к одной и той же положительной константе.

В статистике интервальных данных выборочный доверительный интервал для $Mg(x_1)$ имеет вид

$$[f(y) - N_f(y) - u(\frac{1+\gamma}{2}) \frac{s(g(y))}{\sqrt{n}}; f(y) + N_f(y) + u(\frac{1+\gamma}{2}) \frac{s(g(y))}{\sqrt{n}}].$$

В асимптотике его длина такова:

$$2N_f(x) + 2u(\frac{1+\delta}{2}) \frac{\sigma}{\sqrt{n}}, \quad (8)$$

где $\sigma^2$ - дисперсия $g(x_1)$, в то время как в классической теории математической статистики имеется только второе слагаемое. Соотношение (8) – аналог суммарной ошибки у метрологов [26]. Поскольку первое слагаемое положительно, то оценивание $Mg(x_1)$ с помощью *f(y)* не является состоятельным.

Для аддитивных статистик при больших $n$ максимум (по возможным погрешностям) среднего квадрата отклонения оценки имеет вид

$$\max_{\varepsilon} M[f(y) - Mg(x_1)]^2 = N_f^2(x) + \frac{Dg(x_1)}{n} \qquad (9)$$

с точностью до членов более высокого порядка. Исходя из принципа уравнивания погрешностей в общей схеме устойчивости [3], нецелесообразно второе слагаемое в (9) делать меньше первого за счет увеличения объема выборки $n$. Рациональный объем выборки, т.е. тот объем, при котором равны погрешности оценивания (или проверки гипотез), вызванные погрешностями исходных данных, и статистические погрешности, рассчитанные по обычным правилам математической статистики (при $\varepsilon_i \equiv 0$), для аддитивных статистик согласно (9) имеет вид

$$n_{rat} = \frac{Dg(x_1)}{N_f^2(x)}. \quad (10)$$

В качестве примера рассмотрим экспоненциально распределенные результаты наблюдений $x_i, M(x_1) = D(x_1) = 1$. Оцениваем математическое ожидание с помощью выборочного среднего арифметического при ограничениях на относительную погрешность. Тогда согласно формуле (10)

$$N_f(x) = \delta, \quad n_{rat} = \frac{1}{\delta^2}.$$

В частности, если относительная погрешность измерений $\delta$=10%, то рациональный объем выборки равен 100. Формуле (10) соответствует также рассмотренный выше пример 1.

**Пример 4. Оценивание медианы распределения с помощью выборочной медианы.** Хотя нельзя выделить главный линейный член из-за недифференцируемости функции *f(x)*, выражающей выборочную медиану через элементы выборки, непосредственно из определения нотны следует, что при ограничениях на абсолютные погрешности

$$N_f(x) = \Delta,$$

а при ограничениях на относительные погрешности

$$N_f(x) = \delta x_{med}$$

с точностью до бесконечно малых более высокого порядка, где $x_{med}$ - теоретическая медиана. Доверительный интервал для медианы имеет вид

$$[a_1(x) - N_f(x); a_2(x) + N_f(x)],$$

где $[a_1(x); a_2(x)]$ - доверительный интервал для медианы, вычисленный по классическим правилам непараметрической статистики [27]. Для нахождения рационального объема выборки можно использовать асимптотическую дисперсию выборочной медианы. Она, как известно (см., например, [28, с.178]), равна

$$\sigma^2(M) = \frac{1}{4np^2(x_{med})}.$$

где $p(x_{med})$ - плотность распределения результатов измерений в точке $x_{med}$. Следовательно, рациональный объем выборки имеет вид

$$n_{rat} = \frac{1}{4p^2(x_{med})\Delta^2}, \quad n_{rat} = \frac{1}{4p^2(x_{med})x_{med}^2\delta^2}$$

при ограничениях на абсолютные и относительные погрешности результатов измерений соответственно. Для практического использования этих формул следует оценить плотность распределения результатов измерений в одной точке - теоретической медиане. Это можно сделать с помощью тех или иных непараметрических оценок плотности [27].

Если результаты наблюдений имеют стандартное нормальное распределение с математическим ожиданием 0 и дисперсией 1, то

$$n_{rat} = \frac{\pi}{2\Delta^2} \approx \frac{1{,}57}{\Delta^2}.$$

В этом случае рациональный объем выборки в $\pi/2$ раз больше, чем для оценивания математического ожидания (пример 1 выше). Однако для других распределений рассматриваемое соотношение объемов может быть иным, в частности, меньше 1. Как вытекает из статьи А.Н.Колмогорова 1931 г. [29], рассматриваемое соотношение объемов может принимать любое значение между 0 и3.

**Пример 5. Оценивание коэффициента вариации.** Рассмотрим выборочный коэффициент вариации

$$v = f(y_1, y_2, ..., y_n) = \frac{\left\{\frac{1}{n-1}\sum_{1\le i\le n}(y_i - \bar{y})^2\right\}^{1/2}}{\frac{1}{n}\sum_{1\le i\le n} y_i} = \frac{s(y)}{\bar{y}}.$$

Как нетрудно подсчитать,

$$\frac{\partial f}{\partial x_i} = \frac{n\bar{x}(x_i - \bar{x}) - (n-1)s^2(x)}{n(n-1)(\bar{x})^2 s(x)}.$$

В случае ограничений на относительную погрешность

$$\lim_{n\to\infty} N_f(x) = \frac{\delta}{(M(x_1))^2\sigma} M\,|\,x_1\{[x_1 - M(x_1)]M(x_1) - \sigma^2\}\,|\,.$$

На основе этого предельного соотношения и формулы для асимптотической дисперсии выборочного коэффициента вариации, приведенной в [27], могут быть найдены по описанной выше схеме доверительные границы для теоретического коэффициента вариации и рациональный объем выборки.

***Замечание.*** Отметим, что формулы для рационального объема выборки получены на основе асимптотической теории, а применяются для получения конечных объемов – 36 и 100 в примерах 1-3. Как всегда при использовании асимптотических результатов математической статистики, необходимы дополнительные исследования для изучения точности асимптотических формул при конечных объемах выборок.

Рассмотрим классическую в прикладной математической статистике параметрическую задачу оценивания. Исходные данные – выборка $x1$ , $x2$ , ..., $xn$, состоящая из $n$ действительных чисел. В вероятностной модели простой случайной выборки ее элементы $x1$ , $x2$ , ..., $xn$ считаются набором реализаций $n$ независимых одинаково распределенных случайных величин. Будем считать, что эти величины имеют плотность $f(x)$. В параметрической статистической теории предполагается, что плотность $f(x)$ известна с точностью до конечномерного параметра, т.е., $f(x) = f(x,\theta_0)$ при некотором $\theta_0 \in \Theta \subseteq R^k$. Это, конечно, весьма сильное предположение, которое требует обоснования и проверки; однако в настоящее время параметрическая теория оценивания широко используется в различных прикладных областях.

Все результаты наблюдений определяются с некоторой точностью, в частности, записываются с помощью конечного числа значащих цифр (обычно 2 – 5). Следовательно, все реальные распределения результатов наблюдений дискретны. Обычно считают, что эти дискретные распределения достаточно хорошо приближаются непрерывными. Уточняя это утверждение, приходим к уже рассматривавшейся модели, согласно которой статистику доступны лишь величины

$$yj = xj + \varepsilon j \ , \ j = 1, 2, \ldots, n \ ,$$

где $x_i$ – «истинные» значения, $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ – погрешности наблюдений (включая погрешности дискретизации). В вероятностной модели принимаем, что $n$ пар

$$(x_1, \varepsilon_1), (x_2, \varepsilon_2), ..., (x_n, \varepsilon_n)$$

образуют простую случайную выборку из некоторого двумерного распределения, причем $x1$ , $x2$ , ..., $xn$ - выборка из распределения с плотностью $f(x) = f(x,\theta_0)$. Необходимо учитывать, что $x_i$ и $\varepsilon_i$ - реализации зависимых случайных величин (если считать их независимыми, то распределение $yi$ будет непрерывным, а не дискретным). Поскольку систематическую ошибку, как правило,

нельзя полностью исключить [26, с.141], то необходимо рассматривать случай $M\varepsilon_i \neq 0$. Нет оснований априори принимать и нормальность распределения погрешностей (согласно сводкам экспериментальных данных о разнообразии форм распределения погрешностей измерений, приведенным в [26, с.148] и [27, с.71-77], в подавляющем большинстве случаев гипотеза о нормальном распределении погрешностей оказалась неприемлемой для средств измерений различных типов). Таким образом, все три распространенных представления о свойствах погрешностей не адекватны реальности. Влияние погрешностей наблюдений на свойства статистических моделей необходимо изучать на основе иных моделей, а именно, моделей интервальной статистики.

Пусть $\varepsilon$ - характеристика величины погрешности, например, средняя квадратическая ошибка $\varepsilon = \sqrt{M(\varepsilon_i^2)}$. В классической математической статистике $\varepsilon$ считается пренебрежимо малой ($\varepsilon \to 0$) при фиксированном объеме выборки $n$. Общие результаты доказываются в асимптотике $n \to \infty$. Таким образом, в классической математической статистике сначала делается предельный переход $\varepsilon \to 0$, а затем предельный переход $n \to \infty$. В статистике интервальных данных принимаем, что объем выборки достаточно велик ($n \to \infty$), но всем измерениям соответствует одна и та же характеристика погрешности $\varepsilon \neq 0$. Полезные для анализа реальных данных предельные теоремы получаем при $\varepsilon \to 0$. В статистике интервальных данных сначала делается предельный переход $n \to \infty$, а затем предельный переход $\varepsilon \to 0$. Итак, в обеих теориях используются одни и те же два предельных перехода: $n \to \infty$ и $\varepsilon \to 0$, но в разном порядке. Утверждения обеих теорий принципиально различны.

В дальнейшем изложение идет на примере оценивания параметров гамма-распределения, хотя аналогичные результаты можно получить и для других параметрических семейств, а также для задач проверки гипотез (см. ниже) и т.д. Наша цель – продемонстрировать основные черты подхода статистики интервальных данных. Его разработка была стимулирована подготовкой ГОСТ 11.011-83 [4].

Отметим, что постановки статистики объектов нечисловой природы соответствуют подходу, принятому в общей теории устойчивости [3,27]. В соответствии с этим подходом выборке $x = (x_1, x_2, ..., x_n)$ ставится в соответствие множество допустимых отклонений $G(x)$, т.е. множество возможных значений вектора результатов наблюдений $y = (y_1, y_2, ..., y_n)$. Если известно, что абсолютная погрешность результатов измерений не превосходит $\Delta$, то множество допустимых отклонений имеет вид

$$G(x,\Delta) = \{y :| y_i - x_i |\le \Delta, i = 1,2,...,n\}.$$

Если известно, что относительная погрешность не превосходит $\delta$, то множество допустимых отклонений имеет вид

$$G(x,\delta) = \{y :| \frac{y_i}{x_i} - 1 |\le \delta, i = 1,2,...,n\}.$$

Теория устойчивости позволяет учесть «наихудшие» отклонения, т.е. приводит к выводам типа минимаксных, в то время как конкретные модели погрешностей позволяют делать заключения о поведении статистик «в среднем».

**Оценки параметров гамма-распределения.** Как известно, случайная величина X имеет гамма-распределение, если ее плотность такова [4]:

$$f(x;a,b) = \begin{cases} \dfrac{1}{\Gamma(a)} x^{a-1} b^{-a} \exp\{-\dfrac{x}{b}\}, x > 0, \\ 0, x \le 0, \end{cases}$$

где $a$ – параметр формы, $b$ – параметр масштаба, $\Gamma(a)$ - гамма-функция. Отметим, что есть и иные способы параметризации семейства гамма-распределений [30].

Поскольку $M(X) = ab$, $D(X) = ab^2$, то оценки метода имеют вид

$$\hat{a} = \frac{(\bar{x})^2}{s^2}, \quad \hat{b} = \frac{\bar{x}}{\hat{a}} = \frac{s^2}{\bar{x}},$$

где $\bar{x}$ - выборочное среднее арифметическое, а $s^2$ – выборочная дисперсия. Можно показать, что при больших $n$

$$M(\hat{a}-a)^2=\frac{2a(a+1)}{n},\quad M(\hat{b}-b)^2=\frac{b^2}{n}(2+\frac{3}{a}) \quad (11)$$

с точностью до бесконечно малых более высокого порядка.

Оценка максимального правдоподобия $a^*$ имеет вид [4]:

$$a^*=H(\frac{1}{n}\sum_{1\le i\le n}\ln(\frac{\bar{x}}{x_i})), \quad (12)$$

где $H(\bullet)$ - функция, обратная к функции

$$Q(a)=\ln a-\frac{d\Gamma(a)}{da}\Big/\Gamma(a).$$

При больших $n$ с точностью до бесконечно малых более высокого порядка

$$M(a^*-a)^2=\frac{a}{n(a\psi'(a)-1)},\quad \psi(a)=\frac{d\Gamma(a)}{da}\Big/\Gamma(a).$$

Как и для оценок метода моментов, оценка максимального правдоподобия $b^*$ параметра масштаба имеет вид

$$b^*=\bar{x}/a^*.$$

При больших $n$ с точностью до бесконечно малых более высокого порядка

$$M(b^*-b)^2=\frac{b^2\psi'(a)}{n(a\psi'(a)-1)}.$$

Используя свойства гамма-функции, можно показать [4], что при больших $a$

$$M(a^*-a)^2=\frac{a(2a-1)}{n},\quad M(b^*-b)^2=\frac{2b^2}{n}.$$

с точностью до бесконечно малых более высокого порядка. Сравнивая с формулами (11), убеждаемся в том, что средние квадраты ошибок для оценок метода моментов больше соответствующих средних квадратов ошибок для оценок максимального правдоподобия. Таким образом, с точки зрения классической математической статистики оценки максимального правдоподобия имеют преимущество по сравнению с оценками метода моментов.

**Необходимость учета погрешностей измерений.** Положим

$$v=f(x_1,x_2,...,x_n)=\frac{1}{n}\sum_{1\le i\le n}\ln\left(\frac{\bar{x}}{x_i}\right).$$

Из свойств функции $H(\bullet)$ следует [4, с.14], что при малых v

$$a^*\sim 1/(2v). \quad (13)$$

В силу состоятельности оценки максимального правдоподобия $a^*$ из формулы (13) следует, что $v\to 0$ по вероятности при $a\to\infty$.

Согласно модели статистики интервальных данных результатами наблюдений являются не $x_i$, а $y_i$, вместо $v$ по реальным данным рассчитывают

$$w=f(y_1,y_2,...,y_n)=\frac{1}{n}\sum_{1\le i\le n}\ln\left(\frac{\bar{y}}{y_i}\right).$$

Имеем

$$w-v=\ln\left(\frac{\bar{y}}{\bar{x}}\right)-\frac{1}{n}\sum_{1\le i\le n}\ln\left(1+\frac{\varepsilon_i}{x_i}\right). \quad (14)$$

В силу закона больших чисел при достаточно малой погрешности $\varepsilon$, обеспечивающей возможность приближения $\ln(1+\alpha)\sim\alpha$ для слагаемых в формуле (14), или, что эквивалентно, при достаточно малых предельной абсолютной погрешности $\Delta$ в формуле (1) или достаточно малой предельной относительной погрешности $\delta$ имеем при $n\to\infty$

$$w - v \to \frac{M(\varepsilon_i)}{M(x_i)} - M\left(\frac{\varepsilon_i}{x_i}\right) = c$$

по вероятности (в предположении, что все погрешности одинаково распределены). Таким образом, наличие погрешностей вносит сдвиг, вообще говоря, не исчезающий при росте объема выборки. Следовательно, если $c \neq 0,$ то оценка максимального правдоподобия не является состоятельной. Имеем

$$a^*(y) - a^* \approx -\frac{c}{2v^2},$$

где величина *a*(y)* определена по формуле (12) с заменой $x_i$ на $y_i$, $i$=1,2,…,$n$. Из формулы (13) следует [4], что

$$a^*(y) - a \approx -2(a^*)^2 c, \quad (15)$$

т.е. влияние погрешностей измерений увеличивается по мере роста $a$.

Из формул для $v$ и $w$ следует, что с точностью до бесконечно малых более высокого порядка

$$w - v \approx \sum_{1 \le i \le n} \frac{\partial f}{\partial x_i} \varepsilon_i = \frac{1}{n} \sum_{1 \le i \le n} \left(\frac{1}{\bar{x}} - \frac{1}{x_i}\right) \varepsilon_i. \quad (16)$$

С целью нахождения асимптотического распределения $w$ выделим, используя формулу (16) и формулу для $v$, главные члены в соответствующих слагаемых

$$w = \ln M(x_1) + \frac{1}{n} \sum_{1 \le i \le n} \left\{ \frac{x_i - M(x_1)}{M(x_i)} - \ln x_i + \left(\frac{1}{M(x_1)} - \frac{1}{x_i}\right) \varepsilon_i \right\} + O_p\left(\frac{1}{n}\right). \quad (17)$$

Таким образом, величина $w$ представлена в виде суммы независимых одинаково распределенных случайных величин (с точностью до зависящего от случая остаточного члена порядка $1/n$). В каждом слагаемом выделяются две части – одна, соответствующая Мб и вторая, в которую входят $\varepsilon_i$. На основе представления (17) можно показать, что при $n \to \infty, \varepsilon \to 0$ распределения случайных величин $v$ и $w$ асимптотически нормальны, причем

$$M(w) \approx M(v) + c, \quad D(w) \approx D(v).$$

Из асимптотического совпадения дисперсий $v$ и $w$, вида параметров асимптотического распределения (при $a \to \infty$) оценки максимального правдоподобия *a** и формулы (15) вытекает одно из основных соотношений статистики интервальных данных

$$M(a^*(y) - a)^2 \approx 4a^4 c^2 + \frac{a(2a-1)}{n}. \quad (18)$$

Соотношение (18) уточняет утверждение о несостоятельности *a**. Из него следует также, что не имеет смысла безгранично увеличивать объем выборки $n$ с целью повышения точности оценивания параметра $a$, поскольку при этом уменьшается только второе слагаемое в (18), а первое остается постоянным.

В соответствии с общим подходом статистики интервальных данных в стандарте [4] предлагается определять рациональный объем выборки $n_{rat}$ определять из условия «уравнивания погрешностей» (это условие было впервые предложено в монографии [3]) различных видов в формуле (18), т.е. из условия

$$4a^4 c^2 = \frac{a(2a-1)}{n_{rat}}.$$

Упрощая это уравнение в предположении $a \to \infty,$ получаем, что

$$n_{rat} = \frac{1}{2a^2 c^2}.$$

Согласно сказанному выше, целесообразно использовать лишь выборки с объемами $n \le n_{rat}$ . Превышение рационального объема выборки $n_{rat}$ не дает существенного повышения точности оценивания.

**Применение методов теории устойчивости.** Найдем асимптотическую нотну. Как следует из вида главного линейного члена в формуле (17), решение оптимизационной задачи

$$w - v \to \max, \quad |\varepsilon_i| \le \Delta,$$

соответствующей ограничениям на абсолютные погрешности, имеет вид

$$\varepsilon_i = \begin{cases} \Delta, & \frac{1}{\bar{x}} - \frac{1}{x_i} \ge 0, \\ -\Delta, & \frac{1}{\bar{x}} - \frac{1}{x_i} < 0 \end{cases}.$$

Однако при этом пары $(x_i, \varepsilon_i)$ не образуют простую случайную выборку, т.к. в выражения для $\varepsilon_i$ входит $\bar{x}$. Однако при $n \to \infty$ можно заменить $\bar{x}$ на $M(x_1)$. Тогда получаем, что

$$w - v \approx A\Delta$$

при $a>1$, где

$$A = M\left|\frac{1}{M(x_1)} - \frac{1}{x_1}\right| = \int_0^\infty \left|\frac{1}{ab} - \frac{1}{x}\right| f(x; a, b)dx.$$

Таким образом, с точностью до бесконечно малых более высокого порядка нотна имеет вид

$$N_{a^*}(y) = 2(a^*)^2 c, \quad c = A\Delta.$$

Применим полученные результаты к построению доверительных интервалов. В постановке классической математической статистики (т.е. при $\varepsilon = 0$) доверительный интервал для параметра формы $a$, соответствующий доверительной вероятности $\gamma$, имеет вид [4]

$$[a^* - u\left(\frac{1+\gamma}{2}\right)\sigma^*(a^*); \quad a^* + u\left(\frac{1+\gamma}{2}\right)\sigma^*(a^*)],$$

где $u\left(\frac{1+\gamma}{2}\right)$ - квантиль порядка $\frac{1+\gamma}{2}$ стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1,

$$[\sigma^*(a^*)]^2 = \frac{a^*}{n(a^*\psi'(a^*) - 1)}, \quad \psi(a) = \frac{d\Gamma(a)}{da} \Big/ \Gamma(a).$$

В постановке статистики интервальных данных (т.е. при $\varepsilon \ne 0$) следует рассматривать доверительный интервал

$$[a^* - 2(a^*)^2 |c| - u\left(\frac{1+\gamma}{2}\right)\sigma^*(a^*); \quad a^* + 2(a^*)^2 |c| + u\left(\frac{1+\gamma}{2}\right)\sigma^*(a^*)],$$

где

$$c = \frac{M(\varepsilon_i)}{M(x_i)} - M\left(\frac{\varepsilon_i}{x_i}\right)$$

в вероятностной постановке (пары $(x_i, \varepsilon_i)$ образуют простую случайную выборку) и $c = A\Delta$ в оптимизационной постановке. Как в вероятностной, так и в оптимизационной постановках длина доверительного интервала не стремится к 0 при $n \to \infty$.

Если ограничения наложены на предельную относительную погрешность, задана величина $\delta$, то значение $c$ можно найти с помощью следующих правил приближенных вычислений [32, с.142].

(I) Относительная погрешность суммы заключена между наибольшей и наименьшей из относительных погрешностей слагаемых.

(II) Относительная погрешность произведения и частного равна сумме относительных погрешностей сомножителей или, соответственно, делимого и делителя.

Можно показать, что в рамках статистики интервальных данных с ограничениями на относительную погрешность правила (I) и (II) являются строгими утверждениями при $\delta \to 0$.

Обозначим относительную погрешность некоторой величины $t$ через ОП($t$), абсолютную погрешность – через АП($t$).

Из правила (I) следует, что ОП($\bar{x}$) = $\delta$, а из правила (II) – что

$$ОП\left(\frac{\bar{x}}{x_i}\right) = 2\delta.$$

Поскольку рассмотрения ведутся при $a \to \infty$, то в силу неравенства Чебышева

$$\frac{\bar{x}}{x_i} \to 1 \qquad (19)$$

по вероятности при $a \to \infty$, поскольку и числитель, и знаменатель в (19) с близкой к 1 вероятностью лежат в промежутке $[ab - db\sqrt{a};\ \ ab + db\sqrt{a}]$, где константа d может быть определена с помощью упомянутого неравенства Чебышева.

Поскольку при справедливости (19) с точностью до бесконечно малых более высокого порядка

$$\ln\left(\frac{\bar{x}}{x_i}\right) \approx \frac{\bar{x}}{x_i} - 1,$$

то с помощью трех последних соотношений имеем

$$ОП\left(\frac{\bar{x}}{x_i}\right) = АП\left(\frac{\bar{x}}{x_i}\right) = АП\left(\ln\left(\frac{\bar{x}}{x_i}\right)\right) = 2\delta. \quad (20)$$

Применим еще одно правило приближенных вычислений [32, с.142].

(III) Предельная абсолютная погрешность суммы равна сумме предельных абсолютных погрешностей слагаемых.

Из (20) и правила (III) следует, что

$$АП(v) = 2\delta. \quad (21)$$

Из (15) и (21) вытекает [4, с.44, ф-ла (18)], что

$$АП(a^*) = 4a^2\delta,$$

откуда в соответствии с ранее полученной формулой для рационального объема выборки с заменой $c = 2\delta$ получаем, что

$$n_{rat} = \frac{1}{8a^2\delta^2}.$$

В частности, при $a$ = 5,00, $\delta$ = 0,01 получаем $n_{rat} = 50$, т.е. в ситуации, в которой были получены данные о наработке резцов до предельного состояния [4, с.29], проводить более 50 наблюдений нерационально.

В соответствии с ранее проведенными рассмотрениями асимптотический доверительный интервал для $a$, соответствующий доверительной вероятности $\gamma$ = 0,95, имеет вид

$$[a^* - 4(a^*)^2\delta - 1{,}96\sqrt{\frac{a^*(2a^*-1)}{n}};\quad a^* + 4(a^*)^2\delta + 1{,}96\sqrt{\frac{a^*(2a^*-1)}{n}}].$$

В частности, при $a^* = 5{,}00$, $\delta = 0{,}01$, $n = 50$ имеем асимптотический доверительный интервал [2,12; 7,86] вместо [3,14; 6,86] при $\delta = 0$.

При больших $a$ в силу соображений, приведенных при выводе формулы (19), можно связать между собой относительную и абсолютную погрешности результатов наблюдений $x_i$ :

$$\delta = \frac{\Delta}{M(x_1)} = \frac{\Delta}{ab}. \quad (21)$$

Следовательно, при больших $a$ имеем

$$c = 2\delta = A\Delta, \quad A = \frac{2\delta}{\Delta} = \frac{2}{ab}.$$

Таким образом, проведенные рассуждения дали возможность вычислить асимптотику интеграла, задающего величину $A$.

**Сравнение методов оценивания.** Изучим влияние погрешностей измерений (с ограничениями на абсолютную погрешность) на оценку $\hat{c}$ метода моментов. Имеем

$$АП(\bar{x}) = \Delta, \quad АП((\bar{x})^2) \approx 2\bar{x}\Delta \approx 2ab\Delta.$$

Погрешность $s^2$ зависит от способа вычисления $s^2$. Если используется формула

$$s^2 = \frac{1}{n-1}\sum_{1\le i\le n}(x_i - \bar{x})^2, \quad (22)$$

то необходимо использовать соотношения

$$АП(x_i - \bar{x})^2 = 2\Delta, \quad АП[(x_i - \bar{x})^2] \approx 2\,|\,x_i - \bar{x}\,|\,\Delta.$$

По сравнению с анализом влияния погрешностей на оценку а* здесь возникает новый момент – необходимость учета погрешностей в случайной составляющей отклонения оценки $\hat{a}$ от оцениваемого параметра, в то время как при рассмотрении оценки максимального правдоподобия погрешности давали лишь смещение. Примем в соответствии с неравенством Чебышева

$$|\,x_i - \bar{x}\,| \sim \sqrt{D(x_1)}, \quad (23)$$

тогда

$$АП[(x_i - \bar{x})^2] \sim 2b\sqrt{a}\Delta, \quad АП(s^2) \sim 2b\sqrt{a}\Delta.$$

Если вычислять $s^2$ по формуле

$$s^2 = \frac{1}{n-1}\sum_{1\le i\le n}x_i^2 - \frac{n}{n-1}(\bar{x})^2, \quad (24)$$

то аналогичные вычисления дают, что

$$АП(s^2) \sim 4ab\Delta,$$

т.е. погрешность при больших $a$ существенно больше. Хотя правые части формул (22) и (24) тождественно равны, но погрешности вычислений по этим формулам весьма отличаются. Связано это с тем, что в формуле (24) последняя операция – нахождение разности двух больших чисел, примерно равных по величине (для выборки из гамма-распределения при большом значении параметра формы).

Из полученных результатов следует, что

$$АП(\hat{a}) = АП\left(\frac{(\bar{x})^2}{s^2}\right) \sim \frac{2\Delta}{b}(1+\sqrt{a}).$$

При выводе этой формулы использована линеаризация влияния погрешностей (выделение главного линейного члена). Используя связь (21) между абсолютной и относительной погрешностями, можно записать

$$АП(\hat{a}) \sim 2a(1+\sqrt{a})\delta.$$

Эта формула отличается от приведенной в [4, с.44, ф-ла (19)]

$$АП(\hat{a}) \sim 2a(1+3\sqrt{a})\delta,$$

поскольку в [4] вместо (23) использовалась оценка

$$|\,x_i - \bar{x}\,| < 3\sqrt{D(x_1)}.$$

Используя соотношение (23), мы характеризуем влияние погрешностей «в среднем».

Доверительный интервал, соответствующий доверительной вероятности 0,95, имеет вид

$$[\hat{a} - 2\hat{a}(1+\sqrt{\hat{a}})\delta - 1{,}96\sqrt{\frac{2\hat{a}(\hat{a}+1)}{n}};\ \ \hat{a} + 2\hat{a}(1+\sqrt{\hat{a}})\delta + 1{,}96\sqrt{\frac{2\hat{a}(\hat{a}+1)}{n}}].$$

Если $\hat{a} = 5{,}00,\ \ \delta = 0{,}01,\ \ n = 50,$ то получаем доверительный интервал [2,54; 7,46] вместо [2,86; 7,14] при $\delta = 0$. Хотя при $\delta = 0$ доверительный интервал для $a$ при использовании оценки метода моментов $\hat{a}$ шире, чем при использовании оценки максимального правдоподобия $a^*$, при $\delta = 0{,}01$ результат сравнения длин интервалов противоположен.

Необходимо выбрать способ сравнения двух методов оценивания параметра $a$, поскольку в длины доверительных интервалов входят две составляющие – зависящая от доверительной

вероятности и не зависящая от нее. Выберем $\delta = 0{,}68$, т.е. $u\left(\frac{1+\gamma}{2}\right) = 1{,}00$. Тогда оценке максимального правдоподобия $a^*$ соответствует полудлина доверительного интервала

$$\nu(a^*) = 4a^2\delta + \sqrt{\frac{a(2a-1)}{n}}, \quad (25)$$

а оценке $\hat{a}$ метода моментов соответствует полудлина доверительного интервала

$$\nu(\hat{a}) = 2a(1+\sqrt{a})\delta + \sqrt{\frac{2a(a+1)}{n}}. \quad (26)$$

Ясно, что больших $a$ или больших $n$ справедливо неравенство $\nu(a^*) > \nu(\hat{a})$, т.е. метод моментов лучше метода максимального правдоподобия, вопреки классическим результатам Р.Фишера при $\delta = 0$ [33,с.99].

Из (25) и (26) элементарными преобразованиями получаем следующее правило принятия решений. Если

$$\delta\sqrt{n} \ge \frac{\sqrt{2a(a+1)} - \sqrt{a(2a-1)}}{4a^2 - 2a(1+\sqrt{a})} = B(a),$$

то $\nu(a^*) \ge \nu(\hat{a})$ и следует использовать $\hat{a}$; а если $\delta\sqrt{n} < B(a)$, то $\nu(a^*) < \nu(\hat{a})$ и надо применять $a^*$. Для выбора метода оценивания при обработке реальных данных целесообразно использовать $B(\hat{a})$ (см. раздел 5 в ГОСТ 11.011-83 [4, с.10-11]).

Пример анализа реальных данных опубликован в [4].

На основе рассмотрения проблем оценивания параметров гамма-распределения можно сделать некоторые общие выводы. Если в классической теории математической статистики:

а) существуют состоятельные оценки $a_n$ параметра $a$,

$$\lim_{n\to\infty} M(a_n - a)^2 = 0;$$

б) для повышения точности оценивания объем выборки целесообразно безгранично увеличивать;

в) оценки максимального правдоподобия лучше оценок метода моментов,

то в статистике интервальных данных, учитывающей погрешности измерений, соответственно:

а) не существует состоятельных оценок: для любой оценки $a_n$ существует константа $c$ такая, что

$$\lim_{n\to\infty} M(a_n - a)^2 \ge c > 0;$$

б) не имеет смысла рассматривать объемы выборок, большие «рационального объема выборки» $n_{rat}$;

в) оценки метода моментов в обширной области параметров $(a, n, \delta)$ лучше оценок максимального правдоподобия, в частности, при $a \to \infty$ и при $n \to \infty$.

Ясно, что приведенные выше результаты справедливы не только для рассмотренной задачи оценивания параметров гамма-распределения, но и для многих других постановок прикладной математической статистики.

**Метрологические, методические, статистические и вычислительные погрешности.** Целесообразно выделить ряд видов погрешностей статистических данных. Погрешности, вызванные неточностью измерения исходных данных, называем метрологическими. Их максимальное значение можно оценить с помощью нотны. Впрочем, выше на примере оценивания параметров гамма-распределения показано, что переход от максимального отклонения к реально имеющемуся в вероятностно-статистической модели не меняет выводы (с точностью до умножения предельных значений погрешностей $\Delta$ или $\delta$ на константы). Как правило, метрологические погрешности не убывают с ростом объема выборки.

Методические погрешности вызваны неадекватностью вероятностно-статистической модели, отклонением реальности от ее предпосылок. Неадекватность обычно не исчезает при росте объема выборки. Методические погрешности целесообразно изучать с помощью «общей

схемы устойчивости» [3,27], обобщающей популярную в теории робастных статистических процедур модель засорения большими выбросами. В настоящей главе методические погрешности не рассматриваются.

Статистическая погрешность – это та погрешность, которая традиционно рассматривается в математической статистике. Ее характеристики – дисперсия оценки, дополнение до 1 мощности критерия при фиксированной альтернативе и т.д. Как правило, статистическая погрешность стремится к 0 при росте объема выборки.

Вычислительная погрешность определяется алгоритмами расчета, в частности, правилами округления. На уровне чистой математики справедливо тождество правых частей формул (22) и (24), задающих выборочную дисперсию $s^2$, а на уровне вычислительной математики формула (22) дает при определенных условиях существенно больше верных значащих цифр, чем вторая [34, с.51-52].

Выше на примере задачи оценивания параметров гамма-распределения рассмотрено совместное действие метрологических и вычислительных погрешностей, причем погрешности вычислений оценивались по классическим правилам для ручного счета [32]. Оказалось, что при таком подходе оценки метода моментов имеют преимущество перед оценками максимального правдоподобия в обширной области изменения параметров. Однако, если учитывать только метрологические погрешности, как это делалось выше в примерах 1-5, то с помощью аналогичных выкладок можно показать, что оценки этих двух типов имеют (при достаточно больших $n$) одинаковую погрешность.

Вычислительную погрешность здесь подробно не рассматриваем. Ряд интересных результатов о ее роли в статистике получили Н.Н.Ляшенко и М.С.Никулин [35].

Проведем сравнение методов оценивания параметров в более общей постановке.

В теории оценивания параметров классической математической статистики установлено, что метод максимального правдоподобия, как правило, лучше (в смысле асимптотической дисперсии асимптотического среднего квадрата ошибки), чем метод моментов. Однако в интервальной статистике это, вообще говоря, не так, что продемонстрировано выше на примере оценивания параметров гамма-распределения. Сравним эти два метода оценивания в случае интервальных данных в общей постановке. Поскольку метод максимального правдоподобия – частный случай метода минимального контраста, начнем с разбора этого несколько более общего метода.

**Оценки минимального контраста.** Пусть $X$ – пространство, в котором лежат независимые одинаково распределенные случайные элементы $x_1, x_2, ..., x_n, ...$ Будем оценивать элемент пространства параметров $\Theta$ с помощью функции контраста $f : X \times \Theta \to R^1$. Оценкой минимального контраста называется

$$\theta_n = Arg\min\left\{ \sum_{1 \le i \le n} f(x_i, \theta), \theta \in \Theta \right\}.$$

Если множество $\theta_n$ состоит из более чем одного элемента, то оценкой минимального контраста называют также любой элемент $\theta_n$.

Оценками минимального контраста являются, в частности, многие робастные статистики [3,36]. Эти оценки широко используются в статистике объектов нечисловой природы [3,27], поскольку при $X = \Theta$ переходят в переходят в эмпирические средние, а если $X = \Theta$ - пространство бинарных отношений – в медиану Кемени.

Пусть в $X$ имеется мера $\mu$ (заданная на той же $\sigma$-алгебре, что участвует в определении случайных элементов $x_i$), и $p(x;\theta)$ - плотность распределения $x_i$ по мере $\mu$. Если

$$f(x;\theta) = -\ln p(x;\theta),$$

то оценка минимального контраста переходит в оценку максимального правдоподобия.

Асимптотическое поведение оценок минимального контраста в случае пространств $X$ и $\Theta$ общего вида хорошо изучено [37], в частности, известны условия состоятельности оценок. Здесь ограничимся случаем $X = R^1$, но при этом введя погрешности измерений $\varepsilon_i$. Примем также, что $\Theta = (\theta_{\min}, \theta_{\max}) \subseteq R^1$.

В рассматриваемой математической модели предполагается, что статистику известны лишь искаженные значения $y_i = x_i + \varepsilon_i,\ \ i = 1,2,...,n.$ Поэтому вместо $\theta_n$ он вычисляет

$$\theta_n^* = Arg\min\left\{\sum_{1\le i\le n} f(y_i,\theta), \theta\in\Theta\right\}.$$

Будем изучать величину $\theta_n^* - \theta_n$ в предположении, что погрешности измерений $\varepsilon_i$ малы. Цель этого изучения – продемонстрировать идеи статистики интервальных данных при достаточно простых предположениях. Поэтому естественно следовать условиям и ходу рассуждений, которые обычно принимаются при изучении оценок максимального правдоподобия [38, п.33.3].

Пусть $\theta_0$ - истинное значение параметра, функция $f(x;\theta)$ трижды дифференцируема по $\theta$, причем

$$\left|\frac{\partial^3 f(x;\theta)}{\partial\theta^3}\right| < H(x)$$

при всех $x,\theta.$ Тогда

$$\frac{\partial f(x;\theta)}{\partial\theta} = \frac{\partial f(x;\theta_0)}{\partial\theta} + \frac{\partial^2 f(x;\theta_0)}{\partial\theta^2}(\theta-\theta_0) + \frac{1}{2}\alpha(x)H(x)(\theta-\theta_0)^2, \quad (27)$$

где $|\alpha(x)|<1.$

Используя обозначения векторов $x = (x_1, x_2, ..., x_n),\ y = (y_1, y_2, ..., y_n),$ введем суммы

$$B_0(x) = \frac{1}{n}\sum_{1\le i\le n}\frac{\partial f(x_i;\theta_0)}{\partial\theta},\ \ B_1(x) = \frac{1}{n}\sum_{1\le i\le n}\frac{\partial^2 f(x_i;\theta_0)}{\partial\theta^2},\ \ R(x) = \frac{1}{n}\sum_{1\le i\le n}H(x_i).$$

Аналогичным образом введем функции $B_0(y),\ B_1(y),\ R(y),$ в которых вместо $x_i$ стоят $y_i$, $i$=1,2,…,$n$.

Поскольку в соответствии с теоремой Ферма оценка минимального контраста $\theta_n$ удовлетворяет уравнению

$$\sum_{1\le i\le n}\frac{\partial f(x_i;\theta_n)}{\partial\theta} = 0, \quad (28)$$

то, подставляя в (27) $x_i$ вместо $x$ и суммируя по $i$ = 1,2,…,$n$, получаем, что

$$0 = B_0(x) + B_1(x)(\theta_n-\theta_0) + \frac{\beta R(x)}{2}(\theta_n-\theta_0)^2, \quad |\beta|<1, \quad (29)$$

откуда

$$\theta_n - \theta_0 = \frac{-B_0(x)}{B_1(x) + \frac{\beta R(x)}{2}(\theta_n-\theta_0)}. \quad (30)$$

Решения уравнения (28) будем также называть оценками минимального контраста. Хотя уравнение (28) – лишь необходимое условие минимума, такое словоупотребление не будет вызывать трудностей.

**Теорема 1.** Пусть для любого $x$ выполнено соотношение (27). Пусть для случайной величины $x_1$ с распределением, соответствующим значению параметра $\theta = \theta_0,$ существуют математические ожидания

$$M\frac{\partial f(x_1,\theta_0)}{\partial\theta_0} = 0,\ \ M\frac{\partial^2 f(x_1,\theta_0)}{\partial\theta_0^2} = A \ne 0,\ \ MH(x_1) = M < +\infty. \quad (31)$$

Тогда существуют оценки минимального контраста $\theta_n$ такие, что $\theta_n \to \theta_0$ при $n\to\infty$ (в смысле сходимости по вероятности).

*Доказательство.* Возьмем $\varepsilon > 0$ и $\delta > 0.$ В силу закона больших чисел (теорема Хинчина) существует $n(\varepsilon,\delta)$ такое, что для любого $n > n(\varepsilon,\delta)$ справедливы неравенства

$$P\{|B_0| \ge \delta^2\} < \varepsilon/3, \quad P\{|B_1| < |A|/2\} < \varepsilon/3, \quad P\{R(x) > 2M\} < \varepsilon/3.$$

Тогда с вероятностью не менее $1-\varepsilon$ одновременно выполняются соотношения

$$|B_0| \le \delta^2, \quad |B_1| \ge |A|/2, \quad R(x) \le 2M. \quad (32)$$

При $\theta \in [\theta_0 - \delta; \theta_0 + \delta]$ рассмотрим многочлен второй степени

$$y(\theta) = B_0(x) + B_1(x)(\theta - \theta_0) + \frac{\beta R(x)}{2}(\theta - \theta_0)^2$$

(см. формулу (29)). С вероятностью не менее $1 - \varepsilon$ выполнены соотношения

$$|B_0 + \frac{\beta R(x)}{2}(\theta - \theta_0)^2| \le |B_0| + \frac{R(x)\delta^2}{2} \le \delta^2(M+1), \quad |B_1\delta| \ge \frac{|A|\delta}{2}.$$

Если $0 < 2(M+1)\delta < |A|$, то знак $y(\theta)$ в точках $\theta_1 = \theta_0 - \delta$ и $\theta_2 = \theta_0 + \delta$ определяется знаком линейного члена $B_1(\theta_i - \theta_0)$, $i = 1,2$, следовательно, знаки $y(\theta_1)$ и $y(\theta_2)$ различны, а потому существует $\theta_n \in [\theta_0 - \delta; \theta_0 + \delta]$ такое, что $y(\theta_n) = 0$, что и требовалось доказать.

**Теорема 2.** Пусть выполнены условия теоремы 2 и, кроме того, для случайной величины $x_1$, распределение которой соответствует значению параметра $\theta = \theta_0$, существует математическое ожидание

$$M\left(\frac{\partial f(x_1;\theta_0)}{\partial \theta_0}\right) = \sigma^2.$$

Тогда оценка минимального контраста имеет асимптотически нормальное распределение:

$$\lim_{n\to\infty} P\left\{\sqrt{n}\frac{|A|}{\sigma}(\theta_n - \theta_0) < x\right\} = \Phi(x) \quad (33)$$

для любого *x*, где *Φ(x)* – функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1.

*Доказательство.* Из центральной предельной теоремы вытекает, что числитель в правой части формулы (30) асимптотически нормален с математическим ожиданием 0 и дисперсией $\sigma^2$. Первое слагаемое в знаменателе формулы (30) в силу условий (31) и закона больших чисел сходится по вероятности к $A \ne 0$, а второе слагаемое по тем же основанием и с учетом теоремы 1 – к 0. Итак, знаменатель сходится по вероятности к $A \ne 0$. Доказательство теоремы 2 завершает ссылка на теорему о наследовании сходимости [3,параграф 2.4].

**Нотна оценки минимального контраста.** Аналогично (30) нетрудно получить, что

$$\theta_n^* - \theta_0 = \frac{-B_0(y)}{B_1(y) + \frac{\beta(y)R(y)}{2}(\theta_n^* - \theta_0)}, \quad |\beta(y)| < 1. \quad (34)$$

Следовательно, $\theta_n^* - \theta_n$ есть разность правых частей формул (30) и (34). Найдем максимально возможное значение (т.е. нотну) величины $|\theta_n^* - \theta_n|$ при ограничениях (1) на абсолютные погрешности результатов измерений.

Покажем, что при $\Delta \to 0$ для некоторого $C>0$ нотна имеет вид

$$N_{\theta_n}(x) = \sup_{\{\varepsilon\}} |\theta_n^* - \theta_n| = C\Delta(1 + o(1)). \quad (35)$$

Поскольку $\theta_n^* - \theta_n = (\theta_n^* - \theta_0) + (\theta_0 - \theta_n)$, то из (33) и (35) следует, что

$$\sup_{\{\varepsilon\}} M(\theta_n^* - \theta_n)^2 = \left(C^2\Delta^2 + \frac{\sigma^2}{A^2 n}\right)(1 + o(1)). \quad (36)$$

Можно сказать, что наличие погрешностей $\varepsilon_i$ приводит к появлению систематической ошибки (смещения) у оценки метода максимального правдоподобия, и нотна является максимально возможным значением этой систематической ошибки.

В правой части (36) первое слагаемое – квадрат асимптотической нотны, второе соответствует статистической ошибке. Приравнивая их, получаем рациональный объем выборки

$$n_{rat} = \left(\frac{\sigma}{CA\Delta}\right)^2.$$

Остается доказать соотношение (35) и вычислить $C$. Укажем сначала условия, при которых $\theta_n^* \to \theta_0$ (по вероятности) при $n \to \infty$ одновременно с $\Delta \to 0$.

**Теорема 3.** Пусть существуют константа $\Delta_0$ и функции $g_1(x)$, $g_2(x)$, $g_3(x)$ такие, что при $0 \le \Delta \le \Delta_0$ и $-1 \le \gamma \le 1$ выполнены неравенства (ср. формулу (27))

$$\left|\frac{\partial f(x;\theta_0)}{\partial\theta} - \frac{\partial f(x+\gamma\Delta;\theta_0)}{\partial\theta}\right| \le g_1(x)\Delta,$$

$$\left|\frac{\partial^2 f(x;\theta_0)}{\partial\theta^2} - \frac{\partial^2 f(x+\gamma\Delta;\theta_0)}{\partial\theta^2}\right| \le g_2(x)\Delta, \quad \ldots(37)$$

$$|H(x) - H(x+\gamma\Delta)| \le g_3(x)\Delta$$

при всех $x$. Пусть для случайной величины $x_1$, распределение которой соответствует $\theta = \theta_0$, существуют $m_1 = Mg_1(x_1)$, $m_2 = Mg_2(x_1)$ и $m_3 = Mg_3(x_1)$. Пусть выполнены условия теоремы 1. Тогда $\theta_n^* \to \theta_0$ (по вероятности) при $\Delta \to 0$, $n \to \infty$.

*Доказательство* проведем по схеме доказательства теоремы 1. Из неравенств (37) вытекает, что

$$|B_0(y) - B_0(x)| \le \Delta\left(\frac{1}{n}\sum_{1\le i\le n} g_1(x_i)\right),$$

$$|B_1(y) - B_1(x)| \le \Delta\left(\frac{1}{n}\sum_{1\le i\le n} g_2(x_i)\right), \quad (38)$$

$$|R(y) - R(x)| \le \Delta\left(\frac{1}{n}\sum_{1\le i\le n} g_3(x_i)\right).$$

Возьмем $\varepsilon > 0$ и $\delta > 0$. В силу закона больших чисел (теорема Хинчина) существует $n(\varepsilon,\delta)$ такое, что для любого $n > n(\varepsilon,\delta)$ справедливы неравенства

$$P\left\{|B_0| \ge \frac{\delta^2}{2}\right\} < \frac{\varepsilon}{6}, \quad P\left\{|B_1| < \frac{3|A|}{4}\right\} < \frac{\varepsilon}{6}, \quad P\left\{R(x) > \frac{3M}{2}\right\} < \frac{\varepsilon}{6},$$

$$P\left\{\frac{1}{n}\sum_{1\le i\le n} g_j(x_i) > 2m_j\right\} < \frac{\varepsilon}{6}, \quad j = 1,2,3.$$

Тогда с вероятностью не менее $1-\varepsilon$ одновременно выполняются соотношения

$$|B_0| < \frac{1}{2}\delta^2, \quad |B_1| \ge \frac{3|A|}{4}, \quad R(x) \le \frac{3M}{2}, \quad \frac{1}{n}\sum_{1\le i\le n} g_j(x_i) \le 2m_j, \quad j = 1,2,3.$$

В силу (38) при этом

$$|B_0(y)| < \frac{1}{2}\delta^2 + 2\Delta m_1, \quad |B_1(y)| \ge \frac{3|A|}{4} - 2\Delta m_2, \quad R(y) \le \frac{3M}{2} + 2\Delta m_3.$$

Пусть

$$0 \le \Delta \le \min\left\{\frac{1}{4}\frac{\delta^2}{m_1};\ \frac{1}{8}\frac{|A|}{m_2};\ \frac{1}{4}\frac{M}{m_3}\right\}.$$

Тогда с вероятностью не менее $1-\varepsilon$ одновременно выполняются соотношения (ср. (32))

$$|B_0(y)| \le \delta^2, \quad |B_1(y)| \ge |A|/2, \quad R(y) \le 2M.$$

Завершается доказательство дословным повторением такового в теореме 1, с единственным отличием – заменой в обозначениях $x$ на $y$.

**Теорема 4.** Пусть выполнены условия теоремы 3 и, кроме того, существуют математические ожидания (при $\theta = \theta_0$)

$$M\left|\frac{\partial^2 f(x_1,\theta_0)}{\partial x\partial\theta}\right|, \quad M\left|\frac{\partial^3 f(x_1,\theta_0)}{\partial x\partial\theta^2}\right|. \quad (39)$$

Тогда выполнено соотношение (35) с

$$C = \frac{1}{|A|} M \left| \frac{\partial^2 f(x_1, \theta_0)}{\partial x \partial \theta} \right|. \quad (40)$$

*Доказательство.* Воспользуемся следующим элементарным соотношением. Пусть $a$ и $b$ – бесконечно малые по сравнению с $Z$ и $B$ соответственно. Тогда с точностью до бесконечно малых более высокого порядка

$$\frac{Z+a}{B+b} - \frac{Z}{B} = \frac{aB - bZ}{B^2}.$$

Чтобы применить это соотношение к анализу $\theta_n^* - \theta_n$ в соответствии с (30), (34) и теоремой 2, положим

$$Z = B_0(x), \quad a = B_0(y) - B_0(x), \quad B = B_1(x), \quad b = (B_1(y) - B_1(x)) + \frac{\beta(y) R(y)}{2} (\theta_n^* - \theta_0).$$ В силу условий теоремы 4 при малых $\varepsilon_i$ с точностью до членов более высокого порядка

$$B_0(y) - B_0(x) = \frac{1}{n} \sum_{1 \le i \le n} \frac{\partial^2 f(x_i; \theta_0)}{\partial x_i \partial \theta_0} \varepsilon_i, \quad B_1(y) - B_1(x) = \frac{1}{n} \sum_{1 \le i \le n} \frac{\partial^3 f(x_i; \theta_0)}{\partial x_i \partial \theta_0^2} \varepsilon_i.$$

При $\Delta \to 0$ эти величины бесконечно малы, а потому с учетом сходимости $B_1(x)$ к A и теоремы 3

$$\theta_n^* - \theta_n = \frac{1}{A^2} \{ (B_0(y) - B_0(x)) A - (B_1(y) - B_1(x)) B_0(x) \} = \frac{1}{A^2 n} \sum_{1 \le i \le n} \gamma_i \varepsilon_i$$

с точностью до бесконечно малых более высокого порядка, где

$$\gamma_i = \frac{\partial^2 f(x_i; \theta_0)}{\partial x_i \partial \theta_0} A - \frac{\partial^3 f(x_i; \theta_0)}{\partial x_i \partial \theta_0^2} B_0(x).$$

Ясно, что задача оптимизации

$$\begin{cases} \sum\limits_{1 \le i \le n} \gamma_i \varepsilon_i \to \max \\ |\varepsilon_i| \le \Delta, \quad i = 1,2,...,n, \end{cases} \quad (41)$$

имеет решение

$$\varepsilon_i = \begin{cases} \Delta, & \gamma_i \ge 0, \\ -\Delta, & \gamma_i < 0, \end{cases}$$

при этом максимальное значение линейной формы есть $\Delta \sum\limits_{1 \le i \le n} |\gamma_i|$. Поэтому

$$\sup_{\{\varepsilon\}} |\theta_n^* - \theta_n| = \frac{\Delta}{A^2 n} \sum_{1 \le i \le n} |\gamma_i|. \quad (42)$$

С целью упрощения правой части (42) воспользуемся тем, что

$$\frac{1}{n} \sum_{1 \le i \le n} |\gamma_i| = \frac{|A|}{n} \sum_{1 \le i \le n} \left| \frac{\partial^2 f(x_i; \theta_0)}{\partial x \partial \theta_0} \right| + \alpha \frac{|B_0(x)|}{n} \sum_{1 \le i \le n} \left| \frac{\partial^3 f(x_i; \theta_0)}{\partial x \partial \theta_0^2} \right|, \quad (43)$$

где $|\alpha| \le 1$. Поскольку при $n \to \infty$

$$\frac{1}{n} \sum_{1 \le i \le n} \left| \frac{\partial^3 f(x_i; \theta_0)}{\partial x \partial \theta_0^2} \right| \to M \left| \frac{\partial^3 f(x_1; \theta_0)}{\partial x \partial \theta_0^2} \right| < +\infty, \quad B_0(x) \to 0$$

по вероятности, то второе слагаемое в (43) сходится к 0, а первое в силу закона больших чисел с учетом (39) сходится к $CA^2$, где $C$ определено в (40). Теорема 4 доказана.

**Оценки метода моментов.** Пусть $g: R^k \to R^1, \quad h_j: R^1 \to R^1, j = 1,2,...,k,$ - некоторые функции. Рассмотрим аналоги выборочных моментов

$$m_j = \frac{1}{n} \sum_{1 \le i \le n} h_j(x_i), j = 1,2,...,k.$$

Оценки метода моментов имеют вид

$$\hat{\theta}_n(x) = g(m_1, m_2, ..., m_k)$$

(функции $g$ и $h_j$ должны удовлетворять некоторым дополнительным условиям [39, с.80], которые здесь не приводим). Очевидно, что

$$\hat{\theta}_n(y) - \hat{\theta}_n(x) = \sum_{1 \le j \le k} \frac{\partial g}{\partial m_j}(m_j(y) - m_j(x)),$$

$$m_j(y) - m_j(x) = \frac{1}{n} \sum_{1 \le i \le n} \frac{dh_j(x_i)}{dx_i} \varepsilon_i, \quad j = 1,2,...,k, \quad (44)$$

с точностью до бесконечно малых более высокого порядка, а потому с той же точностью

$$\hat{\theta}_n(y) - \hat{\theta}_n(x) = \frac{1}{n} \sum_{1 \le i \le n} \left( \sum_{1 \le j \le k} \frac{\partial g}{\partial m_j} \frac{dh_j(x_i)}{dx_i} \right) \varepsilon_i. \quad (45)$$

**Теорема 5.** Пусть при $\theta = \theta_0$ существуют математические ожидания

$$M_j = Mm_j = Mh_j(x_1), \quad M\left( \frac{dh_j(x_1)}{dx_1} \right), \quad j = 1,2,...,n,$$

функция $g$ дважды непрерывно дифференцируема в некоторой окрестности точки $(M_1, M_2, ..., M_k)$. Пусть существует функция $t: R^1 \to R^1$ такая, что

$$\sup_{|x-y| \le \Delta} \left| h_j(y) - h_j(x) - \frac{dh_j(x)}{dx}(y - x) \right| \le t(x)\Delta^2, \quad j = 1,2,...,k, \quad (46)$$

причем $Mt(x_1)$ существует. Тогда

$$\sup_{\{\varepsilon\}} | \hat{\theta}_n(y) - \hat{\theta}_n(x) | = C_1 \Delta$$

с точностью до бесконечно малых более высокого порядка, причем

$$C_1 = M \left| \sum_{1 \le j \le k} \frac{\partial g(M_1, M_2, ..., M_k)}{\partial m_j} \frac{dh_j(x_1)}{dx_1} \right|.$$

*Доказательство* теоремы 5 сводится к обоснованию проведенных ранее рассуждений, позволивших получить формулу (45). В условиях теоремы 5 собраны предположения, достаточные для такого обоснования. Так, условие (46) дает возможность обосновать соотношения (44); существование $M\left( \frac{dh_j(x_1)}{dx_1} \right)$ обеспечивает существование $C_1$, и т.д. Завершает доказательство ссылка на решение задачи оптимизации (41) и применение закона больших чисел.

Полученные в теоремах 4 и 5 нотны оценок минимального контраста и метода моментов, асимптотические дисперсии этих оценок (см. теорему 2 и [40] соответственно) позволяют находить рациональные объемы выборок, строить доверительные интервалы с учетом погрешностей измерений, а также сравнивать оценки по среднему квадрату ошибки (36). Подобное сравнение было проведено для оценок максимального правдоподобия и метода моментов параметров гамма-распределения. Установлено, что классический вывод о преимуществе оценок максимального правдоподобия [33, с.99-100] неверен в случае $\Delta > 0$.

### 3.5.3. Интервальные данные в задачах проверки гипотез

С позиций статистики интервальных данных целесообразно изучить все практически используемые процедуры прикладной математической статистики, установить соответствующие нотны и рациональные объемы выборок. Это позволит устранить разрыв между математическими схемами прикладной статистики и реальностью влияния погрешностей наблюдений на свойства статистических процедур. Статистика интервальных данных – часть теории устойчивых статистических процедур, развитой в монографии [3]. Часть, более адекватная реальной статистической практике, чем некоторые другие постановки, например, с засорением нормального распределения большими выбросами.

Рассмотрим подходы статистики интервальных данных в задачах проверки статистических гипотез. Пусть принятие решения основано на сравнении рассчитанного по выборке значения статистики критерия $f = f(y_1, y_2, ..., y_n)$ с граничным значением $C$: если $f>C$, то гипотеза отвергается, если же $f \le C$, то принимается. С учетом погрешностей измерений выборочное значение статистики критерия может принимать любое значение в интервале $[f(y) - N_f(y); f(y) + N_f(y)]$. Это означает, что «истинное» значение порога, соответствующее реально используемому критерию, находится между $C\text{-}N_f(y)$ и $C+N_f(y)$, а потому уровень значимости описанного правила (критерия) лежит между $1 - P(C + N_f(y))$ и $1 - P(C - N_f(y))$, где $P(Z)=P(f<Z)$.

**Пример 1.** Пусть $x_1, x_2, ..., x_n$ - выборка из нормального распределения с математическим ожиданием $a$ и единичной дисперсией. Необходимо проверить гипотезу $H_0$: $a = 0$ при альтернативе $H_1 : a \neq 0$.

Как известно из любого учебного курса математической статистики, следует использовать следует использовать статистику $f = \sqrt{n}\,|\,\bar{y}\,|$ и порог $C = \Phi(1 - \alpha/2)$, где $\alpha$ - уровень значимости, $\Phi(\cdot)$ – функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. В частности, C = 1,96 при $\alpha = 0{,}05$.

При ограничениях (1) на абсолютную погрешность $N_f(y) = \sqrt{n}\Delta$. Например, если $\Delta = 0{,}1$, а $n$ = 100, то $N_f(y)$ = 1,0. Это означает, что истинное значение порога лежит между 0,96 и 2,96, а истинный уровень значимости – между 0,003 и 0,34. Можно сделать и другой вывод: нулевую гипотезу $H_0$ допустимо отклонить на уровне значимости 0,05 лишь тогда, когда $f > 2{,}96$.

Если же $n$ = 400 при $\Delta = 0{,}1$, то $N_f(y)$ =2,0 и $C\text{-}N_f(y)$ = -0,04, в то время как $C+N_f(y)$ =3,96. Таким образом, даже в случае $x = 0$ гипотеза $H_0$ может быть отвергнута только из-за погрешностей измерений результатов наблюдений.

Вернемся к общему случаю проверки гипотез. С учетом погрешностей измерений граничное значение $C_\alpha$ в статистике интервальных данных целесообразно заменить на $C_\alpha + N_f(y)$. Такая замена дает гарантию, что вероятность отклонения нулевой гипотезы $H_0$, когда она верна, не более $\alpha$. При проверке гипотез аналогом статистической погрешности, рассмотренной выше в задачах оценивания, является $C_\alpha$. Суммарная погрешность имеет вид $C_\alpha + N_f(y)$. Исходя из принципа уравнивания погрешностей [3], целесообразно определять рациональный объем выборки из условия

$$C_\alpha = N_f(y).$$

Если $f = |f_1|$, где $f_1$ при справедливости $H_0$ имеет асимптотически нормальное распределение с математическим ожиданием 0 и дисперсией $\sigma^2 / n$, то

$$C_\alpha = u\left(1 - \frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} \quad (47)$$

при больших n, где $u(1-\alpha/2)$ - квантиль порядка $1-\alpha/2$ стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Из (47) вытекает, что в рассматриваемом случае

$$n_{rat} = \left[\frac{u(1-\alpha/2)\sigma}{N_f(y)}\right]^2.$$

В условиях примера 1 $f_1 = \bar{y}$ и

$$n_{rat} = \frac{3{,}84}{\Delta^2} = 384.$$

**Пример 2.** Рассмотрим статистику одновыборочного критерия Стьюдента

$$t = \sqrt{n}\,\frac{\bar{y}}{s(y)} = \frac{\sqrt{n}}{v},$$

где v – выборочный коэффициент вариации. Тогда с точностью до бесконечно малых более высокого порядка нотна для *t* имеет вид

$$N_t(y) = \frac{\sqrt{n}}{v^2} N_v(y),$$

где $N_v(y)$ – рассмотренная ранее нотна для выборочного коэффициента вариации. Поскольку распределение статистики Стьюдента t сходится к стандартному нормальному, то небольшое изменение предыдущих рассуждений дает

$$n_{rat} = \frac{v^4 u^2 (1-\alpha/2)}{N_v^2(y)}.$$

**Пример 3.** Рассмотрим двухвыборочный критерий Смирнова, предназначенный для проверки однородности (совпадения) функций распределения двух независимых выборок [41]. Статистика этого критерия имеет вид

$$D_{mn} = \sup_x |F_m(x) - G_n(x)|,$$

где $F_m(x)$ – эмпирическая функция распределения, построенная по первой выборке объема *m*, извлеченной из генеральной совокупности с функцией распределения *F(x),* а $G_n(x)$ – эмпирическая функция распределения, построенная по второй выборке объема *n,* извлеченной из генеральной совокупности с функцией распределения G(x). Нулевая гипотеза имеет вид $H_0 : F(x) \equiv G(x),$ альтернативная состоит в ее отрицании: $H_1 : F(x) \neq G(x)$ *при некотором x*. Значение статистики сравнивают с порогом $D(\alpha, m, n),$ зависящим от уровня значимости $\alpha$ и объемов выборок *m* и *n*. Если значение статистики не превосходит порога, то принимают нулевую гипотезу, если больше порога – альтернативную. Пороговые значения $D(\alpha, m, n)$ берут из таблиц [42]. Описанный критерий иногда неправильно называют критерием Колмогорова-Смирнова. История вопроса описана в [43].

При ограничениях (1) на абсолютные погрешности и справедливости нулевой гипотезы $H_0 : F(x) \equiv G(x)$ нотна имеет вид (при больших объемах выборок)

$$N_D = \sup_x | F(x+\Delta) - F(x-\Delta) | .$$

Если $F(x)=G(x)=x$ при $0 \le x \le 1$, то $N_D = 2\Delta.$ С помощью условия $C_\alpha = N_f(y)$ при уровне значимости $\alpha = 0{,}05$ и достаточно больших объемах выборок (т.е. используя асимптотическое выражение для порога согласно [42]) получаем, что выборки имеет смысл увеличивать, если

$$\frac{mn}{m+n} \le \frac{0{,}46}{\Delta^2}.$$

Правая часть этой формулы при $\Delta = 0{,}1$ равна 46. Если $m = n$, то последнее неравенство переходит в $n \le 92$.

Теоретические результаты в области статистических методов входят в практику через алгоритмы расчетов, воплощенные в программные средства (пакеты программ, диалоговые системы). Ввод данных в современном статистической программной системе должен содержать запросы о погрешностях результатов измерений. На основе ответов на эти запросы вычисляются нотны рассматриваемых статистик, а затем – доверительные интервалы при оценивании, разброс уровней значимости при проверке гипотез, рациональные объемы выборок. Необходимо использовать систему алгоритмов и программ статистики интервальных данных, «параллельную» подобным системам для классической математической статистики.

### 3.5.4. Линейный регрессионный анализ интервальных данных

Перейдем к многомерному статистическому анализу. Сначала с позиций асимптотической математической статистики интервальных данных рассмотрим оценки метода наименьших квадратов (МНК).

Статистическое исследование зависимостей - одна из наиболее важных задач, которые возникают в различных областях науки и техники. Под словами "исследование зависимостей"

имеется в виду выявление и описание существующей связи между исследуемыми переменными на основании результатов статистических наблюдений. К методам исследования зависимостей относятся регрессионный анализ, многомерное шкалирование, идентификация параметров динамических объектов, факторный анализ, дисперсионный анализ, корреляционный анализ и др. Однако многие реальные ситуации характеризуются наличием данных интервального типа, причем известны допустимые границы погрешностей (например, из технических паспортов средств измерения).

Если какая-либо группа объектов характеризуется переменными $X_1, X_2, ...,X_m$ и проведен эксперимент, состоящий из $n$ опытов, где в каждом опыте эти переменные измеряются один раз, то экспериментатор получает набор чисел: $X_{1j}, X_{2j}, ...,X_{mj}$ $(j = 1,..., n)$.

Однако процесс измерения, какой бы физической природы он ни был, обычно не дает однозначный результат. Реально результатом измерения какой-либо величины $X$ являются два числа: $X_H$— нижняя граница и $X_B$ — верхняя граница. Причем $X_{ИСТ} \in [X_H, X_B]$, где $X_{ИСТ}$ - истинное значение измеряемой величины. Результат измерения можно записать как $X: [X_H, X_B]$. Интервальное число X может быть представлено другим способом, а именно, $X: [X_m, Д_x]$, где $X_H = X_m - Д_x$ , $X_H = X_m + Д_x$ . Здесь $X_m$ - центр интервала (как правило, не совпадающий с $X_{ИСТ}$), а $Д_x$ - максимально возможная погрешность измерения.

**Метод наименьших квадратов для интервальных данных.** Пусть математическая модель задана следующим образом:

$$y = Q(x,b) + e,$$

где $x = (x_1 , x_2,..., x_m)$ - вектор влияющих переменных (факторов), поддающихся измерению; $b = (b_1 , b_2 , ... , b_r )$ - вектор оцениваемых параметров модели; $y$ - отклик модели (скаляр); $Q(x,b)$- скалярная функция векторов $x$ и $b$; наконец, $e$ - случайная ошибка (невязка, погрешность).

Пусть проведено $n$ опытов, причем в каждом опыте измерены (один раз) значения отклика ($y$) и вектора факторов ($x$). Результаты измерений могут быть представлены в следующем виде:

$$X = \{ x_{ij} ; i = 1,n ; j = 1,m \}, Y = (y_1 , y_2 , ..., y_n), E = (e_1,e_2,...,e_n),$$

где $X$ - матрица значений измеренного вектора ($x$) в $n$ опытах; $Y$ - вектор значений измеренного отклика в $n$ опытах; $E$ - вектор случайных ошибок. Тогда выполняется матричное соотношение:

$$Y = Q(X,b) + E ,$$

где $Q(X,b) = (Q(x_1 ,b), Q(x_2 ,b), ..., Q(x_n ,b))^T$, причем $x_1 , x_2 , ..., x_n$ - $m$-мерные вектора, которые составляют матрицу $X = (x_1 , x_2 , ..., x_n )^T$.

Введем меру близости $d(Y,Q)$ между векторами $Y$ и $Q$. В МНК в качестве $d(Y,Q)$ берется квадратичная форма взвешенных квадратов $e_i^2$ невязок $e_i = y_i - Q(x_i ,b)$, т.е.

$$d(Y,Q) = [ Y - Q(X,b)]^T W[Y - Q(X,b)],$$

где $W = \{w_{ij} , i, j = 1,..., n\}$ - матрица весов, не зависящая от $b$. Тогда в качестве оценки $b$ можно выбрать такое $b^*$, при котором мера близости $d(Y,Q)$ принимает минимальное значение, т.е.

$$b^* = \{b : d(Y,Q) \to \min_{\{b\}}\}.$$

В общем случае решение этой экстремальной задачи может быть не единственным. Поэтому в дальнейшем будем иметь в виду одно из этих решений. Оно может быть выражено в виде $b^* = f(X,Y)$, где $f(X,Y) = (f_1(X,Y), f_2(X,Y),..., f_m(X.Y))^T$, причем $f_i(X,Y)$ непрерывны и дифференцируемы по $(X,Y) \in Z$, где $Z$ -область определения функции $f(X,Y)$. Эти свойства функции $f(X,Y)$ дают возможность использовать подходы статистики интервальных данных.

Преимущество метода наименьших квадратов заключается в сравнительной простоте и универсальности вычислительных процедур. Однако не всегда оценка МНК является состоятельной (при функции $Q(X,b)$, не являющейся линейной по векторному параметру $b$), что ограничивает его применение на практике.

Важным частным случаем является линейный МНК, когда $Q(x,b)$ есть линейная функция от $b$:

$$y = b_o x_o + b_1 x_1 + ... + b_m x_m + e = b x^T + e ,$$

где, возможно, $x_o = 1$, а $b_o$ - свободный член линейной комбинации. Как известно, в этом случае МНК-оценка имеет вид:

$$b^* = (X^T WX)^{-1} X^T WY.$$

Если матрица $X^T WX$ не вырождена, то эта оценка является единственной. Если матрица весов $W$

единичная, то

$$b^* = (X^T X)^{-1} X^T Y.$$

Пусть выполняются следующие предположения относительно распределения ошибок $e_i$:
- ошибки $e_i$ имеют нулевые математические ожидания $M\{e_i\} = 0$,
- результаты наблюдений имеют одинаковую дисперсию $D\{e_i\} = y^2$,
- ошибки наблюдений некоррелированы, т.е. $cov\{e_i, e_j\} = 0$.

Тогда, как известно, оценки МНК являются наилучшими линейными оценками, т.е. состоятельными и несмещенными оценками, которые представляют собой линейные функции результатов наблюдений и обладают минимальными дисперсиями среди множества всех линейных несмещенных оценок. Далее именно этот наиболее практически важный частный случай рассмотрим более подробно.

Как и в других постановках асимптотической математической статистики интервальных данных, при использовании МНК измеренные величины отличаются от истинных значений из-за наличия погрешностей измерения. Запишем истинные данные в следующей форме:

$X_R = \{ x_{ij}^R ; i = \overline{1, n} ; j = \overline{1, m} \}, \; Y_R = ( y_1^R , y_2^R ,..., y_n^R )$,

где $R$ - индекс, указывающий на то, что значение истинное. Истинные и измеренные данные связаны следующим образом:

$$X = X_R + \Delta X, \; Y = Y_R + \Delta Y,$$

где $\Delta X = \{\Delta x_{ij}; i = \overline{1,n}; j = \overline{1,m}\}, \Delta Y = (\Delta y_1, \Delta y_2, ..., \Delta y_n)$. Предположим, что погрешности измерения отвечают граничным условиям

$$|\Delta x_{ij}| \le \Delta_j^x \; (i = 1,2,...,n, j = 1,2,...,m), \; ,|\Delta y_i| \le \Delta^y \; (i = 1,2,...,n), \quad (48)$$

аналогичным ограничениям (1).

Пусть множество $W$ возможных значений $(X_R , Y_R)$ входит в $Z$-область определения функции $f(X,Y)$. Рассмотрим $b^{*R}$ - оценку МНК, рассчитанную по истинным значениям факторов и отклика, и $b^*$ - оценку МНК, найденную по искаженным погрешностями данным. Тогда

$$\Delta b^* = b^{*R} - b^* = f(X_R, Y_R) - f(X, Y).$$

Ввести понятие *нотны* придется несколько иначе, чем это было сделано выше, поскольку оценивается не одномерный параметр, а вектор. Положим:

$$n(1) = (\sup \Delta b_1^*, \sup \Delta b_2^*, ..., \sup \Delta b_r^*)^T, \; n(2) = -(\inf \Delta b_1^*, \inf \Delta b_2^*, ..., \inf \Delta b_r^*)^T.$$

Будем называть $n(1)$ нижней *нотной*, а $n(2)$ верхней *нотной*. Предположим, что при безграничном возрастании числа измерений $n$, т.е. при $n \to \infty$, вектора $n(1), n(2)$ стремятся к постоянным значениям $N_i(1), N_i(2)$ соответственно. Тогда $N_i(1)$ будем называть нижней асимптотической *нотной*, а $N_i(2)$ - верхней асимптотической *нотной*.

Рассмотрим доверительное множество $B_б = B_б(n, b^{*R})$ для вектора параметров $b$, т.е. замкнутое связное множество точек в $r$-мерном евклидовом пространстве такое, что $P(b \in B_\alpha) = \alpha$, где б — доверительная вероятность, соответствующая $B_б$ (б ≈ 1). Другими словами, $B_б (n, b^{*R})$ есть область рассеивания (аналог эллипсоида рассеивания) случайного вектора $b^{*R}$ с доверительной вероятностью б и числом опытов $n$.

Из определения верхней и нижней *нотн* следует, что всегда $b^{*R} \in [b^* - n(1); b^* + n(2)]$. В соответствии с определением нижней асимптотической нотны и верхней асимптотической нотны можно считать, что $b^{*R} \in [b^* - N(1); b^* + N(2)]$ при достаточно большом числе наблюдений $n$. Этот многомерный интервал описывает $r$-мерный гиперпараллелепипед $P$.

Каким-либо образом разобьем $P$ на $L$ гиперпараллелепипедов. Пусть $b_k$ - внутренняя точка $k$-го гиперпараллелепипеда. Учитывая свойства доверительного множества и устремляя $L$ к бесконечности, можно утверждать, что $P(b \in C) \ge \alpha$, где

$$C = \lim_{L \to \infty} \bigcup_{1 \le k \le L} B_\alpha (n, b_k).$$

Таким образом, множество $C$ характеризует неопределенность при оценивании вектора параметров $b$. Его можно назвать доверительным множеством в статистике интервальных данных.

Введем некоторую меру *M(X),* характеризующую «величину» множества $X \subseteq R^r$ По определению меры она удовлетворяет условию: если $X = Z \bigcup Y$ и $Z \bigcap Y = 0,$ то *M(X)=M(Z)+M(Y).* Примерами такой меры являются площадь для r = 2 и объем для r = 3. Тогда:

$$M(C) = M(P) + M(F), \qquad (49)$$

где $F = C \setminus P$. Здесь *M(F)* характеризует меру статистической неопределенности, в большинстве случаев она убывает при увеличении числа опытов *n*. В то же время *M(P)* характеризует меру интервальной (метрологической) неопределенности, и, как правило, *M(P)* стремится к некоторой постоянной величине при увеличении числа опытов *n.* Пусть теперь требуется найти то число опытов, при котором статистическая неопределенность составляет *д-ю* часть общей неопределенности, т.е.

$$M(F) = \partial\, M(C), \qquad (50)$$

где $\partial < 1$. Тогда, подставив соотношение (50) в равенство (49) и решив уравнение относительно *n*, получим искомое число опытов. В асимптотической математической статистике интервальных данных оно называется "рациональным объемом выборки". При этом $\partial$ есть "степень малости" статистической неопределенности *M(P)* относительно всей неопределенности. Она выбирается из практических соображений. При использовании "принципа уравнивания погрешностей" согласно [3] имеем $\partial = 1/2$.

**Метод наименьших квадратов для линейной модели.** Рассмотрим наиболее важный для практики частный случай МНК, когда модель описывается линейным уравнением (см. выше).

Для простоты описания преобразований пронормируем переменные $x_{ij}, y_i$. следующим образом:

$$x_{ij}^0 = (x_{ij} - \bar{x}_j)/s(x_j), \;\; y_i^0 = (y_i - \bar{y})/s(y),$$

где

$$\bar{x}_j = \frac{1}{n}\sum_{1\le i\le n} x_{ij}, \;\; s^2(x_j) = \frac{1}{n}\sum_{1\le i\le n}(x_{ij} - \bar{x}_j)^2, \quad \bar{y} = \frac{1}{n}\sum_{1\le i\le n} y_i, \;\; s^2(y) = \frac{1}{n}\sum_{1\le i\le n}(y_i - \bar{y})^2.$$

Тогда

$$\bar{x}_j^0 = 0 \;\; s^2(x_j^0) = \frac{1}{n}\sum_{1\le i\le n}(x_{ij}^0 - \bar{x}_j^0)^2 = 1, \quad \bar{y}^0 = 0 \;\; s^2(y^0) = \frac{1}{n}\sum_{1\le i\le n}(y_i^0 - \bar{y}^0)^2 = 1, \;\; j = 1,2,...,m..$$

В дальнейшем изложении будем считать, что рассматриваемые переменные пронормированы описанным образом, и верхние индексы $^0$ опустим. Для облегчения демонстрации основных идей примем достаточно естественные предположения.

1. Для рассматриваемых переменных существуют следующие пределы:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{1\le i\le n} x_{ij}x_{ik} = 0, \;\; j,k = 1,2,...,m.$$

2. Количество опытов *n* таково, что можно пользоваться асимптотическими результатами, полученными при $n \to \infty$.

3. Погрешности измерения удовлетворяют одному из следующих типов ограничений:

*Тип 1*. Абсолютные погрешности измерения ограничены согласно (48):

*Тип 2*. Относительные погрешности измерения ограничены:

$$|\Delta x_{ij}| \le \delta_j^x |x_{ij}| \;\; (i = 1,2,...,n, j = 1,2,...,m), \;\; ,|\Delta y_i| \le \delta^y |y_i| \;\; (i = 1,2,...,n).$$

*Тип 3.* Ограничения наложены на сумму погрешностей:

$$\sum_{j=1}^{m} |\Delta x_{ij}| \le \alpha_x \;\; (i = 1,2,..., n, j = 1,2,..., m), \;\; |\Delta y_i| \le \alpha_y \;\; (i = 1,2,..., n)).$$

(поскольку все переменные отнормированы, т.е. представляют собой относительные величины, то различие в размерности исходных переменных не влияет на возможность сложения погрешностей).

Перейдем к вычислению нотны оценки МНК. Справедливо равенство:

$$\Delta b^* = b^{*R} - b^* = (X_R^T X_R)^{-1} X_R^T Y_R - (X^T X)^{-1} X^T Y = (X_R^T X_R)^{-1} X_R^T Y_R -$$

$$- ((X_R + \Delta X)^T (X_R + \Delta X))^{-1} (X_R + \Delta X)(Y_R + \Delta Y).$$

Воспользуемся следующей теоремой из теории матриц [14].

**Теорема**. Если функция f(л) разлагается в степенной ряд в круге сходимости |л – л$_0$| < r, т.е.

$$f(\lambda)=\sum_{k=0}^{\infty}\alpha_k(\lambda-\lambda_0)^k,$$

то это разложение сохраняет силу, если скалярный аргумент заменить любой матрицей $A$, характеристические числа которой $л_k$, $k = 1,\ldots,n$, лежат внутри круга сходимости.

Из этой теоремы вытекает, что:

$$(E-A)^{-1}=\sum_{P=0}^{\infty}A^P, \quad если \quad /\lambda_k/<1;\ k=1,...,n.$$

Легко убедиться, что:

$$((X_R+\Delta X)^T(X_R+\Delta X))^{-1}=-Z(E-\Delta\cdot Z)^{-1},$$

$$где\quad Z=-(X_R^TX_R)^{-1},\quad \Delta=X_R^T\Delta X+\Delta X^TX_R+\Delta X^T\Delta X.$$

Это вытекает из последовательности равенств:

$$((X_R+\Delta X)^T(X_R+\Delta X))^{-1}=(X_R^TX_R+X_R^T\Delta X+\Delta X^TX_R+\Delta X^T\Delta X)^{-1}=(X_R^TX_R+\Delta)^{-1}=$$

$$=((E+\Delta(X_R^TX_R)^{-1}X_R^TX_R)^{-1}=(X_R^TX_R)^{-1}(E+\Delta(X_R^TX_R)^{-1})^{-1}=-Z(E-\Delta\cdot Z)^{-1}.$$

Применим приведенную выше теорему из теории матриц, полагая $A$ = Д $Z$ и принимая, что собственные числа этой матрицы удовлетворяют неравенству |л$_k$|<1. Тогда получим:

$$((X_R+\Delta X)^T(X_R+\Delta X))^{-1}=-Z\sum_{P=0}^{\infty}(\Delta\cdot Z)^P=(X^T{}_RX_R)^{-1}\sum_{P=0}^{\infty}(-\Delta\cdot(X^T{}_RX_R)^{-1})^P.$$

Подставив последнее соотношение в заключение упомянутой теоремы, получим:

$$\Delta b^*=(X_R^TX_R)^{-1}X_R^TY_R-((X_R^TX_R)^{-1}\sum_{P}^{\infty}(-\Delta\cdot(X_R^TX_R)^{-1})^P)(X_R+\Delta X)^T(Y_R+\Delta Y)=$$

$$=(X_R^TX_R)^{-1}X_R^TY_R-((X_R^TX_R)^{-1}\sum_{P}^{\infty}(-\Delta\cdot(X_R^TX_R)^{-1})^P)(X_R^TY_R+\Delta X^TY_R+X_R^T\Delta Y+\Delta X^T\Delta Y).$$

Для дальнейшего анализа понадобится вспомогательное утверждение. Исходя из предположений 1-3, докажем, что:

$$(X_R^TX_R)^{-1}\approx\frac{1}{n}E.$$

*Доказательство*. Справедливо равенство

$$X_R^TX_R=n\cdot\begin{pmatrix}\hat{D}(x_1) & \ldots & \hat{\mathrm{cov}}(x_1,x_m)\\ \ldots & \ldots & \ldots\\ \hat{\mathrm{cov}}(x_1,x_m) & \ldots & \hat{D}(x_m)\end{pmatrix}=n\cdot\hat{\mathrm{cov}}(x),$$

где $\hat{D}(x_i)$, $\hat{\mathrm{cov}}(x_i,x_j)$ - состоятельные и несмещенные оценки дисперсий и коэффициентов ковариации, т.е.

$$\hat{\mathrm{D}}(x_i)=D(x_i)+o(1/n),\ \hat{\mathrm{cov}}(x_i,x_j)=\mathrm{cov}(x_i,x_j)+o(1/n),$$

тогда

$$X_R^TX_R=n\cdot\hat{\mathrm{cov}}(x)=n\cdot(\mathrm{cov}(x_i,x_j)+o(1/n)),$$

где

$$o(1/n)=\{a_{ij}=o(1/n)\}\ \ (i=\overline{1,n},j=\overline{1,m}).$$

Другими словами, каждый элемент матрицы, обозначенной как $o(1/n)$, есть бесконечно малая величина порядка $1/n$. Для рассматриваемого случая $cov(x) = E$, поэтому

$$X_R^T X_R = n \cdot \hat{\mathrm{cov}}(x) = n \cdot (E + o(1/n)).$$

Предположим, что *n* достаточно велико и можно считать, что собственные числа матрицы *o(1/n)* меньше единицы по модулю, тогда

$$(X_R^T X_R)^{-1} = \frac{1}{n} \cdot (E + o(1/n))^{-1} \approx \frac{1}{n}(E + o(1/n)) = \frac{1}{n}E + o(1/n^2) \approx \frac{1}{n}E,$$

что и требовалось доказать.

Подставим доказанное асимптотическое соотношение в формулу для приращения *b**, получим

$$\Delta b^* = b^{*R} - \frac{1}{n}\sum_{P=0}^{\infty}(-\Delta \cdot \frac{1}{n})^P (nb^{*R} + \Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y) =$$

$$= b^{*R} - \frac{1}{n}\sum_{P=0}^{\infty}(-(X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X) \cdot (\frac{1}{n})^P (nb^{*R} + \Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y) =$$

Выразим

$$= b^{*R} - \frac{1}{n}(E - (X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X)\frac{1}{n} + (X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X)^2 (\frac{1}{n})^2) \cdot$$

$$\cdot (nb^{*R} + \Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y).$$

*Дb** относительно приращений *ДХ, ДУ* до 2-го порядка

$$\Delta b^* = b^{*R} - \frac{1}{n}(E - (X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X)\frac{1}{n} + (X_R^T \Delta X X_R^T \Delta X + \Delta X^T X_R \Delta X^T X_R +$$

$$+ \Delta X^T X_R X_R^T \Delta X + X_R^T \Delta X \Delta X^T X_R)(\frac{1}{n})^2) \cdot (nb^{*R} + \Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y);$$

$$\Delta b^* = b^{*R} - \frac{1}{n}(E - (X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X)\frac{1}{n}) \cdot (nb^{*R} + \Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y);$$ Перейдем от

$$\Delta b^* = \frac{1}{n}(X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X) b^{*R} - \frac{1}{n}(\Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y) =$$

$$= \frac{1}{n}[(X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X) b^{*R} - (\Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y)].$$

матричной к скалярной форме, опуская индекс (*R*):

$$\Delta b_k^* = \frac{1}{n}\{\sum_j^m \sum_i^n (x_{ik}\Delta x_{ij} + \Delta x_{ik} x_{ij}) b_j^* - \sum_i^n (\Delta x_{ik} y_i + x_{ik}\Delta y_i)\};$$

$$\Delta b_k^* = \frac{1}{n}\{2\sum_i^n x_{ik}\Delta x_{ik} b_k^* + \sum_{j\neq k}^m \sum_i^n (x_{ik}\Delta x_{ij} + \Delta x_{ik} x_{ij}) b_j^* - \sum_i^n (\Delta x_{ik} y_i + x_{ik}\Delta y_i)\} =$$

$$= \frac{1}{n}\{2\sum_i^n x_{ik}\Delta x_{ik} b_k^* + \sum_{j\neq k}^m \sum_i^n [(x_{ik}\Delta x_{ij} + \Delta x_{ik} x_{ij}) b_j^* - \frac{1}{m-1}\Delta x_{ik} y_i] - \sum_i^n x_{ik}\Delta y_i\} =$$ Будем искать

$$= \frac{1}{n}\{\sum_{j\neq k}^m \sum_i^n [\frac{2}{m-1} x_{ik}\Delta x_{ik} b_k^* + (x_{ik}\Delta x_{ij} + \Delta x_{ik} x_{ij}) b_j^* - \frac{1}{m-1}\Delta x_{ik} y_i] - \sum_i^n x_{ik}\Delta y_i\} =$$

$$= \frac{1}{n}\{\sum_{j\neq k}^m \sum_i^n [(\frac{2}{m-1} x_{ik} b_k^* + x_{ij} b_j^* - \frac{1}{m-1} y_i)\Delta x_{ik} - x_{ik} b_j^* \Delta x_{ij}] - \sum_i^n x_{ik}\Delta y_i\}$$

$\max(|Дb_k^*|)$ по $Дx_{ij}$ и $Дy_i$ ($i=1,\ldots, n$ ;$j=1,\ldots, m$). Для этого рассмотрим все три ранее введенных типа ограничений на ошибки измерения.

*Тип 1* (абсолютные погрешности измерения ограничены). Тогда:

$$\max_{\Delta x,\Delta y}(|\Delta b_k^*|)=\frac{1}{n}\{\sum_{j\neq k}^{m}\sum_{i}^{n}[|(\frac{2}{m-1}x_{ik}b_k^*+x_{ij}b_j^*-\frac{1}{m-1}y_i)|\Delta_k^x+|x_{ik}b_j^*|\Delta_j^x]-\sum_{i}^{n}x_{ik}|\Delta y\}. \qquad Tun \quad 2$$

(относительные погрешности измерения ограничены). Аналогично получим:

$$\sum_{j=1}^{m}|\Delta x_{ij}|<\alpha_x \quad (i=1,2,...,n,\ j=1,2,...,m),\ |\Delta y_i|<\alpha_y \quad (i=1,2,...,n)). \qquad Tun \quad 3$$

(ограничения наложены на сумму погрешностей). Предположим, что $|Db_k^*|$ достигает максимального значения при таких значениях погрешностей $Дx_{ij}$ и $Ду_i$, которые мы обозначим как:

$$\{\Delta x_{ij}^*, \quad i=\overline{1,2,...,n};\ j=1,2,...,m\},\ \{\Delta y_i^*, \quad i=1,2,...,n\}.$$

тогда:

$$\max_{\Delta x,\Delta y}(|\Delta b_k^*|)=\frac{1}{n}\{\sum_{j\neq k}^{m}\sum_{i}^{n}[(\frac{2}{m-1}x_{ik}b_k^*+x_{ij}b_j^*-\frac{1}{m-1}y_i)x_{ik}^*+x_{ik}b_j^*x_{ij}^*]-\sum_{i}^{n}x_{ik}y_i^*\}.$$

Ввиду линейности последнего выражения и выполнения ограничения типа 3:

$$\max_{\Delta x,\Delta y}(|\Delta b_k^*|)=\frac{1}{n}\{\sum_{j\neq k}^{m}\sum_{i}^{n}[|\frac{2}{m-1}x_{ik}b_k^*+x_{ij}b^*{}_j-\frac{1}{m-1}y_i|\cdot|\Delta x_{ik}^*|+|x_{ik}b_j^*|\cdot|\Delta x_{ij}^*|]-\sum_{i}^{n}|x_{ik}|\cdot|\Delta y_i^*|\},$$

$$\sum_{j}^{m}|\Delta x_{ij}^*|=\alpha_x \qquad (j=1,2,...,m), \qquad |\Delta y_i^*|=\alpha_y.$$

Для простоты записей выкладок сделаем следующие замены:

$$|\Delta x_{ij}|=\alpha_{ij}\geq 0, \qquad C_k=n\sum_{i}^{n}|x_{ik}|\cdot|\Delta y_i^*|\geq 0,$$

$$K_i^k=\sum_{j\neq k}^{m}|\frac{2}{m-1}x_{ik}b_k^*+x_{ij}b_j^*-\frac{1}{m-1}y_i|\geq 0,$$

$$|x_{ik}b_j^*|=R_{ij}^k\geq 0.$$

Теперь для достижения поставленной цели можно сформулировать следующую задачу, которая разделяется на *m* типовых задач оптимизации:

$$f_k(\{\alpha_{ij}\})\to\max_{\alpha_{ij}} \quad (i=1,2,...,n;\ j=1,2,...,m;\ k=1,2,...,m),$$

*где*

$$f_k(\{\alpha_{ij}\})=\frac{1}{n}\{\sum_{i}^{n}K_i^k\alpha_{ik}+\sum_{j\neq m}^{m}\sum_{i}^{n}R_{ij}^k\alpha_{ij}\}+C_k,$$

при ограничениях

$$\sum_{j}^{m}\alpha_{ij}=\alpha_x \qquad (j=1,2,...,m).$$

Перепишем минимизируемые функции в следующем виде:

$$f_k=\frac{1}{n}\sum_{i}^{n}(K_i^k\alpha_{ik}+\sum_{j\neq m}^{m}R_{ij}^k\alpha_{ij})+C_k=\frac{1}{n}\sum_{i}^{n}f_i^k+C_k.$$

Очевидно, что $f_i^k>0$.

Легко видеть, что

$$n \cdot \max_{\alpha_{ij}}(f_k) = \max_{\alpha_{i1}}(f_1^k) + \max_{\alpha_{i2}}(f_2^k) + \ldots + \max_{\alpha_{in}}(f_n^k) + C_k = \sum_i^n \max_{\alpha_{ii}}(f_i^k) + C_k,$$

*где* $i = 1,2,\ldots,n; j = 1,2\ldots,m.$

Следовательно, необходимо решить *nm* задач

$$\{f_i^k\} \to \max_{\alpha_{ij}} (i = 1.2.,,,,n;\ j = 1.2,\ldots,m;\ k = 1,2,\ldots,m)$$

при ограничениях "типа равенства":

$$\sum_j^m \alpha_{ij} = \alpha_x \qquad (i = 1,2,\ldots,\ n),$$

*где* $$f_i^k = K_i^k \alpha_{ik} + \sum_{j \neq m}^m R_{ij}^k \alpha_{ij} = \sum_j^m S_{ij}^k \alpha_{ij},$$

*причем* $$S_{ij}^k = \begin{cases} K_i^k, \text{ если} & j = k, \\ R_{ij}^k, \text{ если} & j \neq k. \end{cases}$$

Сформулирована типовая задача поиска экстремума функции. Она легко решается. Поскольку

$$\max_{a_{ij}} (f_i^k) = \max_j (S_{ij}^k) \cdot \alpha_x,$$

то максимальное отклонение МНК-оценки k-ого параметра равно

$$\max_{\Delta X, \Delta Y}(|\hat{\Delta b_k}|) = \max_{\alpha_{ij}}(f_k) = \frac{1}{n}\alpha_x \sum_i^n \max_j (S_{ij}^k) + \frac{1}{n} C_k, \quad (i = 1,2,\ldots,n;\ j = 1,2,\ldots,m).$$

Кроме рассмотренных выше трех видов ограничений на погрешности могут представлять интерес и другие, но для демонстрации типовых результатов ограничимся только этими тремя видами.

**Оценивание линейной корреляционной связи.** В качестве примера рассмотрим оценивание линейной корреляционной связи случайных величин $y$ и $x_1$ , $x_2$..., $x_m$ с нулевыми математическими ожиданиями. Пусть эта связь описывается соотношением:

$$y = \sum_{j=1}^m b_j x_j + e,$$

где $b_1$ , $b_2$ ,..., $b_m$ - постоянные, а случайная величина $e$ некоррелирована с $x_1$ , $x_2$..., $x_m$. Допустим, необходимо оценить неизвестные параметры $b_1$ , $b_2$ ,..., $b_m$ по серии независимых испытаний:

$$y_i = \sum_{j=1}^m b_j x_{ij} + e_i, \qquad (i = 1,2,\ldots,\ n).$$

Здесь при каждом $i = 1,2,\ldots,n$ имеем новую независимую реализацию рассматриваемых случайных величин. В этой частной схеме оценки наименьших квадратов $b_1^{*R}$ , $b_2^{*R}$ ,..., $b_m^{*R}$ параметров $b_1$, $b_2$ ,..., $b_m$ являются, как известно, состоятельными [45].

Пусть величины $x_1$ , $x_2$..., $x_m$ в дополнение к попарной независимости имеют единичные дисперсии. Тогда из закона больших чисел [45] следует существование следующих пределов (ср. предположение 1 выше):

$$\lim_{n\to\infty}\{\frac{1}{n}\sum_{i}^{n} x_{ij}^{R}\} = M\{x_j\} = 0 \qquad (j=\overline{1,m}),$$

$$\lim_{n\to\infty}\{\frac{1}{n}\sum_{i}^{n} (x_{ij}^{R} - M\{x_j\})^2\} = D\{x_j\} = 1 \qquad (j=\overline{1,m}),$$

$$\lim_{n\to\infty}\{\frac{1}{n}\sum_{i}^{n} (x_{ij}^{R} - M\{x_j\})(x_{ik}^{R} - M\{x_k\})\} = 0 \qquad (j,k=\overline{1,m}),$$

$$\lim_{n\to\infty}\{\frac{1}{n}\sum_{i}^{n} y_i^{R}\} = M\{y\} = b_1 M\{x_1\} + ... + b_m M\{x_m\} + M\{e\} = 0,$$

$$\lim_{n\to\infty}\{\frac{1}{n}\sum_{i}^{n} (y_i^{R} - M\{y\})^2\} = D\{y\} = b_1^2 + ... + b_m^2 + \sigma^2,$$

где у - среднее квадратическое отклонение случайной величины е.

Пусть измерения производятся с погрешностями, удовлетворяющими ограничениям типа 1, тогда максимальное приращение величины $|Дb^*_k|$, как показано выше, равно:

$\max_{\Delta x,\Delta y}(|\Delta b^*_k|) = \frac{1}{n}\{\sum_{j\neq k}^{m}\sum_{i}^{n}[|\frac{2}{m-1}x_{ik}^{R}b^*_k + x_{ij}^{R}b^*_j - \frac{1}{m-1}y_i^{r}|\cdot\Delta_k^x + |x_{ik}^{R}b^*_j|\cdot\Delta_j^x] + \sum_{i}^{n}|x_{ik}^{R}|\cdot\Delta y\}$. Перейдем к предельному случаю и выпишем выражение для нотны:

$$N_k = \lim_{n\to\infty}\{\max_{\Delta x,\Delta y}(|\Delta b^*_k|)\} =$$

$$= \sum_{j\neq k}^{m}[M\{|\frac{2}{m-1}x_k b_k + x_j b_j - \frac{1}{m-1}y|\}\cdot\Delta_k^x + M\{|x_k b_j|\}\cdot\Delta_j^x + M\{|x_k|\}\cdot\Delta y.$$

В качестве примера рассмотрим случай $m = 2$. Тогда

$$N_1 = M\{|\,2x_1b_1 + x_2b_2 - y\,|\}\Delta_1^x + M\{b_2x_1\}\Delta_2^x + M\{|\,x_1\,|\}\Delta y,$$

$$N_2 = M\{|\,2x_2b_2 + x_1b_1 - y\,|\}\Delta_2^x + M\{b_1x_2\}\Delta_1^x + M\{|\,x_2\,|\}\Delta y.$$

Приведенное выше выражение для максимального приращения метрологической погрешности не может быть использована в случае $m = 1$. Для $m = 1$ выведем выражение для нотны, исходя из соотношения:

$\Delta b^*_k = \frac{1}{n}\{\sum_{j}^{m}\sum_{i}^{n}(x_{ik}\Delta x_{ij} + \Delta x_{ik}x_{ij}),\quad b^*_j - \sum_{i}^{n}(\Delta x_{ik}y_i + x_{ik}\Delta y_i)\}$. Подставив $m = 1$, получим:

$$\Delta b^* = \frac{1}{n}\{\sum_{i}^{n}(2x_i\Delta x_i)b^* - \sum_{i}^{n}(\Delta x_i y_i + x_i\Delta y_i)\} = \frac{1}{n}\{\sum_{i}^{n}((2x_i b^* - y_i)\Delta x_i + x_i\Delta y_i)\}.$$

Следовательно, *нотна* выглядит так:

$$N_f = M\{|2xb^* - y|\}Дx + M\{|x|\}Ду\,.$$

Для нахождения рационального объема выборки необходимо сделать следующее.

*Этап 1*. Выразить зависимость размеров и меры области рассеивания $B_б(n,b)$ от числа опытов $n$ (см. выше).

*Этап 2*. Ввести меру неопределенности и записать соотношение между статистической и интервальной неопределенностями.

*Этап 3*. По результатам этапов 1 и 2 получить выражение для рационального объема выборки.

Для выполнения этапа 1 определим область рассеивания следующим образом. Пусть доверительным множеством $B_б(n,b)$ является $m$-мерный куб со сторонами длиною $2K$, для которого

$$P(b \in B_\alpha(n, b^{*R})) = \alpha..$$

Исследуем случайный вектор $b^*$ и

$$b^{*R} = (X_R^T X_R)^{-1} X_R^T Y_R = (X_R^T X_R)^{-1} X_R^T (X_R b + e) =$$
$$= (X_R^T X_R)^{-1} X_R^T X_R b + (X_R^T X_R)^{-1} X_R^T e = b + (X_R^T X_R)^{-1} X_R^T e.$$

Как известно, если элементы матрицы $A = \{a_{ij}\}$ -случайные, т.е. $A$ – случайная матрица, то ее математическим ожиданием является матрица, составленная из математических ожиданий ее элементов, т.е. $M\{A\} = \{M\{a_{ij}\}\}$.

*Утверждение 1*. Пусть $A = \{ a_{ij} \}$ и $B = \{ b_{ij} \}$ - случайные матрицы порядка ($m$ x $n$) и ($n$ x $r$) соответственно, причем любая пара их элементов $(a_{ij}, b_{kl})$ состоит из независимых случайных величин. Тогда математическое ожидание произведения матриц равно произведению математических ожиданий сомножителей, т.е. $M\{AB\} = M\{A\}\, M\{B\}$.

*Доказательство.* На основании определения математического ожидания матрицы заключаем, что

$$A \cdot B = \{\sum_k^n a_{ik} \cdot b_{kj}\} \rightarrow M\{A \cdot B\} = \{M\{\sum_k^n a_{ik} \cdot b_{kj}\}\} = \{\sum_k^n M\{a_{ik} \cdot b_{kj}\}\},$$

но так как случайные величины $a_{ik}$, $b_{kj}$ независимы, то

$$M\{A \cdot B\} = \{\sum_k^n M\{a_{ik}\} \cdot M\{b_{kj}\}\} = M\{A\} \cdot M\{B\},$$

что и требовалось доказать.

*Утверждение 2.* Пусть $A = \{a_{ij}\}$ и $B = \{b_{ij}\}$ - случайные матрицы порядка ($m$ x $n$) и ($n$ x $r$) соответственно. Тогда математическое ожидание суммы матриц равно сумме математических ожиданий слагаемых:, т.е. $M\{A+B\} = M\{A\} + M\{B\}$.

*Доказательство.* На основании определения математического ожидания матрицы заключаем, что

$$M\{A+B\} = \{M\{a_{ij}+b_{ij}\}\} = \{M\{a_{ij}\} + M\{b_{ij}\}\} = M\{A\} + M\{B\},$$

что и требовалось доказать.

Найдем математическое ожидание и ковариационную матрицу вектора $b^*$ с помощью утверждений 1, 2 и выражения для $b^{*R}$, приведенного выше. Имеем

$$M\{b^{*R}\} = b + M\{(X_R^T X_R)^{-1} X_R^T e\} = b + M\{(X_R^T X_R)^{-1} X_R^T\} \cdot M\{e\}.$$

Но так как $M\{ e \} = 0$, то $M\{b^{*R}\} = b$. Это означает что оценка МНК является несмещенной.

Найдем ковариационную матрицу:

$$D\{b^{*R}\} = M\{(b^{*R} - b)(b^{*R} - b)^T\} = M\{(X_R^T X_R)^{-1} X_R^T \cdot e \cdot e^T \cdot X_R (X_R^T X_R)^{-1}\}.$$

*Можно доказать, что*

$$D\{b^{*R}\} = M\{(X_R^T X_R)^{-1} X_R^T \cdot M\{e \cdot e^T\} \cdot X_R (X_R^T X_R)^{-1}\},$$

*но*

$$M\{e \cdot e^T\} = D\{e\} = \sigma^2 E,$$

*поэтому*

$$D\{\hat{b}_R\} = M\{(X_R^T X_R)^{-1} X_R^T \cdot (\sigma^2 E) \cdot X_R (X_R^T X_R)^{-1}\} = \sigma^2 \cdot M\{(X_R^T X_R)^{-1}\}.$$

Как выяснено ранее, для достаточно большого количества опытов $n$ выполняется приближенное равенство

$$(X_R^T X_R)^{-1} \approx \frac{1}{n} E, \qquad (51)$$

тогда

$$D\{b^{*R}\} = \frac{\sigma^2}{n} E.$$

Осталось определить вид распределения вектора $b^{*R}$. Из выражения для $b^{*R}$, приведенного выше, и асимптотического соотношения (51) следует, что

$$b^{*R} = b + \frac{1}{n} X_R^T e \quad .$$

Можно утверждать, что вектор $b^{*R}$ имеет асимптотически нормальное распределение, т.е.

$$b^{*R} \in N(b, \frac{\sigma^2}{n} E).$$

Тогда совместная функция плотности распределения вероятностей случайных величин $b^{*R}_1$, $b^{*R}_2$, ..., $b^{*R}_m$ будет иметь вид:

$$f(b^{*R}) = \frac{1}{(2\pi)^{m/2} \cdot (\det C)^{1/2}} \cdot \exp[-\frac{1}{2}(b^{*R} - b)^T \cdot C^{-1} \cdot (b^{*R} - b)], \quad (52)$$

где

$$C = D(b^{*R}) = \frac{\sigma^2}{n} E .$$

Тогда справедливы соотношения

$$C^{-1} = \frac{n}{\sigma^2} E \quad , \quad \det C = \det(\frac{n}{\sigma^2} E) = (\frac{\sigma^2}{n})^m .$$

Подставим в формулу (52), получим

$$f(b^{*R}) = \frac{1}{(2\pi)^{m/2} \cdot (\sigma^2/n)^{m/2}} \cdot \exp[-\frac{n}{2\sigma^2}(b^{*R} - b)^T \cdot C^{-1} \cdot (b^{*R} - b)] =$$

$$= \frac{1}{(\sigma\sqrt{2\pi/n})^m} \exp[-\frac{n}{2\sigma^2}(b^{*R} - b)^T \cdot C^{-1} \cdot (b^{*R} - b)] =$$

$$= \frac{1}{(\sigma\sqrt{2\pi/n})^m} \exp[-\frac{n}{2\sigma^2}(\beta_1^2 + \beta_2^2 + ... + \beta_m^2],$$

где

$$\beta_i = b^{*R}_i - b_i, \quad i = 1,2,...m \quad .$$

Вычислим асимптотическую вероятность попадания описывающего реальность вектора параметров $b$ в $m$-мерный куб с длиной стороны, равной $2k$, и с центром $b^{*R}$.

$$P(-k < \beta_1 < k, -k < \beta_2 < k, ..., -k < \beta_m < k) =$$

$$= \frac{1}{(\sigma\sqrt{2\pi/n})^m} \{\int_{-k}^{k} .. \int_{-k}^{k} \exp[-\frac{n}{2\sigma^2}(\beta_1^2 + \beta_2^2 + ... + \beta_m^2)] \cdot d\beta_1 ... d\beta_m\} =$$

$$= \frac{1}{(\sigma\sqrt{2\pi/n})^m} \{\int_{-k}^{k} \exp[-\frac{n}{2\sigma^2}\beta_1^2] d\beta_1 ... \int_{-k}^{k} \exp[-\frac{n}{2\sigma^2}\beta_i^2] d\beta_i\}.$$

Сделаем замену

$$t_i = \sqrt{n/2} \cdot \frac{1}{\sigma} \beta_i \quad , \quad i = 1,2,..., m.$$

Тогда

$$P = P(-k < \beta_1 < k, -k < \beta_2 < k, ..., -k < \beta_m < k,) =$$

$$= \frac{(\sigma\sqrt{2/n})^m}{(\sigma\sqrt{2\pi/n})^m} [\int_{-T}^{T} e^{-t^2} dt]^m = [(1/\sqrt{\pi}) \int_{-T}^{T} e^{-t^2} dt]]^m = [\Phi_0(T)]^m ,$$

где $T = (n/2)^{1/2}(k/y)$, а $\Phi_0(T)$- интеграл Лапласа,

$$\Phi_0(T) = \Phi(\sqrt{2}T) - \Phi(-\sqrt{2}T),$$

где $\Phi(t)$ - функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Из последнего соотношения получаем

$$T = \Phi^{-1}(P^{1/m}),$$

где $\Phi^{-1}(P)$ - обратная функция Лапласа. Отсюда следует, что

$$k = y\,(2/n)^{1/2}\,\Phi^{-1}(P^{1/m}). \qquad (53)$$

Напомним, что доверительная область $B_б\,(n,b)$ - это $m$-мерный куб, длина стороны которого равна $K$, т.е.

$$P(\,b \in B_б\,(n,b)) = P\,(-K < в_1 < K,\ -K < в_2 < K, \ldots,\ -K < в_m < K) = б.$$

Подставляя $P = б$ в формулу (53), получим

$$K = k = y\,(2/n)^{1/2}\,\Phi^{-1}(б^{1/m}). \qquad (54)$$

Соотношение (54) выражает зависимость размеров доверительной области (т.е. длины ребра куба $K$) от числа опытов $n$, среднего квадратического отклонения $y$ ошибки $e$ и доверительной вероятности б. Это соотношение понадобится для определения рационального объема выборки.

Переходим к этапу 2. Необходимо ввести меру разброса (неопределенности) и установить соотношение между статистической и интервальной (метрологической) неопределенностями с соответствии с ранее сформулированным общим подходом.

Пусть $A$ - некоторое измеримое множество точек в $m$-мерном евклидовом пространстве, характеризующее неопределенность задания вектора $a \in A$ . Тогда необходимо ввести некую меру $M(A)$, измеряющую степень неопределенности. Такой мерой может служить $m$-мерный объем $V(A)$ множества $A$ (т.е. его мера Лебега или Жордана), $M(A) = V(A)$.

Пусть $P$ - $m$-мерный параллелепипед, характеризующий интервальную неопределенность. Длины его сторон равны значениям *нотн* $2N_1, 2N_2, \ldots, 2N_m$ , а центр $a$ (точка пересечений диагоналей параллелепипеда) находится в точке $b^{*R}$. Пусть $C$ - измеримое множество точек, характеризующее общую неопределенность. В рассматриваемом случае это $m$-мерный параллелепипед, длины сторон которого равны $2(N_1 + K),\ 2(N_2 + K), \ldots,\ 2(N_m + K)$, а центр находится в точке $b^{*R}$ . Тогда

$$M(P) = V(P) = 2^m N_1 N_2 \ldots N_m, \qquad (55)$$

$$M(C) = V(C) = 2^m (N_1 + K)(N_2 + K)\ldots(N_m + K). \qquad (56)$$

Справедливо соотношение (49), согласно которому $M(C) = M(P) + M(F)$, где множество $F = C \backslash P$ характеризует статистическую неопределенность.

На этапе 3 получаем по результатам этапов 1 и 2 выражение для рационального объема выборки. Найдем то число опытов, при котором статистическая неопределенность составит д 100% от общей неопределенности, т.е. согласно правилу (50)

$$M(F) = M(C) - M(P) = \partial\, M(C) \qquad (57)$$

где $0 < \delta < 1$. Подставив (55) и (56) в (57), получим

$$2^m \prod_{i=1}^{m}(N_i + K) - 2^m \prod_{i=1}^{m}(N_i) = 2^m \delta \prod_{i=1}^{m}(N_i + K).$$

Следовательно,

$$(1-\delta)\prod_{i=1}^{m}(N_i + K) \Big/ \prod_{i=1}^{m}(N_i) = 1.$$

Преобразуем эту формулу:

$$(1-\delta)\prod_{i=1}^{m}(1 + K / N_i) = 1,$$

откуда

$$\prod_{i}^{m}(1 + K / N_i) = 1/(1-\delta).$$

Если статистическая погрешность мала относительно метрологической, т.е. величины *K/Ni* малы, то

$$\prod_i^m (1 + K / N_i) \approx 1 + \sum_i^m (K / N_i).$$

При $m = 1$ эта формула является точной. Из нее следует, что для дальнейших расчетов можно использовать соотношение

$$1 + \sum_i^m (K / N_i) = 1/(1 - \delta).$$

Отсюда нетрудно найти *K*:

$$K = \frac{\delta}{1 - \delta}(1 / \sum_{i=1}^{m} (1 / N_i)). \qquad (58)$$

Подставив в формулу (58) зависимость *K = K(n)*, полученную в формуле (54), находим приближенное (асимптотическое) выражение для рационального объема выборки:

$$n_{рац} = 2(\frac{1-\delta}{\delta}\sigma\sum_{i=1}^{m}(1/N_i)\cdot\Phi^{-1}(\alpha^{1/m}))^2.$$

При $m = 1$ эта формула также справедлива, более того, является точной.

Переход от произведения к сумме является обоснованным при достаточно малом $\delta$, т.е. при достаточно малой статистической неопределенности по сравнению с метрологической. В общем случае можно находить *K* и затем рациональный объем выборки тем или иным численным методом.

**Пример 1.** Представляет интерес определение $n_{рац}$ для случая, когда $m = 2$, поскольку простейшая линейная регрессия с $m = 2$ широко применяется. В этом случае базовое соотношение имеет вид

$$(1 + K/N_1)(1 + K/N_2) = 1/(1 - \delta).$$

Решая это уравнение относительно *K*, получаем

$$K = 0.5\{-(N_1 + N_2) + [(N_1 + N_2)^2 + 4 N_1 N_2 (\delta/(1 - \delta)]^{1/2}\}.$$

Далее, подставив в формулу (54), получим уравнение для рационального объема выборки в случае $m = 2$:

$$y (2/n)^{1/2} \Phi^{-1}(б^{1/2}) = 0.5\{-(N_1 + N_2) + [(N_1 + N_2)^2 + 4 N_1 N_2 (\delta/(1 - \delta)]^{1/2}\}.$$

Следовательно,

$$n_{rat} = \frac{8\left\{\Phi^{-1}(\sqrt{\alpha})\right\}^2}{\left\{-\frac{N_1}{\sigma} - \frac{N_2}{\sigma} + \sqrt{\left(\frac{N_1}{\sigma} + \frac{N_2}{\sigma}\right)^2 + 4\frac{N_1 N_2 \delta}{\sigma^2(1-\delta)}}\right\}^2}.$$

При использовании «принципа уравнивания погрешностей» согласно [3] $\delta = 1/2$. При доверительной вероятности $\alpha = 0{,}95$ имеем $\sqrt{\alpha} = 0{,}9747$ и согласно [42] $\Phi^{-1}(\sqrt{\alpha}) = 1{,}954$. Для этих численных значений

$$n_{rat} = \frac{30{,}545}{\left\{-\frac{N_1}{\sigma} - \frac{N_2}{\sigma} + \sqrt{\left(\frac{N_1}{\sigma} + \frac{N_2}{\sigma}\right)^2 + 4\frac{N_1 N_2}{\sigma^2}}\right\}^2}.$$

Если $\frac{N_1}{\sigma} = \frac{N_2}{\sigma} = 0{,}1$, то $n_{rat} = 4455$. Если же $\frac{N_1}{\sigma} = \frac{N_2}{\sigma} = 0{,}5$, то $n_{rat} = 178$. Если первое из этих чисел превышает обычно используемые объемы выборок, то второе находится в «рабочей зоне» регрессионного анализа.

**Парная регрессия.** Наиболее простой и одновременно наиболее широко применяемый частный случай парной регрессии рассмотрим подробнее. Модель имеет вид

$$y_i = ax_i + b + \varepsilon_i, \ \ i = 1,2,...,n.$$

Здесь $x_i$ – значения фактора (независимой переменной), $y_i$ – значения отклика (зависимой переменной), $\varepsilon_i$ - статистические погрешности, *a, b* - неизвестные параметры, оцениваемые методом наименьших квадратов. Она переходит в модель (используем альтернативную запись линейной модели)

$$y = X\beta + \varepsilon,$$

если положить

$$y = \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1 & 1 \\ ... & ... \\ x_n & 1 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ ... \\ \varepsilon_n \end{pmatrix}, \quad \beta = \begin{pmatrix} a \\ b \end{pmatrix}.$$

Естественно принять, что погрешности факторов описываются матрицей

$$\Delta X = \alpha = \begin{pmatrix} \Delta x_1 & 0 \\ ... & ... \\ \Delta x_n & 0 \end{pmatrix} = \begin{pmatrix} \alpha_1 & 0 \\ ... & ... \\ \alpha_n & 0 \end{pmatrix}.$$

В рассматриваемой модели интервального метода наименьших квадратов

$$X = X_R + \alpha, \quad y = y_R + \gamma \begin{pmatrix} 1 \\ ... \\ 1 \end{pmatrix},$$

где *X, y* – наблюдаемые (т.е. известные статистику) значения фактора и отклика, $X_R$, $y_R$ – истинные значения переменных, $\alpha, \gamma$ - погрешности измерений переменных. Пусть $\beta^*$ - оценка метода наименьших квадратов, вычисленная по наблюдаемым значениям переменных, $\beta_R^*$ - аналогичная оценка, найденная по истинным значениям. В соответствии с ранее проведенными рассуждениями

$$\beta^* - \beta = [-(X_0^T X_0)^{-1} \Delta (X_0^T X_0)^{-1} X_0^T + (X_0^T X_0)^{-1} \alpha^T] y_0 + (X_0^T X_0)^{-1} X_0^T \gamma \quad (59)$$

с точностью до бесконечно малых более высокого порядка по $|\alpha|$ и $|\gamma|$. В формуле (59) использовано обозначение $\Delta = X_0^T \alpha + \alpha^T X_0$. Вычислим правую часть в (59), выделим главный линейный член и найдем нотну.

Легко видеть, что

$$X^T X = \begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix}, \quad (60)$$

где суммирование проводится от 1 до *n*. Для упрощения обозначений в дальнейшем до конца настоящего пункта не будем указывать эти пределы суммирования. Из (60) вытекает, что

$$(X^T X)^{-1} = \begin{pmatrix} n & -\sum x_i \\ -\sum x_i & \sum x_i^2 \end{pmatrix} \Big/ [n \sum x_i^2 - (\sum x_i)^2]. \quad (61)$$

Легко подсчитать, что

$$X^T \alpha + \alpha^T X = \begin{pmatrix} 2\sum x_i \alpha_i & \sum \alpha_i \\ \sum \alpha_i & n \end{pmatrix}. \quad (62)$$

Положим

$$S_0^2 = \frac{1}{n} \sum (x_i - \bar{x})^2.$$

Тогда знаменатель в (61) равен $n^2 S_0^2$. Из (61) и (62) следует, что

$$(X^T X)^{-1} (X^T \alpha + \alpha^T X) = \frac{1}{n^2 S_0^2} \begin{pmatrix} 2n\sum x\alpha - \sum x \sum \alpha & n\sum \alpha \\ -2\sum x \sum x\alpha + \sum x^2 \sum \alpha & -\sum x \sum \alpha \end{pmatrix}. \quad (63)$$

Здесь и далее опустим индекс $i$, по которому проводится суммирование. Это не может привести к недоразумению, поскольку всюду суммирование проводится по индексу $i$ в интервале от 1 до $n$. Из (61) и (63) следует, что

$$(X^TX)^{-1}(X^T\alpha+\alpha^TX)(X^TX)^{-1}=\frac{1}{n^4S_0^4}\begin{pmatrix}A & B\\ C & D\end{pmatrix}, \quad (64)$$

где

$$A=2n^2\sum x\alpha-2n\sum x\sum\alpha,$$
$$B=C=-2n\sum x\sum x\alpha+\Big(\sum x\Big)^2\sum\alpha+n\sum\alpha\sum x^2,$$
$$D=2\Big(\sum x\Big)^2\sum x\alpha-2\sum\alpha\sum x\sum x^2.$$

Наконец, вычисляем основной множитель в (59)

$$(X^TX)^{-1}(X^T\alpha+\alpha^TX)(X^TX)^{-1}X^T=\frac{1}{n^4S_0^4}\begin{pmatrix}z_{11} & z_{12} & \dots & z_{1i} & \dots & z_{1n}\\ z_{21} & z_{22} & \dots & z_{2i} & \dots & z_{2n}\end{pmatrix}, \quad (65)$$

где

$$z_{1i}=Ax_i+B,\ \ z_{2i}=Cx_i+D,\ \ i=1,2,...,n.$$

Перейдем к вычислению второго члена с $\alpha$ в (59). Имеем

$$(X^TX)^{-1}\alpha^T=\frac{1}{n^2S_0^2}\begin{pmatrix}w_{11} & \dots & w_{1i} & \dots & w_{1n}\\ w_{21} & \dots & w_{2i} & \dots & w_{2n}\end{pmatrix}, \quad (67)$$

где

$$w_{1i}=n\alpha_i,\ \ w_{2i}=-\alpha_i\sum x,\ \ i=1,2,...,n.$$

Складывая правые части (65) и (67) и умножая на $y$, получим окончательный вид члена с $\alpha$ в (59):

$$\{(X^TX)^{-1}(X^T\alpha+\alpha^TX)(X^TX)^{-1}X^T+(X^TX)^{-1}\alpha^T\}y=\begin{pmatrix}F\\ G\end{pmatrix}, \quad (68)$$

где

$$\begin{aligned}F&=\Big(\sum xy\Big)\Big(2n^2\sum x\alpha-2n\sum x\sum\alpha\Big)/n^4S_0^4+\Big(\sum y\alpha\Big)/nS_0^2+\\&+\Big(\sum y\Big)\Big(n\sum\alpha\sum x^2+\sum\alpha\Big(\sum x\Big)^2-2n\sum x\sum x\alpha\Big)/n^4S_0^4,\\G&=\Big(\sum xy\Big)\Big(-2n\sum x\sum x\alpha+n\sum\alpha\sum x^2+\sum\alpha\Big(\sum x\Big)^2\Big)/n^4S_0^4-\\&-\Big(\sum y\alpha\Big)\Big(\sum x\Big)/n^2S_0^2+\Big(\sum y\Big)\Big(2\sum x\alpha\Big(\sum x\Big)^2-2\sum\alpha\sum x\sum x^2\Big)/n^4S_0^4.\end{aligned} \quad (69)$$

Для вычисления нотны выделим главный линейный член. Сначала найдем частные производные. Имеем

$$\begin{aligned}\frac{\partial F}{\partial\alpha_j}&=\Big(\sum xy\Big)(2n^2x_j-2n\sum x)/n^4S_0^4+y_j/nS_0^2+\\&+\Big(\sum y\Big)\Big(n\sum x^2+\Big(\sum x\Big)^2-2n\Big(\sum x\Big)x_j\Big)/n^4S_0^4;\\\frac{\partial G}{\partial\alpha_j}&=\Big(\sum xy\Big)\Big(-2n(\sum x)x_j+n\sum x^2+\Big(\sum x\Big)^2\Big)/n^4S_0^4-\\&-y_j\Big(\sum x\Big)/n^2S_0^2+\Big(\sum y\Big)\Big(2x_j\Big(\sum x\Big)^2-2\sum x\sum x^2\Big)/n^4S_0^4.\end{aligned} \quad (70)$$

Если ограничения имеют вид

$$|\alpha_j|\le\Delta,\ \ j=1,2,...,n,$$

то максимально возможное отклонение оценки $a^*$ параметра $a$ из-за погрешностей $\alpha_j$ таково:

$$N_a(x)=\sum_{1\le j\le n}\left|\frac{\partial F}{\partial\alpha_j}\right|\Delta+O(\Delta^2),$$

где производные заданы формулой (70).

**Пример 2.** Пусть вектор *(x,y)* имеет двумерное нормальное распределение с нулевыми математическими ожиданиями, единичными дисперсиями и коэффициентом корреляции $\rho$. Тогда

$$\lim_{\Delta\to 0}\lim_{n\leftarrow\infty}\frac{N_a(x)}{\Delta}=\lim_{n\to\infty}\sum_{1\le j\le n}\left|\frac{\partial F}{\partial\alpha_j}\right|=M\mid 2\rho x+y\mid=\sqrt{\frac{2(1+8\rho^2)}{\pi}}. \quad (71)$$

При этом

$$\lim_{n\to\infty}\frac{\partial G}{\partial\alpha_j}=\rho,$$

следовательно, максимально возможному изменению параметра $b^*$ соответствует сдвиг всех $x_i$ в одну сторону, т.е. наличие систематической ошибки при определении *x*-ов. В то же время согласно (71) значения $\alpha_j$ в асимптотике выбираются по правилу

$$\alpha_j=\begin{cases}\Delta, & 2\rho x_j+y_j>0,\\ -\Delta, & 2\rho x_j+y_j\le 0.\end{cases}$$

Таким образом, максимальному изменению $a^*$ соответствуют не те $\alpha_j$, что максимальному изменению $b^*$. В этом – новое по сравнению с одномерным случаем. В зависимости от вида ограничений на возможные отклонения, в частности, от вида метрики в пространстве параметров, будут «согласовываться» отклонения по отдельным параметрам. Ситуация аналогична той, что возникает в классической математической статистике в связи с оптимальным оцениванием параметров. Если параметр одномерен, то ситуация с оцениванием достаточно прозрачна – есть понятие эффективных оценок, показателем качества оценки является средний квадрат ошибки, а при ее несмещенности – дисперсия. В случае нескольких параметров возникает необходимость соизмерить точность оценивания по разным параметрам. Есть много критериев оптимальности (см., например, [46]), но нет признанных правил выбора среди них.

Вернемся к формуле (59). Интересно, что отклонения вектора параметров, вызванные отклонениями значений факторов $\alpha$ и отклика $\gamma$, входят в (59) аддитивно. Хотя

$$\sup_{\alpha,\gamma}\|\beta^*-\beta\|\ne\sup_{\alpha}\mid\{-(X_0^TX_0)^{-1}\Delta(X_0^TX_0)^{-1}X_0^T+(X_0^TX_0)^{-1}\alpha^T\}y_0\mid+$$

$$+\sup_{\gamma}\mid(X_0^TX_0)^{-1}X_0^T\gamma\mid,$$

но для отдельных компонент (не векторов!) имеет место равенство.

В случае парной регрессии

$$(X_0^TX_0)^{-1}X_0^T\gamma=\frac{1}{n^2S_0^2}\left(\sum\gamma_i(nx_i-\sum x);\ \sum\gamma_i(-x_i\sum x+\sum x^2)\right)^T. \quad (72)$$

Из формул (68), (69) и (72) следует, что

$$\beta^*-\beta=\begin{pmatrix}a^*(X,y)-a^*(X_0,y_0)\\ b^*(X,y)-b^*(X_0,y_0)\end{pmatrix}=\begin{pmatrix}F+F_1\\ G+G_1\end{pmatrix},$$

где *F* и *G* определены в (69), а

$$F_1=\frac{1}{n^2S_0^2}\left(n\sum\gamma x-\sum x\sum\gamma\right),\quad G_1=\frac{1}{n^2S_0^2}\left(\sum\gamma\sum x^2-\sum\gamma x\sum x\right).$$

Итак, продемонстрирована возможность применения основных подходов статистики интервальных данных в регрессионном анализе.

### 3.5.5. Интервальный дискриминантный анализ

Перейдем к задачам классификации в статистике интервальных данных. Как известно [27], важная их часть – задачи дискриминации (диагностики, распознавания образов с учителем). В этих задачах заданы классы (полностью или частично, с помощью обучающих выборок), и необходимо принять решение –к какому этих классов отнести вновь поступающий объект.

В линейном дискриминантном анализе правило принятия решений основано на линейной функции *f(x)* от распознаваемого вектора $x \in R^k$. Рассмотрим для простоты случай двух классов. Правило принятия решений определяется константой *C* – при *f(x)>C* распознаваемый объект относится к первому классу, при *f(x)*$\leq$*C* – ко второму.

В первоначальной вероятностной модели Р.Фишера предполагается, что классы заданы обучающими выборками объемов $N_1$ и $N_2$ соответственно из многомерных нормальных распределений с разными математическими ожиданиями, но одинаковыми ковариационными матрицами. В соответствии с леммой Неймана-Пирсона, дающей правило принятия решений при поверке статистических гипотез, дискриминантная функция является линейной. Для ее практического использования теоретические характеристики распределения необходимо заменить на выборочные. Тогда дискриминантная функция приобретает следующий вид

$$f(x) = \left( x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2) \right)^T S^{-1}(\bar{x}_1 - \bar{x}_2).$$

Здесь $\bar{x}_1$ - выборочное среднее арифметическое по первой выборке $x_\alpha^{(1)}$, $\alpha = 1,2,...,N_1$, а $\bar{x}_2$ - выборочное среднее арифметическое по второй выборке $x_\beta^{(2)}$, $\beta = 1,2,...,N_2$. В роли S может выступать любая состоятельная оценка общей для выборок ковариационной матрицы. Обычно используют следующую оценку, естественным образом сконструированную на основе выборочных ковариационных матриц:

$$S = \frac{\sum_{\alpha=1}^{N_1}(x_\alpha^{(1)} - \bar{x}_1)(x_\alpha^{(1)} - \bar{x}_1)^T + \sum_{\beta=1}^{N_2}(x_\beta^{(2)} - \bar{x}_2)(x_\beta^{(2)} - \bar{x}_2)^T}{N_1 + N_2 - 2}.$$

В соответствии с подходом статистики интервальных данных считаем, что специалисту по анализу данных известны лишь значения с погрешностями

$$y_\alpha^{(1)} = x_\alpha^{(1)} + \varepsilon_\alpha^{(1)}, \ \alpha = 1,2,...,N_1, \quad y_\beta^{(2)} = x_\beta^{(2)} + \varepsilon_\beta^{(2)}, \ \beta = 1,2,...,N_2.$$

Таким образом, вместо *f(x)* статистик делает выводы на основе искаженной линейной дискриминантной функции $f_1(x)$, в которой коэффициенты рассчитаны не по исходным данным $x_\alpha^{(1)}, x_\beta^{(2)}$, а по искаженным погрешностями значениям $y_\alpha^{(1)}, y_\beta^{(2)}$.

Это – модель с искаженными параметрами дискриминантной функции. Следующая модель – такая, в которой распознаваемый вектор *x* также известен с ошибкой. Далее, константа *C* может появляться в модели различными способами. Она может задаваться априори абсолютно точно. Может задаваться с какой-то ошибкой, не связанной с ошибками, вызванными конечностью обучающих выборок. Может рассчитываться по обучающим выборкам, например, с целью уравнять ошибки классификации, т.е. провести плоскость дискриминации через середину отрезка, соединяющего центры классов. Итак – целый спектр моделей ошибок.

На какие статистические процедуры влияют ошибки в исходных данных? Здесь тоже много постановок. Можно изучать влияние погрешностей измерений на значения дискриминантной функции *f*, например, в той точке, куда попадает вновь поступающий объект *x*. Очевидно, случайная величина *f(x)* имеет некоторое распределение, определяемое распределениями обучающих выборок. Выше описана модель Р.Фишера с нормально распределенными совокупностями. Однако реальные данные, как правило, не подчиняются нормальному распределению [27]. Тем не менее линейный статистический анализ имеет смысл и для распределений, не являющихся нормальными (при этом вместо свойств многомерного нормального распределения приходится опираться на многомерную центральную предельную теорему и теорему о наследовании сходимости [3]). В частности, приравняв метрологическую ошибку, вызванную погрешностями исходных данных, и статистическую ошибку, получим условие, определяющее рациональность объемов выборок. Здесь два объема выборок, а не один, как в большинстве рассмотренных постановок статистики интервальных данных. С подобным мы сталкивались ранее при рассмотрении двухвыборочного критерия Смирнова.

Естественно изучать влияние погрешностей исходных данных не при конкретном *x*, а для правила принятия решений в целом. Может представлять интерес изучение характеристик этого

правила по всем $x$ или по какому-либо отрезку. Более интересно рассмотреть показатель качества классификации, связанный с пересчетом на модель линейного дискриминантного анализа [27].

Математический аппарат изучения перечисленных моделей развит выше в предыдущих пунктах настоящей главы. Некоторые результаты приведены в [14]. Из-за большого объема выкладок ограничимся приведенными здесь замечаниями.

### 3.5.6. Интервальный кластер-анализ

Кластер-анализ, как известно [27], имеет целью разбиение совокупности объектов на группы сходных между собой. Многие методы кластер-анализа основаны на использовании расстояний между объектами. (Степень близости между объектами может измеряться также с помощью мер близости и показателей различия, для которых неравенство треугольника выполнено не всегда.) Рассмотрим влияние погрешностей измерения на расстояния между объектами и на результаты работы алгоритмов кластер-анализа.

С ростом размерности $p$ евклидова пространства диагональ единичного куба растет как $\sqrt{p}.$ А какова погрешность определения евклидова расстояния? Пусть двум рассматриваемым векторам соответствуют $X_0 = (x_1, x_2, ..., x_p)$ и $Y_0 = (y_1, y_2, ..., y_p)$ - вектора размерности $p$. Они известны с погрешностями $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_p)$ и $\delta = (\delta_1, \delta_2, ..., \delta_p)$, т.е. статистику доступны лишь вектора $X = X_0 + \varepsilon,\ Y = Y_0 + \delta.$ Легко видеть, что

$$\rho^2(X,Y) = \rho^2(X_0,Y_0) + 2\sum_{1\le i\le p}(x_i - y_i)(\varepsilon_i - \delta_i) + \sum_{1\le i\le p}(\varepsilon_i - \delta_i)^2. \quad (73)$$

Пусть ограничения на абсолютные погрешности имеют вид

$$|\varepsilon_i| \le \Delta,\ |\delta_i| \le \Delta,\ i = 1,2,...,n.$$

Такая запись ограничений предполагает, что все переменные имеют примерно одинаковый разброс. Трудно ожидать этого, если переменные имеют различные размерности. Однако рассматриваемые ограничения на погрешности естественны, если переменные предварительно стандартизованы, т.е. отнормированы (т.е. из каждого значения вычтено среднее арифметическое, а разность поделена на выборочное среднее квадратическое отклонение).

Пусть $p\Delta^2 \to 0.$ Тогда последнее слагаемое в (73) не превосходит $4p\Delta^2,$ поэтому им можно пренебречь. Тогда из (73) следует, что нотна евклидова расстояния имеет вид

$$N_{\rho^2}(X_0,Y_0) = 4\sum_{1\le i\le p}|x_i - y_i|\,\Delta$$

с точностью до бесконечно малых более высокого порядка. Если случайные величины $|x_i - y_i|$ имеют одинаковые математические ожидания и для них справедлив закон больших чисел (эти предположения естественны, если переменные перед применением кластер-анализа стандартизованы), то существует константа $C$ такая, что

$$N_{\rho^2}(X_0,Y_0) = Cp\Delta$$

с точностью до бесконечно малых более высокого порядка при малых $\Delta,$ больших $p$ и $p\Delta^2 \to 0.$

Из рассмотрений настоящего пункта вытекает, что

$$\rho(X,Y) = \rho(X_0,Y_0) + \theta\frac{Cp\Delta}{2\rho(X_0,Y_0)} \quad (74)$$

при некотором $\theta$ таком, что $|\theta| < 1.$

Какое минимальное расстояние является различимым? По аналогии с определением рационального объема выборки при проверке гипотез предлагается уравнять слагаемые в (74), т.е. определять минимально различимое расстояние $\rho_{\min}$ из условия

$$\rho_{\min} = \frac{Cp\Delta}{2\rho_{\min}}, \quad \rho_{\min} = \sqrt{\frac{Cp\Delta}{2}}. \quad (75)$$

Естественно принять, что расстояния, меньшие $\rho_{\min}$, не отличаются от 0, т.е. точки, лежащие на расстоянии $\rho \le \rho_{\min}$, не различаются между собой.

Каков порядок величины $C$? Если $x_i$ и $y_i$ независимы и имеют стандартное нормальное распределение с математическим ожиданием 0 и дисперсией 1, то, как легко подсчитать, $M \mid x_i - y_i \mid = 2/\sqrt{\pi} = 1{,}13$ и соответственно $C = 4{,}51$. Следовательно, в этой модели

$$\rho_{\min} = 1{,}5\sqrt{p\Delta}.$$

Формула (75) показывает, что хотя с ростом размерности пространства $p$ растет диаметр (длина диагонали) единичного куба – естественной области расположения значений переменных, с той же скоростью растет и естественное квантование расстояния с помощью порога неразличимости $\rho_{\min}$, т.е. увеличение размерности (вовлечение новых переменных), вообще говоря, не улучшает возможности кластер-анализа.

Можно сделать выводы и для конкретных алгоритмов. В дендрограммах (например, результатах работы иерархических агломеративных алгоритмах ближнего соседа, дальнего соседа, средней связи) можно порекомендовать склеивать (т.е. объединять) уровни, отличающиеся менее чем на $\rho_{\min}$. Если все уровни склеятся, то можно сделать вывод, что у данных нет кластерной структуры, они однородны. В алгоритмах типа «Форель» центр тяжести текущего кластера определяется с точностью $\pm\Delta$ по каждой координате, а порог для включения точки в кластер (радиус шара $R$) из-за погрешностей исходных данных может измениться согласно (74) на

$$\pm \frac{2{,}25}{R} p\Delta.$$

Поэтому кроме расчетов с $R$ рекомендуется провести также расчеты с радиусами $R_1$ и $R_2$, где

$$R_1 = R\left(1 - \frac{2{,}25}{R^2} p\Delta\right), \quad R_2 = R\left(1 + \frac{2{,}25}{R^2} p\Delta\right),$$

и сравнить полученные разбиения. Быть адекватными реальности могут только выводы, общие для всех трех расчетов. Эти рекомендации развивают общую идею [3] о целесообразности проведения расчетов при различных значениях параметров алгоритмов с целью выделения выводов, инвариантных по отношению к выбору конкретного алгоритма.

### 3.5.7. Статистика интервальных данных и оценки погрешностей характеристик финансовых потоков инвестиционных проектов

Методы статистики интервальных данных оказываются полезными не только в традиционных технических и эконометрических задачах, но и во многих других областях, в экономике и менеджменте, например, в инновационном менеджменте.

Основная идея формулируется так. Все знают, что любое инженерное измерение проводится с некоторой погрешностью. Эту погрешность обычно приводят в документации и учитывают при принятии решений. Ясно, что и любое экономическое измерение также проводится с погрешностью. А вот какова она? Необходимо уметь ее оценивать, поскольку ошибки при принятии экономических решений обходятся дорого.

Например, как принимать решение о выгодности или невыгодности инвестиционного проекта? Как сравнивать инвестиционные проекты между собой? Как известно, для решения этих задач используют такие экономические характеристики, как NPV (Net Present Value) - чистая текущая стоимость (этот термин переводится с английского также как чистый дисконтированный доход, чистое приведенное значение и др.), внутренняя норма доходности, срок окупаемости, показатели рентабельности и др.

С экономической точки зрения инвестиционные проекты описываются финансовыми потоками, т.е. функциями от времени, значениями которых являются платежи (и тогда значения этих функций отрицательны) и поступления (значения функций положительны). Сравнение инвестиционных проектов - это сравнение функций от времени с учетом внешней среды, проявляющейся в виде дисконт-функции (как результата воздействия СТЭП-факторов), и представлений законодателя или инвестора - обычно ограничений на финансовые потоки

платежей и на горизонт планирования. Основная проблема при сравнении инвестиционных проектов такова: *что лучше - меньше, но сейчас, или больше, но потом*? Как правило, чем больше вкладываем сейчас, тем больше получаем в более или менее отдаленном будущем. Вопрос в том, достаточны ли будущие поступления, чтобы покрыть нынешние платежи и дать приемлемую для инвестора прибыль?

В настоящее время широко используются различные теоретические подходы к сравнению инвестиционных проектов и облегчающие расчеты компьютерные системы, в частности, Project Expert, COMFAR, PROPSIN, Альт-Инвест, ТЭО-ИНВЕСТ. Однако ряд важных моментов в них не учтен.

Введем основные понятия. Дисконт-функция как функция от времени показывает, сколько стоит для фирмы 1 руб. в заданный момент времени, если его привести к начальному моменту. Если дисконт-функция - константа для разных отраслей, товаров и проектов, то эта константа называется дисконт-фактором, или просто дисконтом. Дисконт-функция определяется совместным действием различных факторов, в частности, реальной процентной ставки и индекса инфляции. Реальная процентная ставка описывает "нормальный" рост экономики (т.е. без инфляции). В стабильной ситуации доходность от вложения средств в различные отрасли, в частности, в банковские депозиты, примерно одинакова. Сейчас она, по оценке ряда экспертов, около 12%. Итак, нынешний 1 руб. превращается в 1,12 руб. через год, а потому 1 руб. через год соответствует 1/1,12 = 0,89 руб. сейчас - это и есть максимум дисконта.

Обозначим дисконт буквой *C*. Если *q* - банковский процент (плата за депозит), т.е. вложив в начале года в банк 1 руб., в конце года получим (1+ *q*) руб., то дисконт определяется по формуле C=1/(1+q). При таком подходе полагают, что банковские проценты одинаковы во всех банках. Более правильно было бы считать *q*, а потому и *C*, нечисловыми величинами, а именно, интервалами $[q_1; q_2]$ и $[C_1; C_2]$. Следовательно, экономические выводы должны быть исследованы на *устойчивость* (применяют и термин *"чувствительность"*) по отношению к возможным отклонениям.

Как функцию времени *t* дисконт-функцию обозначим *C(t)*. При постоянстве дисконт-фактора имеем *C(t) = Ct*. Если *q* = 0,12, *C* = 0,89, то 1 руб. за 2 года превращается в $1{,}12^2 = 1{,}2544$, через 3 - в 1,4049. Итак, 1 руб., получаемый через 2 года, соответствует 1/1,2544=0,7972 руб., т.е. 79,72 коп. сейчас, а 1 руб., обещанный через 3 года, соответствует 0,71 руб. сейчас. Другими словами, *C(2)* = 0,80, а *C(3)* = 0,71. Если дисконт-фактор зависит от времени, в первый год равен *C1*, во второй - *C2*, в третий - *C3*,..., в *t*-ый год - *Ct*, то *C(t)*=C1C2C3...Ct. .

Рассмотрим характеристики потоков платежей. Срок окупаемости - тот срок, за который доходы покроют расходы. Обычно предполагается, что после этого проект приносит только прибыль. Это верно не всегда. Простейший вариант, для которого не возникает никаких парадоксов, состоит в том, что все инвестиции (капиталовложения) делаются сразу, в начале, а затем инвестор получает только доход. Сложности возникают, если проект состоит из нескольких очередей, вложения распределены во времени. Тогда, например, понятие "срок окупаемости" может быть денежных единиц со временем, т.е. не учитывает дисконтирование. Если неоднозначно - вслед за окупаемостью первой очереди может придти очередь затрат на вторую очередь проекта…

Примитивный способ расчета срока окупаемости состоит в делении объема вложений *A* на ожидаемый ежегодный доход *B*. Тогда срок окупаемости равен *A/B*. Этот способ падение стоимости дисконт-фактор равен *C,* то максимально возможный суммарный доход равен

*BC+BC2+BC3+BC4+BC5+...=BC(1+C+C2+C3+C4+...) = BC / (1-C).*

Если *A/B* меньше *C/(1-C),* то можно рассчитать срок окупаемости проекта, но он будет больше, чем *A/B*. Если же *A/B* больше или равно *C/(1-C),* то проект не окупится никогда. Поскольку максимум *C* равен 0,89, то проект не окупится никогда, если *A/B* не меньше 8,09.

Пусть вложения равны 1 млн. руб., ежегодная прибыль составляет 500 тыс., т.е. *A/B*=2, дисконт-фактор *C*=0.8. При примитивном подходе (при *C*=1) срок окупаемости равен 2 годам. А на самом деле? За *k* лет будет возвращено

$$BC(1+C+C2+C3+C4+...+Ck)=BC(1-Ck+1) / (1-C).$$

Срок окупаемости *k* получаем из уравнения 1=0,5x0,8(1-0,8 *k*+1)/(1- 0,8), откуда k= 2,11. Он оказался равным 2,11 лет, т.е. увеличился примерно на 4 недели. Это немного. Однако если *B* =

0,2, то имеем уравнение 1=0,2x0,8(1-0,8*k*+1)/ (1- 0,8). У этого уравнения нет корней, поскольку *A/B*=5>*C/(1-C)*=0.8/(1-0,8)=4. Проект не окупится никогда. Прибыль можно ожидать лишь при *A/B*<4. Рассмотрим промежуточный случай, *B*=0,33, с "примитивным" сроком окупаемости 3 года. Тогда имеем уравнение 1=0,33x0,8 (1-0,8 *k*+1)/ (1-0,8), откуда *k* = 5,40.

Рассмотрим финансовый поток *a(0), a(1), a(2), a(3), ... , a(t), ....* (для простоты примем, что платежи или поступления происходят раз в год). Выше рассмотрен поток с одним платежом *a(0)=(-A)* и дальнейшими поступлениями *a(1) = a(2) = a(3) = ... = a(t) = .... = B*. Чистая текущая стоимость (Net Present Value, сокращенно NPV), рассчитывается для финансового потока путем приведения затрат и поступлений к начальному моменту времени:

$$NPV = a(0) + a(1)C(1) + a(2)C(2) + a(3)C(3) + ... + a(t)C(t) + ...,$$

где *C(t)* - дисконт-функция. В простейшем случае, когда дисконт-фактор не меняется год от года и имеет вид *C=1/(1+q),* формула для NPV конкретизируется:

$$NPV = NPV(q) = a(0) + a(1)/(1+q) + a(2)/(1+q)2 + a(3)/(1+q)3 + ... + a(t)/(1+q)t + ...$$

Пусть, например, *a(0)= -10, a(1)=3, a(2)=4, a(3)=5.* Пусть q=0,12, тогда

*NPV(0,12)=-10+3x0,89+4x0.80+5x0,71=-10+2,67 + 3,20+3,55= -0,58.*

Итак, проект невыгоден для вложения капитала, поскольку *NPV(0,12)* отрицательно. При отсутствии дисконтирования (при *C = 1, q = 0*) вывод иной:

$$NPV(0) = -10 + 3 + 4 + 5 = 2,$$

проект выгоден.

Срок окупаемости и сам вывод о прибыльности проекта зависят от неизвестного дисконт-фактора *C* или даже от неизвестной дисконт-функции - ибо какие у нас основания считать будущую дисконт-функцию постоянной? Экономическая история России последних лет показывает, что банки часто меняют проценты платы за депозит. Часто предлагают использовать норму дисконта, равную *приемлемой для инвестора норме дохода на капитал.* Это значит, что экономисты явным образом обращаются к инвестору как к эксперту, который должен назвать им некоторое число исходя из своего опыта и интуиции (т.е. экономисты перекладывают свою работу на инвестора). Кроме того, при этом игнорируется изменение указанной нормы во времени,

Приведем пример исследования NPV на устойчивость (чувствительность) к малым отклонениям значений дисконт-функции. Для этого надо найти максимально возможное отклонение NPV при допустимых отклонениях значений дисконт-функции (или, если угодно, значений банковских процентов). В качестве примера рассмотрим

$$NPV = NPV(a(0), a(1), C(1), a(2), C(2), a(3), C(3)) =$$
$$= a(0) + a(1)C(1) + a(2)C(2) + a(3)C(3).$$

Предположим, что изучается устойчивость (чувствительность) для ранее рассмотренных значений

$$a(0)=-10, a(1)=3, a(2)=4, a(3)=5, C(1)=0,89, C(2)=0,80, C(3)=0,71.$$

Пусть максимально возможные отклонения *C(1), C(2), C(3)* равны $\pm$0,05. Тогда, максимум значений NPV равен

$$NPVmax = -10+3x0,94+4x0.85+5x0,76 =$$
$$-10+ 2,82 + 3,40 + 3,80 = 0,02,$$

в то время как минимум значений NPV есть

$$NPVmin = -10+3x0,84+4x0.75+5x0,66 = -10 +2,52 +3,00+3,30 = -1,18.$$

Для NPV получаем интервал от (-1,18) до (+0,02). В нем есть и положительные, и отрицательные значения. Следовательно, нет однозначного заключения - проект убыточен или выгоден. Для принятия решения не обойтись без экспертов.

Для иных характеристик, например, внутренней нормы доходности, выводы аналогичны. Дополнительные проблемы вносит неопределенность горизонта планирования, а также будущая инфляция. Если считать, что финансовый поток должен учитывать инфляцию, то это означает, что до принятия решений об инвестициях необходимо на годы вперед спрогнозировать рост цен, а это до сих пор еще не удавалось ни одной государственной или частной исследовательской структуре. Если же рост цен не учитывать, то отдаленные во времени доходы могут "растаять" в огне инфляции. На практике риски учитывают, увеличивая *q* на десяток-другой процентов.

### 3.5.8. Место статистики интервальных данных (СИД) в прикладной статистике

Кратко рассмотрим положение статистики интервальных данных среди других методов описания неопределенностей и анализа данных.

**Нечеткость и СИД.** С формальной точки зрения описание нечеткости интервалом – это частный случай описания ее нечетким множеством. В СИД функция принадлежности нечеткого множества имеет специфический вид – она равна 1 в некотором интервале и 0 вне его. Такая функция принадлежности описывается всего двумя параметрами (границами интервала). Эта простота описания делает математический аппарат СИД гораздо более прозрачным, чем аппарат теории нечеткости в общем случае. Это, в свою очередь, позволяет продвинуться дальше, чем при использовании функций принадлежности произвольного вида.

**Интервальная математика и СИД.** Можно было бы сказать, что СИД – часть интервальной математики, что СИД так соотносится с прикладной математической статистикой, как интервальная математика – с математикой в целом. Однако исторически сложилось так, что интервальная математика занимается прежде всего вычислительным погрешностями. С точки зрения интервальной математики две формулы для выборочной дисперсии, рассмотренные выше, имеют разные погрешности. А с точки зрения СИД эти две формулы задают одну и ту же функцию, и поэтому им соответствуют совпадающие нотны и рациональные объемы выборок. Интервальная математика прослеживает процесс вычислений, СИД этим не занимается. Необходимо отметить, что типовые постановки СИД могут быть перенесены в другие области математики, и, наоборот, вычислительные алгоритмы прикладной математической статистики и СИД заслуживают изучения. Однако и то, и другое – скорее дело будущего. Из уже сделанного отметим применение методов СИД при анализе такой характеристики финансовых потоков, как *NPV* – чистая текущая стоимость [27].

**Математическая статистика и СИД.** Как уже отмечалось, математическая статистика и СИД отличаются тем, в каком порядке делаются предельные переходы $n \to \infty$ и $\Delta \to 0$. При этом СИД переходит в математическую статистику при $\Delta = 0$. Правда, тогда исчезают основные особенности СИД: нотна становится равной 0, а рациональный объем выборки – бесконечности. Рассмотренные выше методы СИД разработаны в предположении, что погрешности малы (но не исчезают) и объем выборки велик. СИД расширяет классическую математическую статистику тем, что в исходных статистических данных каждое число заменяет интервалом. С другой стороны, можно считать СИД новым этапом развития математической статистики.

**Статистика объектов нечисловой природы и СИД.** Статистика объектов нечисловой природы (СОНП) расширяет область применения классической математической статистики путем включения в нее новых видов статистических данных [27]. Естественно, при этом появляются новые виды алгоритмов анализа статистических данных и новый математический аппарат (в частности, происходит переход от методов суммирования к методам оптимизации). С точки зрения СОНП частному виду новых статистических данных – интервальным данным – соответствует СИД. Напомним, что одно из двух основных понятий СИД – нотна – определяется как решение оптимизационной задачи. Однако СИД, изучая классические методы прикладной статистики применительно к интервальным данным, по математическому аппарату ближе к классике, чем другие части СОНП, например, статистика бинарных отношений.

**Робастные методы статистики и СИД.** Если понимать робастность согласно [3] как теорию устойчивости статистических методов по отношению к допустимым отклонениям исходных данных и предпосылок модели, то в СИД рассматривается одна из естественных постановок робастности. Однако в массовом сознании специалистов термин «робастность» закрепился за моделью засорения выборки большими выбросами (модель Тьюки-Хубера), хотя эта модель не имеет большого практического значения [27]. К этой модели СИД не имеет отношения.

**Теория устойчивости и СИД.** Общей схеме устойчивости [3] математических моделей социально-экономических явлений и процессов по отношению к допустимым отклонениям исходных данных и предпосылок моделей СИД полностью соответствует. Он посвящен математико-статистическим моделям, используемым при анализе статистических данных, а допустимые отклонения – это интервалы, заданные ограничениями на погрешности. СИД можно рассматривать как пример теории, в которой учет устойчивости позволил сделать нетривиальные выводы. Отметим, что с точки зрения общей схемы устойчивости [3] устойчивость по Ляпунову в

теории дифференциальных уравнений – весьма частный случай, в котором из-за его конкретности удалось весьма далеко продвинуться.

**Минимаксные методы, типовые отклонения и СИД.** Постановки СИД относятся к минимаксным. За основу берется максимально возможное отклонение. Это – подход пессимиста, используемый, например, в теории антагонистических игр. Использование минимаксного подхода позволяет подозревать СИД в завышении роли погрешностей измерения. Однако примеры изучения вероятностно-статистических моделей погрешностей, проведенные, в частности, при разработке методов оценивания параметров гамма-распределения [4], показали, что это подозрение не подтверждается. Влияние погрешностей измерений по порядку такое же, только вместо максимально возможного отклонения (нотны) приходится рассматривать математическое ожидание соответствующего отклонения (см. выше). Подчеркнем, что применение в СИД вероятностно-статистических моделей погрешностей не менее перспективно, чем минимаксных.

**Подход научной школы А.П. Вощинина и СИД.** Если в математической статистике неопределенность только статистическая, то в научной школе А.П. Вощинина - только интервальная. Можно сказать, что СИД лежит между классической прикладной математической статистикой и областью исследований научной школы А.П. Вощинина. Другое отличие состоит в том, что в этой школе разрабатывают новые методы анализа интервальных данных, а в СИД в настоящее время изучается устойчивость классических статистических методов по отношению к малым погрешностям. Подход СИД оправдывается распространенностью этих методов, однако в дальнейшем следует переходить к разработке новых методов, специально предназначенных для анализа интервальных данных.

**Анализ чувствительности и СИД.** При анализе чувствительности, как и в СИД, рассчитывают производные по используемым переменным, или непосредственно находят изменения при отклонении переменной на $\pm$10% от базового значения. Однако этот анализ делают по каждой переменной отдельно. В СИД все переменные рассматриваются совместно, и находится максимально возможное отклонение (нотна). При малых погрешностях удается на основе главного члена разложения функции в многомерный ряд Тейлора получить удобную формулу для нотны. Можно сказать, что СИД – это многомерный анализ чувствительности.

**Литература**

1. Дискуссия по анализу интервальных данных // Заводская лаборатория. 1990. Т.56. No.7, с.75-95.
2. Сборник трудов Международной конференции по интервальным и стохастическим методам в науке и технике (ИНТЕРВАЛ-92). Тт. 1,2. - М.: МЭИ, 1992, 216 с., 152 с.
3. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука, 1979. 296 с.
4. ГОСТ 11.011-83. Прикладная статистика. Правила определения оценок и доверительных границ для параметров гамма-распределения. - М.: Изд-во стандартов, 1984, 53 с.
5. Orlov A.I. // Interval Computations, 1992, No.1(3), p.44-52.
6. Орлов А.И. // Наука и технология в России. 1994. No.4(6). С.8-9.
7. Шокин Ю.И. Интервальный анализ. Новосибирск: Наука, 1981, 112 с.
8. Орлов А.И. - В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. Пермь: Изд-во Пермского государственного университета, 1990, с..89-99.
9. Орлов А.И. - В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. Пермь: Изд-во Пермского государственного университета, 1991, с.77-86.
10. Орлов А.И. - В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. Пермь: Изд-во Пермского государственного университета, 1988, с.45-55.
11. Орлов А.И. - В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. - Пермь: Изд-во Пермского государственного университета, 1995, с.114-124.
12. Орлов А.И. - В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. Пермь: Пермский государственный университет, 1993, с.149-158.
13. Биттар А.Б. Метод наименьших квадратов для интервальных данных. Дипломная работа. - М.: МЭИ, 1994. 38 с.
14. Пузикова Д.А. // Наука и технология в России. 1995. No.2(8). С.12-13.
15. Орлов А.И. // Надежность и контроль качества, 1991, № 8, с.3-8.

16. Орлов А.И. // Заводская лаборатория. 1998. Т.64. № 3. С.52-60.
17. Вощинин А.П. Метод оптимизации объектов по интервальным моделям целевой функции. - М.: МЭИ, 1987. 109 с.
18. Вощинин А.П., Сотиров Г.Р. Оптимизация в условиях неопределенности. - М.: МЭИ - София: Техника, 1989. 224 с.
19. Вощинин А.П., Акматбеков Р.А. Оптимизация по регрессионным моделям и планирование эксперимента. - Бишкек: Илим, 1991. 164 с.
20. Вощинин А.П. // Заводская лаборатория. 2000. Т.66, № 3. С.51-65.
21. Вощинин А.П. // Заводская лаборатория. 2002. Т.68, № 1. С.118-126.
22. Дывак Н.П. Разработка методов оптимального планирования эксперимента и анализа интервальных данных. Автореф. дисс. канд.. технич. наук. - М.: МЭИ, 1992. 20 с.
23. Симов С.Ж. Разработка и исследование интервальных моделей при анализе данных и проектировании экспертных систем. Автореф. дисс. канд. технич. наук. - М.: МЭИ, 1992. 20 с.
24. Орлов А.И. // Заводская лаборатория. 1999. Т.65. № 7. С.46-54.
25. Орлов А.И. // Заводская лаборатория. 1991. Т.57. № 7. С.64-66.
26. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. – Л.: Энергоатомиздат, 1985. 248 с.
27. Орлов А.И. Эконометрика. – М.: Экзамен, 2002. 576 с.
28. Дейвид Г. Порядковые статистики. – М.: Наука, 1979.
29. Колмогоров А.Н. Метод медианы в теории ошибок. – В кн.: Колмогоров А.Н. Теория вероятностей и математическая статистика: [Сб. статей]. – М.: Наука, 1986. – С.111-114.
30. Орлов А.И. Об оценивании параметров гамма-распределения. - Журнал "Обозрение прикладной и промышленной математики". 1997. Т.4. Вып.3. С.471-482.
31. Гнеденко Б.В., Хинчин А.Я. Элементарное введение в теорию вероятностей. – М.: Наука, 1970.
32. Бронштейн И.Н., Семендяев К.А. Справочник по математике для инженеров и учащихся втузов. – М.-Л.: ГИТТЛ, 1945.
33. Кендалл М., Стьюарт А. Статистические выводы и связи. – М.: Наука, 1973. 900 с.
34. Рекомендации. Прикладная статистика. Методы обработки данных. Основные требования и характеристики. – М.: ВНИИС, 1987.
35. Ляшенко Н.Н., Никулин М.С. Машинное умножение и деление независимых случайных величин // Записки научных семинаров Ленингр. Отделения Математического ин-та АН СССР, 1986, Т.153.
36. Хьюбер П. Робастность в статистике. – М.: Мир, 1984. 303 с.
37. Орлов А.И. Асимптотика решений экстремальных статистических задач // Анализ нечисловых данных в системных исследованиях. Сб. трудов. Вып.10. – М.: ВНИИ системных исследований АН СССР, 1982. – С.4-12.
38. Крамер Г. Математические методы статистики. – М.: Мир, 1975. 648 с.
39. Боровков А.А. Математическая статистика. – М.: Наука, 1984. 472 с.
40. Кендалл М., Стьюарт А. Теория распределений. – М.: Наука, 1966.
41. Смирнов Н.В. Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках // Бюллетень МГУ. Сер.А. 1939. Т.2. №2.
42. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.
43. Орлов А.И. О критериях Колмогорова и Смирнова // Заводская лаборатория. 1995. Т.61. No.7. С.59-61.
44. Гантмахер Ф.Р. Теория матриц. - М.: Наука, 1966. -576 с.
45. Розанов Ю.А. Теория вероятностей, случайные процессы и математическая статистика. - М.: Наука, 1989. - 320 с.
46. Налимов В.В., Голикова Т.И. Логические основания планирования эксперимента. – М.: Металлургия, 1976. 128 с.

**Контрольные вопросы и задачи**

1. Покажите на примерах, что в задачах принятия решений исходные данные часто имеют интервальный характер.

2. В чем особенности подхода статистики интервальных данных в задачах оценивания параметров?
3. В чем особенности подхода статистики интервальных данных в задачах проверки гипотез?
4. Какие новые нюансы проявляются в статистике интервальных данных при переходе к многомерным задачам?
5. Выполните операции над интервальными числами:
Вариант 1 - а)[1,2]+[3,4], б)[4,5]-[2,3], в)[3,4]х[5,7], г)[10,20]:[4,5];
Вариант 2 - а)[0,2]+[3,5], б)[3,5]-[2,4], в)[2,4]х[5,8], г)[15,25]:[1,5].
6. Выпишите формулу для асимптотической нотны (ошибки по абсолютной величине не превосходят константы $t$, предполагающейся малой) для функции

$$f(x_1,x_2) = 5\,(x_1)^2 + 10\,(x_2)^2 + 7\,x_1x_2.$$

Вычислите асимптотическую нотну в точке $(x_1,x_2) = (1,2)$ при $t = 0{,}1$.
7. Выпишите формулу для асимптотической нотны (ошибки по абсолютной величине не превосходят константы $t$, предполагающейся малой) для функции

$$f(x_1,x_2) = 4\,(x_1)^2 + 12\,(x_2)^2 - 3\,x_1x_2.$$

Вычислите асимптотическую нотну в точке $(x_1,x_2) = (2,1)$ при $t = 0{,}05$.

**Темы докладов, рефератов, исследовательских работ**

1. Классическая математическая статистика как предельный случай статистики интервальных данных.
2. Концепция рационального объема выборки.
3. Сравнение методов оценивания параметров и характеристик распределений в статистике интервальных данных и в классической математической статистике.
4. Подход к проверке гипотез в статистике интервальных данных.
5. Метод наименьших квадратов для интервальных данных.
6. Различные способы учета погрешностей исходных данных в статистических процедурах.
7. Статистика интервальных данных как часть теории устойчивости (с использованием монографии [3]).

## Часть 4. Заключение. Современная прикладная статистика

Подведем итоги и наметим перспективы развития методов прикладной статистики. В настоящем «Заключении» обсуждаются тенденции развития статистических методов, выделяются пять основных «точек роста». В связи с внедрением современных методов прикладной статистики обосновывается полезность понятия "высокие статистические технологии". Рассматриваются технологии использования компьютеров в прикладной статистике. Обсуждаются основные нерешенные проблемы прикладной статистики.

### 4.1. Точки роста

Отечественная литература по прикладной статистике столь же необозрима, как и мировая. Только в секции "Математические методы исследования" журнала "Заводская лаборатория" с 1960-х годов опубликовано более 1000 статей. Не будем даже пытаться перечислять коллективы исследователей или основные монографии в этой области. Отметим только одно издание. По нашему мнению, наилучшей отечественной книгой по прикладной статистике является сборник статистических таблиц Л.Н. Большева и Н.В.Смирнова [1] с подробными комментариями, играющими роль сжатого учебника и справочника.

Выделим и обсудим "точки роста" прикладной статистики, те их направления, которые представляются перспективными в будущем, в XXI веке, но пока в большинстве учебных изданий отодвинуты на задний план традиционными постановками.

При описании современного этапа развития статистических методов целесообразно выделить пять актуальных направлений, в которых развивается современная прикладная статистика, т.е. пять "точек роста": непараметрика (т.е. непараметрическая статистика), робастность, бутстреп, статистика интервальных данных, статистика нечисловых данных (в несколько иной терминологии - статистика объектов нечисловой природы). Обсудим их.

**Непараметрическая статистика**. В первой трети XX в., одновременно с параметрической статистикой, в работах Спирмена и Кендалла появились первые непараметрические методы, основанные на коэффициентах ранговой корреляции, носящих ныне имена этих статистиков. Но непараметрика, не делающая нереалистических предположений о том, что функции распределения результатов наблюдений принадлежат тем или иным параметрическим семействам распределений, стала заметной частью статистики лишь со второй трети XX века. В 30-е годы появились работы А.Н.Колмогорова и Н.В.Смирнова, предложивших и изучивших статистические критерии, носящие в настоящее время их имена. Эти критерии основаны на использовании так называемого эмпирического процесса. (Как известно, эмпирический процесс – это разность между эмпирической и теоретической функциями распределения, умноженная на квадратный корень из объема выборки.) В работе А.Н.Колмогорова 1933 г. изучено предельное распределение супремума модуля эмпирического процесса, называемого сейчас критерием Колмогорова. Затем Н.В. Смирнов исследовал супремум и инфимум эмпирического процесса, а также интеграл (по теоретической функции распределения) квадрата эмпирического процесса.

Следует отметить, что встречающееся иногда в литературе словосочетание "критерий Колмогорова-Смирнова" некорректно, поскольку эти два статистика никогда не печатались вместе и не изучали один и тот же критерий схожими методами. Корректно сочетание "критерий типа Колмогорова-Смирнова", применяемое для обозначения критериев, основанных на использовании супремума функций от эмпирического процесса [2].

После второй мировой войны развитие непараметрической статистики пошло быстрыми темпами. Большую роль сыграли работы американского статистика Ф. Вилкоксона и его школы. К настоящему времени с помощью непараметрических методов можно решать практически тот же круг статистических задач, что и с помощью параметрических. Однако для обеспечения широкого внедрения непараметрических методов необходимо провести еще целый комплекс теоретических и пилотных (т.е. пробных) прикладных работ. Все большую роль играют непараметрические оценки плотности, непараметрические методы регрессии и распознавания образов (дискриминантного

анализа). В нашей стране непараметрические методы получили достаточно большую известность после выхода в 1965 г. первого издания упомянутого выше сборника статистических таблиц Л.Н. Большева и Н.В.Смирнова [1], содержащего подробные таблицы для основных непараметрических критериев.

Тем не менее параметрические методы всё еще популярнее непараметрических, особенно среди тех прикладников, кто слабо знаком со статистическими методами. Неоднократно публиковались экспериментальные данные, свидетельствующие о том, что распределения реально наблюдаемых случайных величин, в частности, ошибок измерения, в подавляющем большинстве случаев отличны от нормальных (гауссовских). Тем не менее теоретики продолжают строить и изучать статистические модели, основанные на гауссовости, а практики - применять подобные методы и модели. Другими словами, "ищут под фонарем, а не там, где потеряли".

**Устойчивость статистических процедур (робастность)**. Если в параметрических постановках на вероятностные модели статистических данных накладываются слишком жесткие требования - их функции распределения должны принадлежать определенному параметрическому семейству, то в непараметрических, наоборот, излишне слабые - требуется лишь, чтобы функции распределения были непрерывны. При этом игнорируется априорная информация о том, каков "примерный вид" распределения. Априори можно ожидать, что учет этого "примерного вида" улучшит показатели качества статистических процедур. Развитием этой идеи является теория устойчивости (робастности) статистических процедур, в которой предполагается, что распределение исходных данных мало отличается от некоторого параметрического семейства. За рубежом эту теорию разрабатывали П. Хубер, Ф. Хампель и многие другие. Из монографий на русском языке, трактующих о робастности и устойчивости статистических процедур, самой ранней и наиболее общей была книга [3], следующей - монография [4]. Частными случаями реализации идеи робастности (устойчивости) статистических процедур являются статистика объектов нечисловой природы и статистика интервальных данных.

Имеется большое разнообразие моделей робастности в зависимости от того, какие именно отклонения от заданного параметрического семейства допускаются. Среди теоретиков наиболее популярной оказалась модель выбросов, в которой исходная выборка "засоряется" малым числом "выбросов", имеющих принципиально иное распределение. Однако эта модель представляется "тупиковой", поскольку в большинстве случаев большие выбросы либо невозможны из-за ограниченности шкалы прибора либо интервала изменения измеряемой величины, либо от них можно избавиться, применяя лишь статистики, построенные по центральной части вариационного ряда. Кроме того, в подобных моделях обычно считается известной частота засорения, что в сочетании со сказанным выше делает их малопригодными для практического использования.

Более перспективным представляется, например, модель малых отклонений распределений, в которой расстояние между распределением каждого элемента выборки и базовым распределением не превосходит заданной малой величины, и модель статистики интервальных данных.

**Бутстреп (размножение выборок)**. Другое из упомянутых выше направлений - бутстреп - связано с интенсивным использованием возможностей компьютеров. Основная идея состоит в том, чтобы теоретическое исследование заменить вычислительным экспериментом. Например, вместо описания выборки распределением из параметрического семейства строим большое число "похожих" выборок, т.е. "размножаем" выборку. Затем вместо оценивания характеристик (и параметров) и проверки гипотез на основе свойств теоретического распределения решаем эти задачи вычислительным методом, рассчитывая интересующие нас статистики по каждой из "похожих" выборок и анализируя полученные при этом распределения. Например, вместо того, чтобы теоретическим путем находить распределение статистики, доверительные интервалы и другие характеристики, моделируют большое число выборок, похожих на исходную, затем рассчитывают соответствующие значения интересующей исследователя статистики и изучают их эмпирическое распределение. Квантили этого распределения задают доверительные интервалы, и т.д.

Термин "бутстреп" мгновенно получил широкую известность после первой же статьи Б.Эфрона 1979 г. по этой тематике. Он сразу же стал обсуждаться в массе публикаций, в том числе и научно-популярных. В "Заводской лаборатории" № 10 за 1987 г. была помещена подборка статей по

бутстрепу. На русском языке выпущен сборник статей Б. Эфрона [5]. Основная идея бутстрепа по Б. Эфрону состоит в том, что методом Монте-Карло (статистических испытаний) многократно извлекаются выборки из эмпирического распределения. Эти выборки, естественно, являются вариантами исходной, напоминают ее.

Сама по себе идея "размножения выборок" была известна гораздо раньше. Одна из статей Б. Эфрона в сборнике [5] называется так: "Бутстреп-методы: новый взгляд на метод складного ножа". Упомянутый "метод складного ножа" (*jackknife*) предложен М. Кенуем еще в 1949 г., за 30 лет до появления статьи Б.Эфрона. "Размножение выборок" при этом осуществляется путем исключения одного наблюдения. Таким путем для выборки объема $n$ получаем $n$ "похожих" на нее выборок объема $(n - 1)$ каждая. Если же исключать по 2 наблюдения, то число "похожих" выборок возрастает до $n(n - 1)/2$ объема $(n - 2)$ каждая.

Преимущества и недостатки бутстрепа как статистического метода в сравнении с рядом аналогичных методов обсуждаются ниже. Необходимо подчеркнуть, что бутстреп по Эфрону - лишь один из вариантов методов "размножения выборки" (*resampling*), и, на наш взгляд, не самый удачный. Метод "складного ножа" представляется более полезным. На его основе можно сформулировать следующую простую практическую рекомендацию.

Предположим, что Вы по выборке делаете какие-либо статистические выводы. Вы хотите узнать также, насколько эти выводы устойчивы. Если у Вас есть другие (контрольные) выборки, описывающие то же явление, то Вы можете применить к ним ту же статистическую процедуру и сравнить результаты. А если таких выборок нет? Тогда Вы можете их построить искусственно. Берете исходную выборку и исключаете один элемент. Получаете похожую выборку (она взята из того же распределения, только объем на единицу меньше). Затем возвращаете этот элемент выборки и исключаете другой. Получаете вторую похожую выборку. Поступая таким образом со всеми элементами исходной выборки, получаете столько выборок, похожих на исходную, каков ее объем. Остается обработать их тем же способом, что и исходную, и изучить устойчивость получаемых выводов - разброс оценок параметров, частоты принятия или отклонения гипотез и т.д.

Можно изменять не выборку, а сами данные. Поскольку всегда имеются погрешности измерения, то реальные данные - это не числа, а интервалы (результат измерения плюс-минус погрешность). Нужна статистическая теория анализа таких данных.

**Статистика интервальных данных**. Перспективное и быстро развивающееся направление последних лет - статистика интервальных данных. Речь идет о развитии методов прикладной математической статистики в ситуации, когда статистические данные - не числа, а интервалы, в частности, порожденные наложением ошибок измерения на значения случайных величин.

Статистика интервальных данных идейно связана с интервальной математикой, в которой в роли чисел выступают интервалы. Это направление математики является дальнейшим развитием известных правил приближенных вычислений, посвященных выражению погрешностей суммы, разности, произведения, частного через погрешности тех чисел, над которыми осуществляются перечисленные операции. К настоящему времени удалось решить, в частности, ряд задач теории интервальных дифференциальных уравнений, в которых коэффициенты, начальные условия и решения описываются с помощью интервалов.

Одна из ведущих научных школ в области статистики интервальных данных - это школа проф. А.П. Вощинина, активно работающая с конца 70-х годов. В частности, ее представителями изучены проблемы регрессионного анализа, планирования эксперимента, сравнения альтернатив и принятия решений в условиях интервальной неопределенности.

Рассмотрим другое направление в статистике интервальных данных, которое также представляется перспективным. В нем развиваются асимптотические методы статистического анализа интервальных данных при больших объемах выборок и малых погрешностях измерений. В отличие от классической математической статистики, сначала устремляется к бесконечности объем выборки и только потом - уменьшаются до нуля погрешности. В частности, с помощью такой асимптотики в начале 1980-х годов были сформулированы правила выбора метода оценивания параметров гамма-распределения в ГОСТ 11.011-83 [6].

В рамках рассматриваемого научного направления разработана общая схема исследования, включающая расчет нотны (максимально возможного отклонения статистики, вызванного интервальностью исходных данных) и рационального объема выборки (превышение которого не дает существенного повышения точности оценивания). Она применена к оцениванию математического ожидания, дисперсии, коэффициента вариации, параметров гамма-распределения и характеристик аддитивных статистик, при проверке гипотез о параметрах нормального распределения, в том числе с помощью критерия Стьюдента, а также гипотезы однородности с помощью критерия Смирнова. Разработаны подходы к рассмотрению интервальных данных в основных постановках регрессионного, дискриминантного и кластерного анализов. В частности, изучено влияние погрешностей измерений и наблюдений на свойства алгоритмов регрессионного анализа, разработаны способы расчета нотн и рациональных объемов выборок, введены и исследованы новые понятия многомерных и асимптотических нотн, доказаны соответствующие предельные теоремы. Начата разработка интервального дискриминантного анализа, в частности, рассмотрено влияние интервальности данных на введенный в главе 3.2 показатель качества классификации. Изучено асимптотическое поведение оценок метода моментов и оценок максимального правдоподобия (а также более общих оценок минимального контраста), проведено асимптотическое сравнение этих методов в случае интервальных данных. Найдены общие условия, при которых, в отличие от классической математической статистики, метод моментов дает более точные оценки, чем метод максимального правдоподобия (глава 3.5).

В области асимптотической статистики интервальных данных российская наука имеет мировой приоритет. Во все виды статистического программного обеспечения включают алгоритмы интервальной статистики, "параллельные" обычно используемым алгоритмам прикладной математической статистики. Это позволяет в явном виде учесть наличие погрешностей у результатов наблюдений.

**Статистика объектов нечисловой природы как часть прикладной статистики.** Напомним, что согласно общепринятой в настоящее время классификации статистических методов прикладная статистика делится на следующие четыре области:

статистика (числовых) случайных величин;

многомерный статистический анализ;

статистика временных рядов и случайных процессов;

статистика объектов нечисловой природы.

Первые три из этих областей являются классическими. Они были хорошо известны еще в первой половине XX в. Остановимся на четвертой, сравнительно недавно вошедшей в массовое сознание специалистов. Ее именуют также статистикой нечисловых данных или попросту нечисловой статистикой. Анализ динамики развития прикладной статистики приводит к выводу, что в XXI в. она станет центральной областью прикладной статистики, поскольку содержит наиболее общие подходы и результаты.

Исходный объект в прикладной математической статистике - это выборка. В вероятностной теории статистики выборка - это совокупность независимых одинаково распределенных случайных элементов. Какова природа этих элементов? В классической математической статистике элементы выборки - это числа. В многомерном статистическом анализе - вектора. А в нечисловой статистике элементы выборки - это объекты нечисловой природы, которые нельзя складывать и умножать на числа. Другими словами, объекты нечисловой природы лежат в пространствах, не имеющих векторной структуры. Примерами объектов нечисловой природы являются:

значения качественных признаков, т.е. результаты кодировки объектов с помощью заданного перечня категорий (градаций);

упорядочения (ранжировки) образцов продукции (при оценке её технического уровня и конкурентоспособности)) или заявок на проведение научных работ (при проведении конкурсов на выделение грантов), описывающие мнения экспертов;

классификации, т.е. разбиения совокупности объектов на группы сходных между собой (кластеры);

толерантности, т.е. бинарные отношения, описывающие сходство объектов между собой, например, сходство тематики научных работ, которое оценивается экспертами с целью рационального формирования экспертных советов внутри определенной области науки;

результаты парных сравнений или контроля качества продукции по альтернативному признаку ("годен" - "брак"), т.е. последовательности из 0 и 1;

множества (обычные или нечеткие), например, зоны, пораженные коррозией; топокарты, полученные при кинетокардиографии; перечни возможных причин аварии, составленные экспертами независимо друг от друга; нечеткие экспертные оценки качества газовых плит;

слова, предложения, тексты;

вектора, координаты которых - совокупность значений разнотипных признаков, например, результат составления статистического отчета о научно-технической деятельности (т.н. форма № 1-наука) или заполненная компьютеризированная история болезни, в которой часть признаков носит качественный характер, а часть - количественный;

ответы на вопросы экспертной, маркетинговой или социологической анкеты, часть из которых носит количественный характер (возможно, интервальный), часть сводится к выбору одной из нескольких подсказок, а часть представляет собой тексты; и т.д.

Интервальные данные также можно рассматривать как пример объектов нечисловой природы, а именно, как частный случай нечетких множеств.

С начала 70-х годов под влиянием запросов прикладных исследований в социально-экономических, технических, медицинских науках в России активно развивается статистика объектов нечисловой природы, известная также как статистика нечисловых данных или нечисловая статистика. В создании этой сравнительно новой области эконометрики и прикладной математической статистики приоритет принадлежит российским ученым.

Большую роль сыграл основанный в 1973 г. научный семинар "Экспертные оценки и анализ данных". В 1960-е годы советское научное сообщество стало интересоваться методами экспертных оценок (об их истории и современном состоянии см. [7, 8]). Как следствие, началось знакомство с конкретными математизированными теориями, связанными с этими методами. Речь идет о репрезентативной теории измерений, ставшей известной в нашей стране по статье П. Суппеса и Дж. Зинеса в сборнике [9] и книге И. Пфанцагля [10], о теории нечеткости, современный этап которой начался с работ Л.А.Заде [11], теории парных сравнений, описанной в монографии Г.Дэвида [12]. К этому кругу идей примыкают теория случайных множеств (см., например, книгу Ж. Матерона [13]) и методы многомерного шкалирования (описаны, в частности, в монографиях А.Ю.Терехиной [14] и В.Т.Перекреста [15]). Но наибольшее влияние оказали идеи американского исследователя проф. Дж. Кемени, который аксиоматически ввел расстояние между ранжировками (теперь оно именуется в литературе расстоянием Кемени) и предложил использовать в качестве средней величины решение оптимизационной задачи (теперь - медиана Кемени). Его скромная книжка [16], написанная в соавторстве с Дж. Снеллом, породила большой поток исследований.

В течение 1970-х годов на основе запросов теории экспертных оценок (а также социологии, экономики, техники и медицины) развивались конкретные направления статистики объектов нечисловой природы. Были установлены связи между конкретными видами таких объектов, разработаны для них вероятностные модели. Научные итоги этого периода подведены в монографиях [3, 17, 18].

Следующий этап - выделение статистики объектов нечисловой природы в качестве самостоятельного направления в прикладной статистике, ядром которого являются методы статистического анализа данных произвольной природы. Программа развития этого нового научного направления впервые была сформулирована в статье [19]. Реализация этой программы была осуществлена в 1980-е годы. Для работ этого периода характерна сосредоточенность на внутренних проблемах нечисловой статистики. Ссылки на конкретные монографии, сборники, статьи и иные публикации нескольких десятков авторов приведены в главе 3.4. Отметим лишь сборник научных статей [20], полностью посвященный нечисловой статистике.

К 1990-м годам статистика объектов нечисловой природы с теоретической точки зрения была достаточно хорошо развита, основные идеи, подходы и методы были разработаны и изучены

математически, в частности, доказано достаточно много теорем. Однако она оставалась недостаточно апробированной на практике. Это было связано как с ее сравнительной молодостью, так и с общеизвестными особенностями организации науки в 1980-е годы, когда отсутствовали достаточно сильные стимулы к тому, чтобы теоретики занялись широким внедрением своих результатов. И в 1990-е годы наступило время от теоретических математико-статистических исследований перейти к применению полученных результатов на практике.

Важно отметить, что в статистике нечисловых данных, как и в других областях прикладной статистики и прикладной математики вообще, одна и та же математическая схема может с успехом применяться и в технических исследованиях, и в менеджменте, и в экономике, и в геологии, и в медицине, и в социологии, и для анализа экспертных оценок, и во многих иных областях. А потому ее лучше всего формулировать и изучать в наиболее общем виде, для объектов произвольной природы.

**Основные идеи статистики объектов нечисловой природы.** В чем принципиальная новизна нечисловой статистики? Для классической математической статистики характерна операция сложения. При расчете выборочных характеристик распределения (выборочное среднее арифметическое, выборочная дисперсия и др.), в регрессионном анализе и других областях этой научной дисциплины постоянно используются суммы. Математический аппарат - законы больших чисел, Центральная предельная теорема и другие теоремы - нацелены на изучение сумм. В нечисловой же статистике нельзя использовать операцию сложения, поскольку элементы выборки лежат в пространствах, где нет операции сложения. Методы обработки нечисловых данных основаны на принципиально ином математическом аппарате - на применении различных расстояний в пространствах объектов нечисловой природы.

Кратко рассмотрим несколько идей, развиваемых в статистике объектов нечисловой природы для данных, лежащих в пространствах произвольного вида. Решаются классические задачи описания данных, оценивания, проверки гипотез - но для неклассических данных, а потому неклассическими методами.

Первой обсудим проблему определения средних величин. В рамках репрезентативной теории измерений удается указать вид средних величин, соответствующих тем или иным шкалам измерения (см. главу 2.1). В классической математической статистике эмпирические и теоретические средние величины вводят с помощью операций сложения (выборочное среднее арифметическое, математическое ожидание) или упорядочения (выборочная и теоретическая медианы). В пространствах произвольной природы средние значения нельзя определить с помощью операций сложения или упорядочения. Теоретические и эмпирические средние приходится вводить как решения экстремальных задач. Для теоретического среднего это - задача минимизации математического ожидания (в классическом смысле) расстояния от случайного элемента со значениями в рассматриваемом пространстве до фиксированной точки этого пространства (минимизируется указанная функция от этой точки). Для эмпирического среднего математическое ожидание берется по эмпирическому распределению, т.е. берется сумма расстояний от некоторой точки до элементов выборки и затем минимизируется по этой точке. При этом как эмпирическое, так и теоретическое средние как решения экстремальных задач могут быть не единственными элементами пространства, а описываться множествами таких элементов, которые могут оказаться и пустыми. Несмотря на возможность неоднозначности или пустоты решений экстремальных задач, удалось сформулировать и доказать законы больших чисел для средних величин, определенных указанным образом, т.е. установить сходимость эмпирических средних к теоретическим.

Хорошая теория дает больше того, что от нее вначале ожидалось. Удалось установить, что методы доказательства законов больших чисел допускают существенно более широкую область применения, чем та, для которой они были разработаны. А именно, с помощью этих методов удалось изучить асимптотику решений экстремальных статистических задач, к которым, как известно, сводится большинство постановок прикладной статистики. В частности, кроме законов больших чисел установлена и состоятельность оценок минимального контраста, в том числе оценок максимального правдоподобия и робастных оценок. К настоящему времени подобные оценки изучены также и в интервальной статистике.

В статистике в пространствах произвольной природы большую роль играют непараметрические оценки плотности, используемые, в частности, в различных алгоритмах регрессионного, дискриминантного, кластерного анализов. В нечисловой статистике предложен и изучен ряд типов непараметрических оценок плотности в пространствах произвольной природы, в частности, доказана их состоятельность, изучена скорость сходимости и установлен примечательный факт совпадения наилучшей скорости сходимости в произвольном случае с той, которая имеет быть в классической математико-статистической теории для числовых случайных величин.

Дискриминантный, кластерный, регрессионный анализы в пространствах произвольной природы основаны либо на параметрической теории - и тогда применяется подход, связанный с асимптотикой решения экстремальных статистических задач - либо на непараметрической теории - и тогда используются алгоритмы на основе непараметрических оценок плотности.

Для проверки гипотез могут быть использованы статистики интегрального типа, в частности, типа омега-квадрат. Любопытно, что предельная теория таких статистик, построенная первоначально в классической постановке [21], приобрела естественный (завершенный, изящный) вид именно для пространств произвольного вида [22], поскольку при этом удалось провести рассуждения, опираясь на базовые математические соотношения, а не на те частные (с общей точки зрения), что были связаны с конечномерным пространством.

Представляют практический интерес результаты, связанные с конкретными областями статистики нечисловых данных. В частности, со статистикой нечетких и случайных множеств (напомним, что теория нечетких множеств в определенном смысле сводится к теории случайных множеств), с непараметрической теорией парных сравнений, с аксиоматическим введением метрик в конкретных пространствах объектов нечисловой природы, и с рядом других конкретных постановок.

Для анализа нечисловых, в частности, экспертных данных весьма важны методы классификации. С другой стороны, наиболее естественно ставить и решать задачи классификации, основанные на использовании расстояний или показателей различия, в рамках статистики нечисловых данных. Это касается как распознавания образов с учителем (другими словами, дискриминантного анализа), так и распознавания образов без учителя (т.е. кластерного анализа).

Статистические методы анализа нечисловых данных особенно хорошо приспособлены для применения в экономике, социологии и экспертных оценках, поскольку в этих областях от 50% до 90% данных являются нечисловыми.

Итак, статистика нечисловых данных является центром прикладной статистики. А ее теоретическая основа – статистика в пространствах произвольной природы – является стержнем математической статистики.

**Другие точки роста.** Выше рассмотрены пять "точек роста" прикладной статистики. Разумеется, они не исчерпывают все многообразие фронта научных исследований в рассматриваемых областях. Кроме того, мы почти не затронули разнообразные применения статистических методов в конкретных прикладных исследованиях и разработках. Много интересных проблем есть в планировании экспериментов, особенно кинетических (см., например, [23]), при анализе проблем надежности, в новых статистических методах управления качеством продукции [7], в том числе в связи с идеями Г. Тагути, при анализе рисков, в вопросах экологии и промышленной безопасности и др.

В течение последних более чем 60 лет в России наблюдается огромный разрыв между государственной статистикой и научным сообществом специалистов по статистическим методам (подробнее об этом см. [24] и приложение 2). В учебнике по истории статистики [25] даже не упоминаются имена членов-корреспондентов АН СССР Н.В.Смирнова и Л.Н. Большева! А ведь они – единственные представители именно математической статистики как таковой в Академии наук в XX в. (еще ряд членов Академии наук имели математическую статистику среди своих интересов, но Н.В. Смирнов и Л.Н. Большев занимались практически только ею).

### 4.2. Высокие статистические технологии

При практическом использовании методов прикладной статистики применяются не отдельные методы описания данных, оценивания, проверки гипотез, а развернутые цельные процедуры - так называемые "статистические технологии". Понятие "статистическая технология" аналогично понятию "технологический процесс" в теории и практике организации производства.

**Статистические технологии.** Статистический анализ конкретных данных, как правило, включает в себя целый ряд процедур и алгоритмов, выполняемых последовательно, параллельно или по более сложной схеме. В частности, с точки зрения организатора прикладного статистического исследования можно выделить следующие этапы:

- планирование статистического исследования (включая разработку анкет, бланков наблюдения и учета и других форм сбора данных; их апробацию; подготовку сценариев интервью и анализа данных и т.п.);
- организация сбора необходимых статистических данных по оптимальной или рациональной программе (планирование выборки, создание организационной структуры и подбор команды статистиков, подготовка кадров, которые будут заниматься сбором данных, а также контролеров данных и т.п.);
- непосредственный сбор данных и их фиксация на тех или иных носителях (с контролем качества сбора и отбраковкой ошибочных данных по соображениям предметной области);
- первичное описание данных (расчет различных выборочных характеристик, функций распределения, непараметрических оценок плотности, построение гистограмм, корреляционных полей, различных таблиц и диаграмм и т.д.),
- оценивание тех или иных числовых или нечисловых характеристик и параметров распределений (например, непараметрическое интервальное оценивание коэффициента вариации или восстановление зависимости между откликом и факторами, т.е. оценивание функции),
- проверка статистических гипотез (иногда их цепочек - после проверки предыдущей гипотезы принимается решение о проверке той или иной последующей гипотезы; например, после проверки адекватности линейной регрессионной модели и отклонения этой гипотезы может проверяться адекватность квадратичной модели),
- более углубленное изучение, т.е. одновременное применение различных алгоритмов многомерного статистического анализа, алгоритмов диагностики и построения классификации, статистики нечисловых и интервальных данных, анализа временных рядов и др.;
- проверка устойчивости полученных оценок и выводов относительно допустимых отклонений исходных данных и предпосылок используемых вероятностно-статистических моделей, в частности, изучение свойств оценок методом размножения выборок и другими численными методами;
- применение полученных статистических результатов в прикладных целях, т.е. для формулировки выводов в терминах содержательной области (например, для диагностики конкретных материалов, построения прогнозов, выбора инвестиционного проекта из предложенных вариантов, нахождения оптимальных режима осуществления технологического процесса, подведения итогов испытаний образцов технических устройств и др.),
- составление итоговых отчетов, в частности, предназначенных для тех, кто не является специалистами в статистических методах анализа данных, в том числе для руководства - "лиц, принимающих решения".

Возможны и иные структуризации различных статистических технологий. Важно подчеркнуть, что квалифицированное и результативное применение статистических методов - это отнюдь не проверка одной отдельно взятой статистической гипотезы или оценка характеристик или параметров одного заданного распределения из фиксированного семейства. Подобного рода операции - только отдельные кирпичики, из которых складывается статистическая технология.

Итак, процедура статистического анализа данных – это информационный технологический процесс, другими словами, та или иная информационная технология. Статистическая информация подвергается разнообразным операциям (последовательно, параллельно или по более сложным схемам). В настоящее время об автоматизации всего процесса статистического анализа данных говорить было бы несерьезно, поскольку имеется слишком много нерешенных проблем, вызывающих дискуссии среди статистиков. Наличие разногласий – причина того, что так

называемые «экспертные системы в области статистического анализа данных» пока не стали рабочим инструментом статистиков.

**Проблема "стыковки" алгоритмов.** В литературе статистические технологии рассматриваются явно недостаточно. В частности, обычно все внимание сосредотачивается на том или ином элементе технологической цепочки, а переход от одного элемента к другому остается в тени. Между тем проблема "стыковки" статистических алгоритмов, как известно, требует специального рассмотрения (см. главу 2.3.5), поскольку в результате использования предыдущего алгоритма зачастую нарушаются условия применимости последующего. В частности, результаты наблюдений могут перестать быть независимыми, может измениться их распределение и т.п.

Так, вполне резонной выглядит рекомендация: сначала разбейте данные на однородные группы, а потом в каждой из групп проводите статистическую обработку, например, регрессионный анализ. Однако эта рекомендация под кажущейся прозрачностью содержит подводные камни. Действительно, как поставить задачу в вероятностно-статистических терминах? Если, как обычно, примем, что исходные данные - это выборка, т.е. совокупность независимых одинаково распределенных случайных элементов, то классификация приведет к разбиению этих элементов на группы. В каждой группе элементы будут зависимы между собой, а их распределение будет зависеть от группы, куда они попали. Отметим, что в типовых ситуациях границы классов стабилизируются, а это значит, что асимптотически элементы кластеров становятся независимыми. Однако их распределение не может быть нормальным. Например, если исходное распределение было нормальным, то распределения в классах будет усеченным нормальным. Это означает, что необходимо пользоваться непараметрическими методами.

Разберем другой пример. При проверке статистических гипотез большое значение имеют такие хорошо известные характеристики статистических критериев, как уровень значимости и мощность. Методы их расчета и использования при проверке одной гипотезы обычно хорошо известны. Если же сначала проверяется одна гипотеза, а потом с учетом результатов ее проверки (конкретнее, если первая гипотеза принята) - вторая, то итоговую процедуру также можно рассматривать как проверку некоторой (более сложной) статистической гипотезы. Она имеет характеристики (уровень значимости и мощность), которые, как правило, нельзя простыми формулами выразить через характеристики двух составляющих гипотез, а потому они обычно неизвестны. Лишь в некоторых простых случаях характеристики итоговой процедуры можно рассчитать. В результате итоговую процедуру нельзя рассматривать как научно обоснованную, она относится к эвристическим алгоритмам. Конечно, после соответствующего изучения, например, методом Монте-Карло, она может войти в число научно обоснованных процедур прикладной статистики.

**Термин "высокие статистические технологии".** Термин "высокие технологии" популярен в современной научно-технической литературе. Он используется для обозначения наиболее передовых технологий, опирающихся на последние достижения научно-технического прогресса. Есть такие технологии и среди технологий статистического анализа данных - как в любой интенсивно развивающейся научно-практической области.

Примеры высоких статистических технологий и входящих в них алгоритмов анализа данных, подробный анализ современного состояния и перспектив развития даны выше при обсуждении “точек роста”. В качестве "высоких статистических технологий" были выделены технологии непараметрического анализа данных; устойчивые (робастные) технологии; технологии, основанные на размножении выборок, на использовании достижений статистики нечисловых данных и статистики интервальных данных.

Обсудим пока не вполне привычный термин "высокие статистические технологии". Каждое из трех слов несет свою смысловую нагрузку.

"Высокие", как и в других областях, означает, что статистическая технология опирается на современные достижения статистической теории и практики, в частности, на достижения теории вероятностей и прикладной математической статистики. При этом "опирается на современные научные достижения" означает, во-первых, что математическая основа технологии получена сравнительно недавно в рамках соответствующей научной дисциплины, во-вторых, что алгоритмы

расчетов разработаны и обоснованы в соответствии в нею (а не являются т.н. "эвристическими"). Со временем новые подходы и результаты могут заставить пересмотреть оценку применимости и возможностей технологии, привести к замене ее на более современную. В противном случае "высокие статистические технологии" переходят в "классические статистические технологии", такие, как метод наименьших квадратов. Итак, высокие статистические технологии - плоды недавних серьезных научных исследований. Здесь два ключевых понятия - "молодость" технологии (во всяком случае, не старше 50 лет, а лучше - не старше 10 или 30 лет) и опора на "высокую науку".

Термин "статистические" привычен, но коротко разъяснить его нелегко. Проще сослаться на введение и все содержание настоящего учебника, на энциклопедию [26], книги [1, 7] и др. В частности, статистические данные – это результаты измерений, наблюдений, испытаний, анализов, опытов. А "статистические технологии" - это технологии анализа статистических данных.

Наконец, редко используемый применительно к статистике термин "технологии". Статистический анализ данных, как правило, включает в себя целый ряд процедур и алгоритмов, выполняемых последовательно, параллельно или по более сложной схеме. Структура типовой статистической технологии описана выше. Обработка статистических данных - это информационный технологический процесс.

**Всегда ли нужны "высокие статистические технологии"?** "Высоким статистическим технологиям" противостоят, естественно, "низкие статистические технологии" (а между ними расположены "классические статистические технологии"). "Низкие статистические технологии" - это те технологии, которые не соответствуют современному уровню науки и практики. Обычно они одновременно и устарели, и не вполне адекватны сути решаемых статистических задач.

Примеры таких технологий неоднократно критически рассматривались, в том числе и на страницах этой книги. Достаточно вспомнить критику использования критерия Стьюдента для проверки однородности при отсутствии нормальности и равенства дисперсии. Или применение критерия Вилкоксона для проверки совпадения теоретических медиан или функций распределения двух выборок. Или использование классических процентных точек критериев Колмогорова и омега-квадрат в ситуациях, когда параметры оцениваются по выборке и эти оценки подставляются в "теоретическую" функцию распределения. На первый взгляд вызывает удивление устойчивость «низких статистических технологий», их постоянное возрождение во все новых статьях, монографиях, учебниках. Поэтому, как ни странно, наиболее "долгоживущими" оказываются не работы, посвященные новым научным результатам, а публикации, разоблачающие ошибки, типа статьи [27]. Прошло около 20 лет с момента ее публикации, но она по-прежнему актуальна, поскольку ошибочное применение критериев Колмогорова и омега-квадрат по-прежнему распространено.

Целесообразно отметить по крайней мере четыре обстоятельства, которые определяют эту устойчивость ошибок. Во-первых, прочно закрепившаяся традиция. Так, многие учебники по курсам типа "Общей теории статистики", если беспристрастно проанализировать их содержание, состоят в основном из введения в прикладную статистику. Иногда изложение идет в стиле "низких статистических технологий", т.е. на уровне 1950-х годов, а во многом и на уровне начала XX в. К "низкой" прикладной статистике добавлена некоторая информация о деятельности органов Госкомстата РФ. Новое поколение, обучившись «низким» подходам, идеям, алгоритмам, их использует, а с течением времени и достижением должностей, ученых званий и степеней – пишет новые учебники со старыми ошибками.

Второе обстоятельство связано с большими трудностями при оценке экономической эффективности применения статистических методов вообще и при оценке вреда от применения ошибочных методов в частности. (А без такой оценки как докажешь, что "высокие статистические технологии" лучше "низких"?) При оценке вреда от применения ошибочных методов приходится учитывать, что общий успех в конкретной инженерной или научной работе вполне мог быть достигнут вопреки применению ошибочных методов, за счет "запаса прочности" других составляющих общей работы. Например, преимущество одного технологического приема над другим можно продемонстрировать как с помощью критерия Крамера-Уэлча проверки равенства

математических ожиданий (что правильно), так и с помощью двухвыборочного критерия Стьюдента (что, вообще говоря, неверно, т.к. обычно не выполняются условия применимости этого критерия - нет ни нормальности распределения, ни равенства дисперсий).

Третье существенное обстоятельство – трудности со знакомством с высокими статистическими технологиями. В нашей стране в силу ряда исторических обстоятельств развития статистических методов в течение последних 10 лет только журнал "Заводская лаборатория" предоставлял такие возможности. К сожалению, поток современных отечественных и переводных статистических книг, выпускавшихся ранее, в частности, издательствами "Наука", "Мир", “Финансы и статистика”, практически превратился в узкий ручеек… Возможно, более существенным является влияние естественной задержки во времени между созданием "новых статистических технологий" и написанием полноценной и объемной учебной и методической литературы. Она должна позволять знакомиться с новой методологией, новыми методами, теоремами, алгоритмами, методами расчетов и интерпретации их результатов, статистическими технологиями в целом не по кратким оригинальным статьям, а при обычном вузовском и последипломном обучении.

И, наконец, наиболее важное. Всегда ли нужны высокие статистические технологии? Приведем аналогию - нужна ли современная сельскохозяйственная техника для обработки приусадебного участка? Нужны ли трактора и комбайны? Может быть, достаточно технологий, основанных на использовании лопаты? Вернемся к данным государственной статистики. Применяются статистические технологии первичной обработки (описания) данных, основанные на построении разнообразных таблиц, диаграмм, графиков. Большинство потребителей статистической информации это представление данных удовлетворяет. Итак, чтобы высокие статистические технологии успешно использовались, необходимы два условия:

- чтобы они были *объективно* нужны для решения практической задачи;
- чтобы потенциальный пользователь технологий *субъективно* понимал это.

Таким образом, весь арсенал реально используемых в настоящее время эконометрических и статистических технологий можно распределить по трем потокам:

- высокие статистические технологии;
- классические статистические технологии,
- низкие статистические технологии.

Под классическими статистическими технологиями, как уже отмечалось, понимаем технологии почтенного возраста, сохранившие свое значение для современной статистической практики. Таковы технологии на основе метода наименьших квадратов (включая методы точечного оценивания параметров прогностической функции, непараметрические методы доверительного оценивания параметров и прогностической функции в целом, проверок различных гипотез о них), статистик типа Колмогорова, Смирнова, омега-квадрат, непараметрических коэффициентов корреляции Спирмена и Кендалла (относить их только к методам анализа ранжировок - значит делать уступку "низким статистическим технологиям") и многих других статистических процедур.

**Основная современная проблема в области статистических технологий** состоит в том, чтобы в конкретных эконометрических исследованиях использовались только технологии первых двух типов.

Каковы возможные пути решения этой проблемы? Бороться с конкретными невеждами - дело почти безнадежное. Конечно, необходима демонстрация квалифицированного применения высоких статистических технологий. В 1960-70-х годах этим занималась Лаборатория статистических методов акад. А.Н. Колмогорова в МГУ им. М.В. Ломоносова. В секции "Математические методы исследования" журнала "Заводская лаборатория" за последние 40 лет опубликовано более 1000 статей в стиле "высоких статистических технологий". В настоящее время действует Институт высоких статистических технологий и эконометрики МГТУ им. Н.Э.Баумана и целый ряд других научных коллективов, работающих на уровне "высоких статистических технологий".

Очевидно, самое основное - это обучение. Какие бы новые научные результаты ни были получены, если они остаются неизвестными студентам, то новое поколение исследователей и инженеров, экономистов и менеджеров, других специалистов вынуждено осваивать их поодиночке, в порядке самообразования, а то и переоткрывать заново. Т.е. зачастую новые научные результаты

практически исчезают из оборота научной и практической информации, едва появившись.
Как ни странно это может показаться, избыток научных публикаций превратился в тормоз развития науки. По нашим данным, к настоящему времени по статистическим технологиям опубликовано не менее миллиона статей и книг, в основном во второй половине XX в., из них не менее 100 тысяч являются актуальными для современного специалиста. При этом реальное число публикаций, которые способен освоить исследователь за свою профессиональную жизнь, по нашей оценке, не превышает 2-3 тысяч. Итак, каждый специалист в области прикладной статистики знаком не более чем с 2-3% актуальных для него литературных источников. Поскольку существенная часть публикаций заражена "низкими статистическими технологиями", то исследователь-самоучка, увы, имеет мало шансов выйти на уровень "высоких статистических технологий". С подтверждениями этого печального вывода постоянно приходится сталкиваться. Одновременно приходится констатировать, что масса полезных результатов погребена в изданиях прошлых десятилетий и имеет мало шансов пробиться в ряды используемых в настоящее время "высоких статистических технологий" без специально организованных усилий современных специалистов.

Итак, основное - обучение. Несколько огрубляя, можно сказать так: что попало в учебные курсы и соответствующие учебные пособия - то сохраняется, что не попало - то пропадает.

**Необходимость высоких статистических технологий.** Может возникнуть естественный вопрос: зачем нужны высокие статистические технологии, разве недостаточно обычных статистических методов? Специалисты по прикладной статистике справедливо считают и доказывают своими теоретическими и прикладными работами, что совершенно недостаточно. Так, совершенно очевидно, что многие данные в информационных системах имеют нечисловой характер, например, являются словами или принимают значения из конечных множеств. Нечисловой характер имеют и упорядочения, которые дают эксперты или менеджеры, например, выбирая главную цель, следующую по важности и т.д. Значит, нужна статистика нечисловых данных. Мы ее построили. Далее, многие величины известны не абсолютно точно, а с некоторой погрешностью - от и до. Другими словами, исходные данные - не числа, а интервалы. Нужна статистика интервальных данных. Мы ее развиваем. В широко известной монографии по контроллингу [28] на с.138 хорошо сказано: "Нечеткая логика - мощный элегантный инструмент современной науки, который на Западе (и на Востоке - в Японии, Китае - А.О.) можно встретить в десятках изделий - от бытовых видеокамер до систем управления вооружениями, - у нас до самого последнего времени был практически неизвестен". Напомним, первая монография российского автора по теории нечеткости содержит основы высоких статистических технологий, связанные с анализом выборок нечетких множеств (см. книгу [29]). Ни статистики нечисловых данных, ни статистики интервальных данных, ни статистики нечетких данных не было и не могло быть в классической статистике. Все это - высокие статистические технологии. Они разработаны за последние десятилетия. А обычные вузовские курсы по общей теории статистики и по математической статистике разбирают научные результаты, полученные в первой половине XX века.

Важная и весьма перспективная часть прикладной статистики - применение высоких статистических технологий к анализу конкретных данных, что зачастую требует дополнительных теоретических исследований по доработке статистических технологий применительно к конкретной ситуации. Большое значение имеют конкретные статистические модели, например, модели экспертных оценок или эконометрики качества. И конечно, такие конкретные применения, как расчет и прогнозирование индекса инфляции. Сейчас уже многим экономистам и менеджерам ясно, что годовой бухгалтерский баланс предприятия может быть использован для оценки его финансово-хозяйственной деятельности только с привлечением данных об инфляции.

**Институт высоких статистических технологий и эконометрики.** Опишем опыт внедрения «высоких статистических технологий». Организованный нами в 1989 г. Институт высоких статистических технологий и эконометрики (ИВСТЭ) действует на базе кафедры ИБМ-2 “Экономика и организация производства” Московского государственного технического университета им. Н.Э.Баумана. Институт на хоздоговорных и госбюджетных началах занимается развитием, изучением и внедрением эконометрики и "высоких статистических технологий", т.е. наиболее современных технологий анализа экономических, технических, социологических, медицинских данных,

ориентированных на использование в условиях современного производства и экономики.
Основной интерес представляют применения "высоких статистических технологий" для анализа конкретных экономических данных, т.е. в эконометрике. Наиболее перспективным представляется применение "высоких статистических технологий" для поддержки принятия управленческих решений, прежде всего в таком новом (для России) современном направлении экономической науки и практики, как контроллинг (см., например, монографию [28]).

Вначале Институт действовал как Всесоюзный центр статистических методов и информатики Центрального правления Всесоюзного экономического общества. В 1990-1992 гг. было выполнено более 100 хоздоговорных работ, в том числе для НИЦентра по безопасности атомной энергетики, ВНИИ нефтепереработки, ПО “Пластик”, ЦНИИ черной металлургии им. Бардина, НИИ стали, ВНИИ эластомерных материалов и изделий, НИИ прикладной химии, ЦНИИ химии и механики, НПО “Орион”, ВНИИ экономических проблем развития науки и техники, ПО “Уралмаш”, “АвтоВАЗ”, МИИТ, Казахского политехнического института, Донецкого государственного университета, Института питания (Алма-Ата) и многих других.

Затем Институт разрабатывал эконометрические методы анализа нечисловых данных, а также процедуры расчета и прогнозирования индекса инфляции и валового внутреннего продукта. ИВСТЭ развивал методологию построения и использования математических моделей процессов налогообложения (для Министерства налогов и сборов РФ), методологию оценки рисков реализации инновационных проектов высшей школы (для Министерства промышленности, науки и технологий РФ). Институт оценивал влияние различных факторов на формирование налогооблагаемой базы ряда налогов (для Минфина РФ), прорабатывал перспективы применения современных статистических и экспертных методов для анализа данных о научном потенциале (для Министерства промышленности, науки и технологий РФ). Важное направление связано с эколого-экономической тематикой - разработка методологического, программного и информационного обеспечения анализа рисков химико-технологических объектов (для Международного научно-технического центра), методов использования экспертных оценок в задачах экологического страхования (совместно с Институтом проблем рынка РАН). Институт проводил маркетинговые исследования (в частности, для *Institute for Market Research GfK MR*, Промрадтехбанка, фирм, торгующих растворимым кофе, программным обеспечением, оказывающих образовательные услуги). Интерес вызывали работы Института по прогнозированию социально-экономического развития России методом сценариев, по экономико-математическому моделированию развития малых предприятий и созданию современных систем информационной поддержки принятия решений для таких организаций.

Институт ведет фундаментальные исследования в области высоких статистических технологий и эконометрики, в частности, в рамках МГТУ им. Н.Э. Баумана и Российского фонда фундаментальных исследований. Информация об Институте представлена на сайте в ИНТЕРНЕТе (http://antorlov.nm.ru, зеркала http://antorlov.euro.ru, http://www.newtech.ru/~orlov), который в 2000 г. посетили более 10000 пользователей. Институтом издается компьютерный еженедельник «Эконометрика» (около 1000 подписчиков). Архив выпусков газеты "Эконометрика" можно рассматривать как хрестоматию по различным разделам эконометрики, а также по высоким статистическим технологиям.

Термин "эконометрика" пока мало известен в России. А между тем в мировой науке эконометрика занимает достойное место. Напомним, что Нобелевские премии по экономике получили эконометрики Ян Тильберген, Рагнар Фриш, Лоуренс Клейн, Трюгве Хаавельмо. В 2000 г. к ним добавились еще двое - Джеймс Хекман и Даниель Мак-Фадден. Выпускается ряд научных журналов, полностью посвященных эконометрике, в том числе: *Journal of Econometrics* (Швеция), Econometric Reviews (США), *Econometrica* (США), Sankhya (*Indian Journal of Statistics. Ser.D. Quantitative Economics.* Индия), *Publications Econometriques* (Франция). Применение эконометрики дает заметный экономический эффект. Например, в США - не менее 20 миллиардов долларов ежегодно только в области статистического контроля качества.

Однако в нашей стране по ряду причин прикладная статистика и эконометрика не были сформированы как самостоятельные направления научной и практической деятельности, в отличие, например, от Польши, не говоря уже об англосаксонских странах. В результате специалистов в

области прикладной статистики и эконометрики у нас на порядок меньше, чем в США и Великобритании (Американская статистическая ассоциация включает более 20000 членов).

**О подготовке специалистов по высоким статистическим технологиям.** Приходится с сожалением констатировать, что в России плохо налажена подготовка специалистов по высоким статистическим технологиям. В курсах по теории вероятностей и математической статистике обычно даются лишь классические основы этих дисциплин, разработанные в первой половине XX в., а преподаватели-математики свою научную деятельность предпочитают посвящать доказательству теорем, имеющих лишь внутриматематическое значение, а не развитию высоких статистических технологий. В настоящее время появилась надежда на эконометрику. В России начинают развертываться эконометрические исследования и преподавание эконометрики. Экономисты, менеджеры и инженеры, прежде всего специалисты по контроллингу, должны быть вооружены современными средствами информационной поддержки, в том числе высокими статистическими технологиями и эконометрикой. Очевидно, преподавание должно идти впереди практического применения. Ведь как применять то, чего не знаешь?

Приведем два примера - отрицательный и положительный, - показывающие связь преподавания с внедрением передовых технологий.

Один раз - в 1990-1992 гг. мы уже обожглись на недооценке необходимости предварительной подготовки тех, для кого предназначены современные программные продукты. Наш коллектив (Всесоюзный центр статистических методов и информатики Центрального Правления Всесоюзного экономического общества) разработал систему диалоговых программных систем обеспечения качества продукции. Их созданием руководили ведущие специалисты страны. Но распространение программных продуктов шло на 1-2 порядка медленнее, чем мы ожидали. Причина стала ясна не сразу. Как оказалось, работники предприятий просто не понимали возможностей разработанных систем, не знали, какие задачи можно решать с их помощью, какой экономический эффект они дадут. А не понимали и не знали потому, что в вузах никто их не учил статистическим методам управления качеством. Без такого систематического обучения нельзя обойтись - сложные концепции "на пальцах" за пять минут не объяснишь.

Есть и противоположный пример - положительный. В середине 1980-х годов в советской средней школе ввели новый предмет "Информатика". И сейчас молодое поколение превосходно владеет компьютерами, мгновенно осваивая быстро появляющиеся новинки, и этим заметно отличается от тех, кому за 30-40 лет.

Если бы удалось ввести в средней школе курс теории вероятностей и статистики, то ситуация с внедрением высоких статистических технологий могла бы быть резко улучшена. Такой курс есть в Японии и США, Швейцарии, Кении и Ботсване, почти во всех странах (и ЮНЕСКО проводит всемирные конференции по преподаванию статистики в средней школе – см. сборник докладов [30]) Надо, конечно, добиться того, чтобы этот курс был построен на высоких статистических технологиях, а не на низких. Другими словами, он должен отражать современные достижения, а не концепции пятидесятилетней или столетней давности.

### 4.3. Компьютеры в прикладной статистике

**Методы статистических испытаний (Монте-Карло).** Многие информационные технологии в области прикладной статистики опираются на использование методов статистических испытаний. Этот термин применяется для обозначения компьютерных технологий, в которых в модель реального явления или процесса искусственно вводится большое число случайных элементов. Обычно моделируется последовательность независимых одинаково распределенных случайных величин или же последовательность, построенная на ее основе, например, последовательность накапливающихся (кумулятивных) сумм.

Необходимость в методе статистических испытаний возникает потому, что чисто теоретические методы дают точное решение, как правило, лишь в исключительных случаях. Либо тогда, когда исходные случайные величины имеют вполне определенные функции распределения,

например, нормальные, чего, как правило, не бывает. Либо когда объемы выборок очень велики (с практической точки зрения - бесконечны).

Не только в задачах обработки данных возникает необходимость в методе статистических испытаний. Она не менее актуальна и при экономико-математическом моделировании технических, социально-экономических, медицинских и иных процессов. Представим себе всем знакомый объект - торговый зал самообслуживания по продаже продовольственных товаров. Сколько нужно работников в зале, сколько касс? Необходимо просчитать загрузку в разное время суток, в разные сезоны года, с учетом замены товаров и смены сотрудников. Нетрудно увидеть, что теоретическому анализу подобная система не поддается, а компьютерному - вполне.

Методы статистических испытаний стали развиваться после второй мировой войны с появлением компьютеров. Второе название - методы Монте-Карло - они получили по наиболее известному игорному дому, а точнее, по его рулетке, поскольку исходный материал для получения случайных чисел с произвольным распределением - это случайные натуральные числа.

В методах статистических испытаний можно выделить две составляющие. Базой являются датчики псевдослучайных чисел. Результатом работы таких датчиков являются последовательности чисел, которые обладают некоторыми свойствами последовательностей случайных величин (в смысле теории вероятностей). Надстройкой являются различные алгоритмы, использующие последовательности псевдослучайных чисел.

Что же это могут быть за алгоритмы? Приведем примеры. Пусть мы изучаем распределение некоторой статистики при заданном объеме выборки. Тогда естественно много раз (например, 100000 раз) смоделировать выборку заданного объема (т.е. набор независимых одинаково распределенных случайных величин) и рассчитать значение статистики. Затем по 100000 значениям статистики можно достаточно точно построить функцию распределения изучаемой статистики, оценить ее характеристики. Однако эта схема годится лишь для так называемой "свободной от распределения" статистики, распределение которой не зависит от распределения элементов выборки. Если же такая зависимость есть, то одной точкой моделирования не обойдешься, придется много раз моделировать выборку, беря различные распределения, меняя параметры. Чтобы общее время моделирования было приемлемым, возможно, придется сократить число моделирований в одной точке, зато увеличив общее число точек. Точность моделирования может быть оценена по общим правилам выборочных обследований.

Второй пример - частично описанное выше моделирование работы торгового зала самообслуживания по продаже продовольственных товаров. Здесь одна последовательность псевдослучайных чисел описывает интервалы между появлениями покупателей, вторая, третья и т.д. связаны с выбором ими первого, второго и т.д. товаров в зале (например, число - номер в перечне товаров). Короче, все действия покупателей, продавцов, работников предприятия разбиты на операции, каждая операция, в продолжительности или иной характеристике которой имеется случайность, моделируется с помощью соответствующей последовательности псевдослучайных чисел. Затем итоги работы сотрудников торговой организации и зала в целом выражаются через характеристики случайных величин. Формулируется критерий оптимальности, решается задача оптимизации и находятся оптимальные значения параметров. В частности, оптимальные планы статистического контроля строятся на основе вероятностно-статистических моделей [7].

**Датчики псевдослучайных чисел.** Теперь обсудим свойства датчиков псевдослучайных чисел. Здесь стоит слово "псевдослучайные", а не "случайные". Это весьма важно. Дело в том, что за последние 50 лет обсуждались в основном три принципиально разных варианта получения последовательностей чисел, которые в дальнейшем использовались в методах статистических испытаний.

Первый - таблица случайных чисел. К сожалению, объем любой таблицы конечен, и сколько-нибудь сложные расчеты с ее помощью невозможны. Через некоторое время приходится повторяться. Кроме того, обычно обнаруживались те или иные отклонения от случайности.

Второй - физические датчики случайных чисел. Основной недостаток - нестабильность, непредсказуемые отклонения от заданного распределения (обычно - равномерного).

Третий - расчетный. В простейшем случае каждый следующий член последовательности рассчитывается по предыдущему. Например, так:

$$z_{n+1} \equiv Mz_n \,(\mathrm{mod}\, P),$$

где $z_0$ - начальное значение (заданное целое положительное число), $M$ - параметр алгоритма (заданное целое положительное число), $P=2^m$, где $m$ - число двоичных разрядов представления чисел, с которыми манипулирует компьютер. Знак $\equiv$ здесь означает теоретико-числовую операцию сравнения, т.е. взятие дробной части от $\frac{Mz_n}{P}$ и отбрасывание целой части.

В настоящее время применяется именно третий вариант. Совершенно ясно, что он не соответствует интуитивному представлению о случайности. Например, интуитивно очевидно, что по предыдущему элементу случайной последовательности с независимыми элементами нельзя предсказать значение следующего элемента. А приведенная выше формула как раз и дает способ такого предсказания. Расчетный путь получения последовательности псевдослучайных чисел противоречит не только интуиции, но и подходу к определению случайности на основе теории алгоритмов, развитому акад. А.Н. Колмогоровым и его учениками в 1960-х годах. Однако во многих прикладных задачах он работает, и это основное.

Методу статистических испытаний посвящена обширная литература (см., например, монографии [31-33]). Время от времени обнаруживаются недостатки у популярных датчиков псевдослучайных чисел. Так, например, в середине 1980-х годов выяснилось, что для одного из наиболее известных датчиков три последовательных значения связаны линейной зависимостью

$$Z_{n+2} = aZ_{n+1} + bZ_n, \quad n = 1,2,...$$

После этого в 1985 г. в журнале "Заводская лаборатория" началась дискуссия о качестве датчиков псевдослучайных чисел, которая продолжалась до 1993 г. и закончилась статьей проф. С.М.Ермакова [34] и нашим комментарием.

Итоги можно подвести так. Во многих случаях решаемая методом статистических испытаний задача сводится к оценке вероятности попадания в некоторую область в многомерном пространстве *фиксированной* размерности. Тогда из чисто математических соображений теории чисел следует, что с помощью датчиков псевдослучайных чисел поставленная задача решается корректно. Сводка соответствующих математических обоснований приведена, например, в работе С.М. Ермакова [34].

В других случаях приходится рассматривать вероятности попадания в области в пространствах *переменной* размерности. Типичным примером является ситуация, когда на каждом шагу проводится проверка, и по ее результатам либо остаемся в данном пространстве, либо переходим в пространство большей размерности. Например, в главе 3.2 при оценивании степени многочлена либо останавливались на данной степени, либо увеличивали степень, переходя в параметрическое пространство большей размерности. Так вот, вопрос об обоснованности применения метода статистических испытаний (а точнее, о свойствах датчиков псевдослучайных чисел) в случае пространств переменной размерности остается в настоящее время открытым. О важности этой проблемы говорил академик РАН Ю.В. Прохоров на Первом Всемирном Конгрессе Общества математической статистики и теории вероятностей им. Бернулли (Ташкент, 1986 г.).

**Имитационное моделирование.** Поскольку постоянно говорим о моделировании, приведем несколько общих формулировок.

Модель в общем смысле (обобщенная модель) - это создаваемый с целью получения и (или) хранения информации специфический объект (в форме мысленного образа, описания знаковыми средствами либо материальной системы), отражающей свойства, характеристики и связи объекта-оригинала произвольной природы, существенные для задачи, решаемой субъектом (это определение взято из монографии [35, с.44]).

Например, в менеджменте производственных систем используют:

- модели технологических процессов (контроль и управление по технико-экономическим критериям, АСУ ТП - автоматизированные системы управления технологическими процессами);

- модели управления качеством продукции (в частности, модели оценки и контроля надежности);

- модели массового обслуживания (теории очередей);

- модели управления запасами (в современной терминологии - модели логистики, т.е. теории и практики управления материальными, финансовыми и информационными потоками);

- имитационные и эконометрические модели деятельности предприятия (как единого целого) и управления им (АСУ предприятием) и др.

Согласно академику РАН Н.Н. Моисееву [36, с.213], имитационная система - это совокупность моделей, имитирующих протекание изучаемого процесса, объединенная со специальной системой вспомогательных программ и информационной базой, позволяющих достаточно просто и оперативно реализовать вариантные расчеты. Другими словами, имитационная система - это совокупность имитационных моделей. А имитационная модель предназначена для ответов на вопросы типа: "Что будет, если…" Что будет, если параметры примут те или иные значения? Что будет с ценой на продукцию, если спрос будет падать, а число конкурентов расти? Что будет, если государство резко усилит вмешательство в экономику? Что будет, если остановку общественного транспорта перенесут на 100 м дальше от входа в торговый зал, о котором шла речь выше, и поток покупателей резко упадет? Кроме компьютерных моделей, на вопросы подобного типа часто отвечают эксперты при использовании метода сценариев [7, 8].

При имитационном моделировании часто используется метод статистических испытаний (Монте-Карло). Теорию и практику машинных имитационных экспериментов с моделями экономических систем еще 30 лет назад подробно разобрал Т. Нейлор в обширной классической монографии [37]. Вернемся к внутристатистическому применению датчиков псевдослучайных чисел.

**Методы размножения выборок (бутстреп-методы).** Прикладная статистика бурно развивается последние десятилетия. Серьезным (хотя, разумеется, не единственным и не главным) стимулом является стремительно растущая производительность вычислительных средств. Поэтому понятен острый интерес к статистическим методам, интенсивно использующим компьютеры. Одним из таких методов является так называемый "бутстреп", предложенный в 1977 г. Б.Эфроном из Станфордского университета (США).

Сам термин "бутстреп" - это английское слово "*bootstrap*", записанное русскими буквами. Оно буквально означает что-то вроде: "вытягивание себя (из болота) за шнурки от ботинок". Термин специально придуман и заставляет вспомнить о подвигах барона Мюнхгаузена.

В истории прикладной статистики было несколько более или менее успешно осуществленных рекламных кампаний. В каждой из них "раскручивался" тот или иной метод, который, как правило, отвечал нескольким условиям:

- по мнению его пропагандистов, полностью решал актуальную научную задачу;

- был понятен (при постановке задачи, при ее решении и при интерпретации результатов) широким массам потенциальных пользователей;

- использовал современные возможности вычислительной техники.

Пропагандисты метода, как правило, избегали беспристрастного сравнения его возможностей с возможностями иных эконометрических методов. Если сравнения и проводились, то с заведомо слабым "противником".

В нашей стране в условиях отсутствия систематического образования в области прикладной статистики подобные рекламные кампании находили особо благоприятную почву, поскольку у большинства затронутых ими специалистов не было достаточных знаний в области методологии построения моделей прикладной статистики для того, чтобы составить самостоятельное квалифицированное мнение.

Речь идет о таких методах и постановках, как бутстреп, нейронные сети, метод группового учета аргументов, робастные оценки по Тьюки-Хуберу, асимптотика пропорционального роста числа параметров и объема данных и др. Бывают локальные всплески энтузиазма, например, московские социологи в 1980-х годах пропагандировали так называемый "детерминационный анализ" - простой эвристический метод анализа таблиц сопряженности. Хотя в Новосибирске в это время давно уже было разработано продвинутое математическое и программное обеспечение анализа векторов

разнотипных признаков, включающее в себя «детерминационный анализ» как весьма частный случай.

Однако даже на фоне всех остальных рекламных кампаний судьба бутстрепа исключительна. Во-первых, признанный его автор Б. Эфрон с самого начала признавался, что он ничего принципиально нового не сделал. Его исходная статья (первая в сборнике [5]) называлась: "Бутстреп-методы: новый взгляд на методы складного ножа". Тем самым Б.Эфрон честно признавал первенство за М. Кенуем – автором методов «складного ножа». Во вторых, сразу появились статьи и дискуссии в научных изданиях, публикации рекламного характера, и даже в научно-популярных журналах. Бурные обсуждения на конференциях, спешный выпуск книг. В 1980-е годы финансовая подоплека всей этой активности, связанная с выбиванием грантов на научную деятельность, содержание учебных заведений и т.п. была мало понятна отечественным специалистам.

В чем основная идея группы методов "размножения выборок", наиболее известным представителем которых является бутстреп?

Пусть дана выборка $x_1, x_2, x_3, ... x_{k-1}, x_k, x_{k+1}, ..., x_{n-1}, x_n$. . В вероятностно-статистической теории предполагаем, что это - набор независимых одинаково распределенных случайных величин. Пусть эконометрика интересует некоторая статистика $f_n(x_1, x_2, ..., x_n)$. Как изучить ее свойства? Подобными проблемами мы занимались на протяжении всей книги и знаем, насколько это непросто. Идея, которую предложил в 1949 г. М. Кенуй (это и есть "метод складного ножа") состоит в том, чтобы из одной выборки сделать много, исключая из нее по одному наблюдению (и возвращая ранее исключенные). Перечислим выборки, которые получаются из исходной:

$$x_2, x_3, ...x_{k-1}, x_k, x_{k+1}, ..., x_{n-1}, x_n.;$$
$$x_1, x_3, ...x_{k-1}, x_k, x_{k+1}, ..., x_{n-1}, x_n.;$$
$$x_1, x_2, x_4, ...x_{k-1}, x_k, x_{k+1}, ..., x_{n-1}, x_n.;$$
$$\dots$$
$$x_1, x_2, x_3, ...x_{k-1}, x_{k+1}, ..., x_{n-1}, x_n.;$$
$$\dots$$
$$x_1, x_2, x_3, ...x_{k-1}, x_k, x_{k+1}, ..., x_{n-2}, x_n.;$$
$$x_1, x_2, x_3, ...x_{k-1}, x_k, x_{k+1}, ..., x_{n-1}.$$

Всего *n* новых (размноженных) выборок объемом (*n*-1) каждая. По каждой из них можно рассчитать значение интересующей эконометрика статистики (с уменьшенным на 1 объемом выборки):

$$f_{n-1,1}(\omega) = f_{n-1}(x_2, x_3, ...x_{k-1}, x_k, x_{k+1}, ..., x_{n-1}, x_n);$$
$$f_{n-1,2}(\omega) = f_{n-1}(x_1, x_3, ...x_{k-1}, x_k, x_{k+1}, ..., x_{n-1}, x_n);$$
$$f_{n-1,3}(\omega) = f_{n-1}(x_1, x_2, x_4, ...x_{k-1}, x_k, x_{k+1}, ..., x_{n-1}, x_n);$$
$$\dots$$
$$f_{n-1,k}(\omega) = f_{n-1}(x_1, x_2, x_3, ...x_{k-1}, x_{k+1}, ..., x_{n-1}, x_n);$$
$$\dots$$
$$f_{n-1,n-1}(\omega) = f_{n-1}(x_1, x_2, x_3, ...x_{k-1}, x_k, x_{k+1}, ..., x_{n-2}, x_n);$$
$$f_{n-1,n}(\omega) = f_{n-1}(x_1, x_2, x_3, ...x_{k-1}, x_k, x_{k+1}, ..., x_{n-1}).$$

Полученные значения статистики позволяют судить о ее распределении и о характеристиках распределения - о математическом ожидании, медиане, квантилях, разбросе, среднем квадратическом отклонении. Значения статистики, построенные по размноженным подвыборкам, не являются независимыми. Однако, как мы видели в главе 3.2 на примере ряда статистик, возникающих в методе наименьших квадратов и в кластер-анализе (при обсуждении возможности объединения двух кластеров), при росте объема выборки влияние зависимости может ослабевать, а потому со значениями статистик типа $f_{n-1,k}(\omega), \quad k = 1,2,...,n,$ можно обращаться как с независимыми случайными величинами.

Однако и без всякой вероятностно-статистической теории разброс величин $f_{n-1,k}(\omega), \quad k = 1,2,...,n,$ дает наглядное представление о том, какую точность может дать рассматриваемая статистическая оценка.

Сам М. Кенуй и его последователи использовали размножение выборок в основном для построения оценок с уменьшенным смещением. А вот Б. Эфрон предложил новый способ размножения выборок, существенно использующий датчики псевдослучайных чисел. А именно, он предложил строить новые выборки, *моделируя выборки из эмпирического распределения.* Другими словами, Б. Эфрон предложил взять конечную совокупность из *n* элементов исходной выборки $x_1, x_2, x_3, ... x_{k-1}, x_k, x_{k+1}, ..., x_{n-1}, x_n$. и с помощью датчика псевдослучайных чисел сформировать из нее любое число размноженных выборок. Процедура, хотя и нереальна без ЭВМ, проста с точки зрения программирования. По сравнению с описанной выше процедурой Кенуя появляются новые недостатки - неизбежные совпадения элементов размноженных выборок и зависимость от качества датчиков псевдослучайных чисел. Однако существует математическая теория, позволяющая (при некоторых предположениях и безграничном росте объема выборки) обосновать процедуры бутстрепа (см. сборник статей [5]).

Есть много способов развития идеи размножения выборок (см., например, статью [38]). Можно по исходной выборке построить эмпирическую функцию распределения, а затем каким-либо образом от кусочно-постоянной функции перейти к непрерывной функции распределения, например, соединив точки $(x(i); \frac{i}{n}), i = 1,2,...n,$ отрезками прямых. Другой вариант - перейти к непрерывному распределению, построив непараметрическую оценку плотности. После этого рекомендуется брать размноженные выборки из этого непрерывного распределения (являющегося состоятельной оценкой исходного), непрерывность защитит от совпадений элементов в этих выборках.

Другой вариант построения размноженных выборок - более прямой. Исходные данные не могут быть определены совершенно точно и однозначно. Поэтому предлагается к исходным данным добавлять малые независимые одинаково распределенные погрешности. При таком подходе соединяем вместе идеи устойчивости и бутстрепа. При внимательном анализе многие идеи прикладной статистики тесно друг с другом связаны (см. статью [38]).

В каких случаях целесообразно применять бутстреп, а в каких - другие методы прикладной статистики? В период рекламной кампании встречались, в том числе в научно-популярных журналах, утверждения о том, что и для оценивания математического ожидания полезен бутстреп. Как показано в статье [38], это совершенно не так. При росте числа испытаний методом Монте-Карло бутстреп-оценка приближается к классической оценке - среднему арифметическому результатов наблюдений. Другими словами, бутстреп-оценка отличается от классической оценки только шумом псевдослучайных чисел.

Аналогичной является ситуация и в ряде других случаев. Там, где эконометрическая теория хорошо развита, где найдены методы анализа данных, в том или иной смысле близкие к оптимальным, бутстрепу делать нечего. А вот в новых областях со сложными алгоритмами, свойства которых недостаточно ясны, он представляет собой ценный инструмент для изучения ситуации.

**Компьютерная статистика в контроллинге.** В качестве примера применения компьютерной статистики рассмотрим конкретную прикладную область – контроллинг, т.е. современный подход к управлению организацией [28]. Контроллеру и сотрудничающему с ним статистику нужна разнообразная экономическая и управленческая информация, не менее нужны удобные инструменты ее анализа. Следовательно, информационная поддержка контроллинга необходима для успешной работы контроллера. Без современных компьютерных инструментов анализа и управления, основанных на продвинутых эконометрических и экономико-математических методах и моделях, невозможно эффективно принимать управленческие решения. Недаром специалисты по контроллингу большое внимание уделяют проблемам создания, развития и применения компьютерных систем поддержки принятия решений. Высокие статистические технологии и

эконометрика - неотъемлемые части любой современной системы поддержки принятия экономических и управленческих решений.

Важная часть прикладной статистики - применение высоких статистических технологий к анализу конкретных экономических данных. Такие исследования зачастую требуют дополнительной теоретической работы по "доводке" статистических технологий применительно к конкретной ситуации. Большое значение для контроллинга имеют не только общие методы, но и конкретные эконометрические модели, например, вероятностно-статистические модели тех или иных процедур экспертных оценок или эконометрики качества, имитационные модели деятельности организации, прогнозирования в условиях риска. И конечно, такие конкретные применения, как расчет и прогнозирование индекса инфляции. Сейчас уже многим специалистам ясно, что годовой, квартальный или месячный бухгалтерский баланс предприятия может быть использован для оценки его финансово-хозяйственной деятельности только с привлечением данных об инфляции. Различные области экономической теории и практики в настоящее время еще далеко не согласованы. При оценке и сравнении инвестиционных проектов принято использовать такие характеристики, как чистая текущая стоимость, внутренняя норма доходности, основанные на введении в рассмотрение изменения стоимости денежной единицы во времени (это осуществляется с помощью дисконтирования). А вот при анализе финансово-хозяйственной деятельности организации на основе данных бухгалтерской отчетности изменение стоимости денежной единицы во времени по традиции не учитывают.

Специалисты по контроллингу должны быть вооружены современными средствами информационной поддержки, в том числе средствами на основе высоких статистических технологий и эконометрики. Очевидно, преподавание должно идти впереди практического применения. Ведь как применять то, чего не знаешь?

Статистические технологии применяют для анализа данных двух принципиально различных типов. Один из них - это результаты измерений (наблюдений, испытаний, анализов, опытов и др.) различных видов, например, результаты управленческого или бухгалтерского учета, данные Госкомстата и др. Короче, речь идет об объективной информации. Другой - это оценки экспертов, на основе своего опыта и интуиции делающих заключения относительно экономических явлений и процессов. Очевидно, это - субъективная информация. В стабильной экономической ситуации, позволяющей рассматривать длинные временные ряды тех или иных экономических величин, полученных в сопоставимых условиях, данные первого типа вполне адекватны. В быстро меняющихся условиях приходятся опираться на экспертные оценки. Такая новейшая часть прикладной статистики, как статистика нечисловых данных, была создана как ответ на запросы теории и практики экспертных оценок.

Для решения каких экономических задач может быть полезна прикладная статистика? Практически для всех, использующих конкретную информацию о реальном мире. Только чисто абстрактные, отвлеченные от реальности исследования могут обойтись без нее. В частности, прикладная статистика необходима для прогнозирования, в том числе поведения потребителей, а потому и для планирования. Выборочные исследования, в том числе выборочный контроль, основаны на прикладной статистике. Но планирование и контроль - основа контроллинга. Поэтому прикладная статистика - важная составляющая инструментария контроллера, воплощенного в компьютерной системе поддержки принятия решений. Прежде всего оптимальных решений, которые предполагают опору на адекватные модели прикладной статистики. В производственном менеджменте это может означать, например, использование моделей экстремального планирования эксперимента (судя по накопленному опыту их практического использования, такие модели позволяют повысить выход полезного продукта на 30-300%).

Высокие статистические технологии предполагают адаптацию применяемых методов к меняющейся ситуации. Например, параметры прогностического индекса меняются вслед за изменением характеристик используемых для прогнозирования величин. Таков метод экспоненциального сглаживания. В соответствующем алгоритме расчетов значения временного ряда используются с весами. Веса уменьшаются по мере удаления в прошлое. Многие методы дискриминантного анализа основаны на применении обучающих выборок. Например, для построения

рейтинга надежности банков можно с помощью экспертов составить две обучающие выборки - надежных и ненадежных банков. А затем с их помощью решать для вновь рассматриваемого банка, каков он - надежный или ненадежный, а также оценивать его надежность численно, т.е. вычислять значение рейтинга.

Один из способов построения адаптивных статистических моделей - нейронные сети (см., например, монографию [39]). При использовании нейронных сетей упор делается не на формулировку адаптивных алгоритмов анализа данных, а - в большинстве случаев - на построение виртуальной адаптивной структуры. Термин "виртуальная" означает, что "нейронная сеть" - это специализированная компьютерная программа, "нейроны" используются лишь при общении человека с компьютером. Методология нейронных сетей идет от идей кибернетики 1940-50-х годов. В компьютере создается модель мозга человека (весьма примитивная с точки зрения физиолога). Основа модели - весьма простые базовые элементы, называемые нейронами. Они соединены между собой, так что нейронные сети можно сравнить с хорошо знакомыми экономистам и инженерам блок-схемами. Каждый нейрон находится в одном из заданного множества состояний. Он получает импульсы от соседей по сети, изменяет свое состояние и сам рассылает импульсы. В результате состояние множества нейтронов изменяется, что соответствует проведению статистических вычислений.

Нейроны обычно объединяются в слои (как правило, два-три). Среди них выделяются входной и выходной слои. Перед началом решения той или иной задачи производится настройка. Во-первых, устанавливаются связи между нейронами, соответствующие решаемой задаче. Во-вторых, проводится обучение, т.е. через нейронную сеть пропускаются обучающие выборки, для элементов которых требуемые результаты расчетов известны. Затем параметры сети модифицируются так, чтобы получить максимальное соответствие выходных значений заданным величинам.

С точки зрения точности расчетов (и оптимальности в том или ином статистическом смысле) нейронные сети не имеют преимуществ перед другими адаптивными системами прикладной статистики. Однако они более просты для восприятия. Надо отметить, что в прикладной статистике используются и модели, промежуточные между нейронными сетями и "обычными" системами регрессионных уравнений (одновременных и с лагами). Они тоже используют блок-схемы, как, например, универсальный метод моделирования связей социально-экономических факторов ЖОК (этот метод описан в [7]).

Профессионалу в области контроллинга полезны многочисленные интеллектуальные инструменты анализа данных, относящиеся к высоким статистическим технологиям и эконометрике. В частности, заметное место в математико-компьютерном обеспечении принятия решений в контроллинге занимают методы теории нечеткости.

### 4.4. Основные нерешенные проблемы прикладной статистики

За последние тридцать лет выявился целый ряд нерешенных проблем прикладной статистики, как чисто научных, так и научно-организационных. Обсудим пять из них:

влияние отклонений от традиционных предпосылок вероятностно-статистических моделей на свойства статистических процедур;

оправданность использования асимптотических теоретических результатов прикладной математической статистики при конечных объемах выборок;

формулировки и обоснования правил выбора одного из многих критериев для проверки конкретных гипотез;

конкретные способы организации теоретических работ в области прикладной статистики;

организация и проведение прикладных работ с использованием статистических методов.

Приводимые ниже соображения отнюдь не претендует на решение перечисленных проблем. Их цель гораздо скромнее - обратить внимание на существование ряда нерешенных проблем в надежде, что коллективными усилиями удастся продвинуться в их решении.

**Влияние отклонений от традиционных предпосылок.** В вероятностной теории статистических методов выборка обычно моделируется как конечная последовательность

независимых одинаково распределенных случайных величин или векторов. Часто предполагается, что эти величины или вектора имеют нормальное распределение.

На основе сформулированных классических предпосылок построено огромное здание классической математической статистики с большим числом теорем. Оно за последнее столетие обросло горой учебников и программных продуктов.

Однако при внимательном взгляде совершенно ясна нереалистичность классических предпосылок. Независимость результатов измерений обычно принимается "из общих предположений", между тем во многих случаях очевидна их коррелированность [38]. Одинаковая распределенность результатов измерений также вызывает сомнения из-за изменения во времени свойств измеряемых образцов, средств измерения и психофизического состояния специалистов, проводящих измерения (наблюдения, испытания, анализы, опыты). Даже обоснованность самой возможности применения вероятностных моделей также часто вызывает сомнения, например, при моделировании уникальных измерений (теорию вероятностей обычно привлекают при изучении массовых явлений). И уж совсем редко распределения результатов измерений можно считать нормальными (см. главу 3.1).

Итак, методы классической математической статистики обычно используют вне сферы их обоснованной применимости. Каково влияние отклонений от традиционных предпосылок на статистические выводы? В настоящее время об этом имеются лишь отрывочные сведения. Приведем три примера.

*Пример 1.* Построение доверительного интервала для математического ожидания обычно проводят с использованием распределения Стьюдента (при справедливости гипотезы нормальности). Как следует из Центральной Предельной Теоремы (ЦПТ) теории вероятностей, в асимптотике (т.е. при большом объеме выборки) такие расчетные методы дают правильные результаты. А именно, из ЦПТ вытекает использование квантилей нормального распределения, а из классической теории - квантилей распределения Стьюдента, но при росте объема выборки квантили распределения Стьюдента стремятся к соответствующим квантилям нормального распределения.

*Пример 2.* Для проверки однородности двух независимых выборок (на самом деле - для проверки равенства математических ожиданий) обычно рекомендуют использовать двухвыборочный критерий Стьюдента. Что будет при отклонении от нормальности распределений, из которых взяты выборки? Если объемы выборок равны или если дисперсии результатов наблюдений в выборках совпадают, то в асимптотике (когда объемы выборок безгранично возрастают) классический метод является корректным. Если же объемы выборок существенно отличаются и их дисперсии различны, то двухвыборочную статистику Стьюдента применять нельзя. Поскольку проверка равенства дисперсий - более сложная задача, чем проверка равенства математических ожиданий, то для выборок разного объема использовать двухвыборочную статистику Стьюдента не следует, лучше применять критерий Крамера-Уэлча, как это подробно обосновано в главе 3.1.

*Пример 3.* В задаче отбраковки (исключения) резко выделяющихся наблюдений (выбросов) расчетные методы, основанные на нормальности, являются крайне неустойчивыми по отношению к отклонениям от нормальности, что полностью лишает эти методы научной обоснованности (подробнее см. главу 2.3).

Примеры 1-3 показывают весь спектр возможных свойств классических расчетных методов в случае отклонения от нормальности. Методы примера 1 оказываются вполне пригодными при таких отклонениях, примера 2 - пригодными в некоторых случаях, примера 3 - полностью непригодными.

Итак, имеется *необходимость изучения свойств расчетных методов классической математической статистики, опирающихся на предположение нормальности, в ситуациях, когда это предположение не выполнено.* Аппаратом для такого изучения наряду с методом Монте-Карло (статистических испытаний) могут послужить предельные теоремы теории вероятностей (и опирающиеся на них асимптотические методы математической статистики), прежде всего ЦПТ, поскольку интересующие нас расчетные методы обычно используют разнообразные суммы.

Пока подобное изучение не проведено, остается неясной научная ценность, например, применения факторного анализа к векторам из переменных, принимающих небольшое число градаций и к тому же измеренных в порядковой шкале. Этот пример показывает важность еще

одного направления исследований - изучения свойств алгоритмов, предназначенных для анализа числовых данных, в случаях, когда данные измерены в шкалах, отличных от абсолютной, в частности, в порядковой шкале.

Из большого числа возможных постановок, относящихся к изучению влияния отклонений от традиционных предпосылок, укажем лишь на то, что реальные данные имеют небольшое число значащих цифр (обычно от 2 до 5), в то время как в классической математической статистике используются непрерывные случайные величины, для которых вероятность получения подобного результата наблюдения равна 0. Действительно, вероятность того, что хотя бы один элемент выборки из распределения с непрерывной функцией распределение попадет в заданное счетное множество, в частности, в множество рациональных чисел, равна 0 (согласно классическим свойствам вероятностной меры). Событиями, имеющими вероятность 0, принято пренебрегать. Следовательно, с точки зрения классической математической статистики любыми реальными данными нужно пренебречь! Выходов из этого парадокса несколько. Один из них - бурно развивающаяся в настоящее время статистика интервальных данных (см. главу 3.5), другой - использование классических поправок Шеппарда для сгруппированных данных [39, 40]. Здесь еще много работы. Так, даже для такого широко используемого статистического показателя, как коэффициент корреляции, поправки на группировку (поправки Шеппарда) были получены сравнительно недавно - лишь в 1980 г. [41].

Почему на первый план выдвинуто изучение классических алгоритмов, а не построение новых, специально предназначенных для работы в условиях отклонения от классических предпосылок? Во-первых, потому, что классические алгоритмы в настоящее время наиболее распространены (благодаря сложившейся системе образования как прикладников, так и математиков). Во-вторых, более новые подходы зачастую методологически уязвимы. Так, известная робастная модель засорения Тьюки-Хубера (см. главу 2.2) нацелена на борьбу с большими выбросами, которые зачастую физически невозможны из-за ограниченности интервала возможных значений измеряемой характеристики, в котором работает конкретное средство измерения. Следовательно, модель Тьюки-Хубера имеет скорее теоретическое значение, чем практическое. Сказанное, конечно, не означает, что следует прекратить разработку, изучение и внедрение непараметрических и устойчивых методов, выделенных выше как "точки роста" современных эконометрики и прикладной статистики.

**Использование асимптотических результатов при конечных объемах выборок.** Как отмечено выше, изучение классических алгоритмов во многих случаях может быть проведено с помощью асимптотических методов математической статистики, в частности, с помощью ЦПТ и методов наследования сходимости (см. главу 1.4). Отрыв классической математической статистики от нужд прикладных исследований проявился, в частности, в том, что в распространенных монографиях недостает математического аппарата, необходимого, в частности, для изучения двухвыборочных статистик. Суть в том, что переходить к пределу приходится не по одному параметру, а по двум – объемам двух выборок. Пришлось разработать соответствующую теорию – теорию наследования сходимости, впервые изложенную в монографии [3, п.2.4].

Однако применять результаты подобного изучения придется при конечных объемах выборок. Возникает целый букет проблем, связанных с таким переходом. Часть из них обсуждалась в главе 1.4.7 в связи с изучением свойств статистик, построенных по выборкам из конкретных распределений.

Однако при обсуждении влияния отклонений от исходных предположений на свойства статистических процедур возникают дополнительные проблемы. Какие отклонения считать типичными? Ориентироваться ли на наиболее "вредные" отклонения, в наибольшей степени искажающие свойства алгоритмов, или же сосредоточить внимание на "типичных" отклонениях?

При первом подходе получаем гарантированный результат, но "цена" этого результата может быть излишне высокой. В качестве примера укажем на универсальное неравенство Берри-Эссеена для погрешности в ЦПТ [42, 43]. Совершенно справедливо подчеркивает академик РАН А.А. Боровков [43, с,172], что «скорость сходимости в реальных задачах, как правило, оказывается лучше».

При втором подходе возникает вопрос, какие отклонения считать "типичными". Попытаться ответить на этот вопрос можно, анализируя большие массивы реальных данных. Вполне естественно, что ответы различных исследовательских групп будут различаться.

Одна из ложных идей - использование при анализе возможных отклонений только какого-либо конкретного параметрического семейства. Например, семейств распределений Вейбулла-Гнеденко, экспоненциальных, нормальных, трехпараметрического семейства гамма - распределений и др. Как уже отмечалось во введении к настоящему учебнику, еще в 1927 г. акад. АН СССР С.Н. Бернштейн обсуждал методологическую ошибку, состоящую в сведении всех эмпирических распределений к четырехпараметрическому семейству Пирсона [44]. Однако и до сих пор параметрические методы статистики весьма популярны, особенно среди прикладников, и вина за это заблуждение лежит, прежде всего, на преподавателях статистических методов.

**Выбор одного из многих критериев для проверки конкретной гипотезы.** Во многих случаях для решения конкретной практической задачи разработано много методов, и специалист по прикладной статистике стоит перед проблемой: какой из них предложить прикладнику для анализа конкретных данных?

В качестве примера рассмотрим задачу проверки однородности двух независимых выборок. Как известно (см. главу 3.1), для ее решения можно предложить массу критериев: Стьюдента, Крамера-Уэлча, Лорда, хи-квадрат, Вилкоксона (Манна-Уитни), Ван–дер- Вардена, Сэвиджа, Н.В.Смирнова, типа омега-квадрат (Лемана-Розенблатта), Г.В. Мартынова и др. Какой выбрать?

Естественным образом приходит в голову идея "голосования": провести проверку по многим критериям, а затем принять решение "по большинству голосов". С точки зрения статистической теории такая процедура приводит попросту к построению еще одного критерия, который априори ничем не лучше прежних (но и не хуже), но более труден для изучения. С другой стороны, если совпадают решения по всем рассмотренным статистическим критериям, исходящим из различных принципов, то в соответствии с концепцией устойчивости, впервые развитой в монографии [3] (см. также главу 1.4), это повышает доверие к полученному общему решению.

Распространено, особенно среди математиков, ложное и вредное мнение о необходимости поиска оптимальных методов, решений и т.д. Дело в том, что оптимальность обычно исчезает при отклонении от исходных предпосылок. Так, среднее арифметическое в качестве оценки математического ожидания является оптимальной оценкой тогда и только тогда, когда исходное распределение - нормальное (см., например, монографию [45]), в то время как состоятельной оценкой - всегда, лишь бы математическое ожидание существовало. С другой стороны, для любого произвольно взятого метода оценивания или проверки гипотез обычно можно так сформулировать понятие оптимальности, чтобы рассматриваемый метод стал оптимальным – с этой специально выбранной точки зрения. Возьмем, например, выборочную медиану как оценку математического ожидания. Она, разумеется, оптимальна, хотя и в другом смысле, чем среднее арифметическое (оптимальное для нормального распределения). А именно, для распределения Лапласа выборочная медиана является оценкой максимального правдоподобия, а потому оптимальной - в том смысле, в каком оптимальной является любая оценка максимального правдоподобия. Соответствующее понятие оптимальности требует аккуратных формулировок, оно строго изложено в монографии [46]. Как известно, оценки максимального правдоподобия удобны при теоретических рассмотрениях, а при анализе конкретных экономических, технических и иных данных следует применять одношаговые оценки (см. об этом главу 2.2).

Проиллюстрируем сказанное примером. Критерии однородности двух выборок были проанализированы в монографии [47]. Естественных подходов к сравнению критериев несколько - на основе асимптотической относительной эффективности по Бахадуру, Ходжесу-Леману, Питмену и др. И выяснилось, что каждый обычно используемый критерий однородности является оптимальным при соответствующей альтернативе или подходящем распределении на множестве альтернатив. При этом математические рассуждения обычно опираются на альтернативу сдвига, сравнительно редко встречающуюся в практике анализа реальных статистических данных (в связи с критерием Вилкоксона эта альтернатива обсуждалась в главе 3.1). Итог печален - блестящая математическая техника, продемонстрированная в монографии [47], не позволяет дать рекомендации для выбора

критерия проверки однородности при анализе реальных данных. Другими словами, с точки зрения работы прикладника, т.е. с точки зрения применимости полученных результатов при анализе конкретных данных, монография [47] бесполезна. Блестящее владение математикой и огромное трудолюбие, продемонстрированные автором этой монографии, увы, ничего не принесли практике.

Конечно, каждый практически работающий статистик так или иначе решает для себя проблему выбора статистического критерия. На основе ряда методологических соображений в главе 3.1 мы остановили свой выбор на состоятельном против любой альтернативы критерии типа омега-квадрат (Лемана-Розенблатта). Однако остается чувство неудовлетворенности в связи с недостаточной теоретической обоснованностью этого выбора.

**Организация теоретических работ в области прикладной статистики.** Выше продемонстрирована необходимость большой теоретической работы по развитию нацеленных на практическое использование методов прикладной статистики. В статье [48] 1992 г. обоснован вывод о необходимости создания сети научно-исследовательских организаций, которая выполняла бы такую работу. Как известно, количество научных работников к настоящему времени сократилось в несколько раза по сравнению с началом 1990-х годов, так что на осуществление в ближайшие годы сформулированной в [48] научно-организационной программы надеяться не приходится.

Приходится с сожалением констатировать, что в рамках научной специальности "теория вероятностей и математическая статистика" наблюдается четко выраженное игнорирование проблем статистического анализа реальных данных и уход в глубь узкоматематических исследований, которые заведомо ничего не могут дать практике. Причины этого явления, типичного для математических дисциплин, обсуждались во введении к настоящему учебнику. Поэтому нет оснований ожидать, что при "естественном ходе событий" будут получены существенные продвижения в рассмотренных выше нерешенных проблемах прикладной статистики.

Помочь может выделение государственными структурами системы грантов, направленных на поддержку работ в области нерешенных проблем прикладной статистики. Принципиальным шагом явилось бы официальное выделение государственными органами прикладной статистики как самостоятельного научного направления. Отличного как от чисто математических дисциплин типа "теории вероятностей и математической статистики", так и от, например, ветви экономической теории, известной в официальных кругах под названием "статистика" (см. приложение 2 ниже).

**О прикладных работах с использованием методов прикладной статистики.** Проблемы организации теоретических работ в области прикладной статистики лишь в перспективе важны для практической работы. Как правило, те, кто обрабатывает реальные данные, недостаточно знакомы с теоретическими основами алгоритмов и тем более не следят за событиями "на переднем крае" обсуждаемой научно-практической дисциплины. Это вполне естественно, поскольку основная специальность у таких специалистов - иная.

Несколько огрубляя, можно сказать, что реально используется только то, что имеется в учебниках и справочниках, в широко распространенных программных продуктах, а научные публикации с точки зрения прикладника представляют собой "информационный шум". Ситуация усугубляется традиционным ненормальным положением в отечественной статистике [49].

К сожалению, учебная и научная литература на русском языке (как, впрочем, и на иных языках) по прикладной статистике в целом далека от совершенства, переполнена устаревшими методологическими подходами и прямыми ошибками. До сих пор наилучшим изданием остаются "Таблицы математической статистики" Л.Н. Большева и Н.В.Смирнова [1], созданные еще в 1960-х годах.

Хотя студенты почти всех специальностей изучают в конце курса высшей математики раздел "теория вероятностей и математическая статистика", реально они знакомятся лишь с некоторыми основными понятиями и результатами, которых явно не достаточно для практической работы. С некоторыми математическими методами исследования студенты встречаются в специальных курсах (например, таких, как "Прогнозирование и технико-экономическое планирование", "Технико-экономический анализ", "Контроль качества продукции", "Маркетинг", "Контроллинг", "Математические методы прогнозирования", «Статистика» и др. – в случае студентов экономических специальностей), однако изложение в большинстве случаев носит весьма сокращенный и

рецептурный характер. В результате подавляющую часть специалистов по прикладной статистике следует считать самоучками.

Поэтому большое значение имеет введение в технических вузах курса "Прикладная статистика", а на экономических факультетах таких вузов и в экономических вузах – курса «Эконометрика», поскольку эконометрика – это, как известно, статистический анализ конкретных экономических данных (см. [7]). Это естественно делать, например, в рамках подпрограммы "Технологии подготовки кадров для национальной технологической базы" федеральной целевой программы "Национальная технологическая база". Естественно, что курсы "Прикладная статистика" и «Эконометрика» должны быть обеспечены соответствующими учебниками и учебными пособиями, методическими материалами и обучающими компьютерными системами.

Только через систему образования можно поднять уровень массового применения прикладной статистики и сократить отставание от "переднего края" теории. А это отставание в настоящее время составляет не менее 20 (но и не более 100) лет.

## Литература

1. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1965 (1-е изд.), 1968 (2-е изд.), 1983 (3-е изд.).
2. Орлов А.И. О критериях Колмогорова и Смирнова. – Журнал «Заводская лаборатория». 1995. Т.61. № 7. С.59-61.
3. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука, 1979. - 296 с.
4. Смоляк С.А., Титаренко Б.П. Устойчивые методы оценивания: Статистическая обработка неоднородных совокупностей. - М;: Статистика, 1980. - 208 с.
5. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. - М.: Финансы и статистика, 1988. - 263 с.
6. ГОСТ 11.011-83. Прикладная статистика. Правила определения оценок и доверительных границ для параметров гамма-распределения. - М.: Изд-во стандартов. 1984. - 53 с.
7. Орлов А.И. Эконометрика. Учебник для вузов. Изд. 2-е, исправленное и дополненное. - М.: Изд-во "Экзамен", 2003. – 576 с.
8. Орлов А.И., Федосеев В.Н. Менеджмент в техносфере: Учеб. пособие для студ. высш. учеб. заведений. – М.: Издательский центр «Академия», 2003. – 384 с.
9. Суппес П., Зинес Дж. Основы теории измерений. - В сб.: Психологические измерения. -М: Мир,1967. С. 9-110.
10. Пфанцагль И. Теория измерений. - М.: Мир, 1976. - 166 с.
11. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. - М.: Мир, 1976. - 168 с.
12. Дэвид Г. Метод парных сравнений. - М.: Статистика, 1978. - 144 с.
13. Матерон Ж. Случайные множества и интегральная геометрия. - М.: Мир, 1978. - 318 с.
14. Терехина А.Ю. Анализ данных методами многомерного шкалирования. - М.: Наука, 1986. - 168 с.
15. Перекрест В.Т. Нелинейный типологический анализ социально-экономической информации: Математические и вычислительные методы. - Л.: Наука, 1983. - 176 с.
16. Кемени Дж., Снелл Дж. Кибернетическое моделирование: Некоторые приложения. - М.: Советское радио, 1972. - 192 с.
17. Тюрин Ю.Н., Литвак Б.Г., Орлов А.И., Сатаров Г.А., Шмерлинг Д.С. Анализ нечисловой информации. - М.: Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1981. - 80 с.
18. Литвак Б.Г. Экспертная информация: Методы получения и анализа. - М.: Радио и связь, 1982. - 184 с.
19. Орлов А.И. Статистика объектов нечисловой природы и экспертные оценки. - В сб.: Экспертные оценки. Вопросы кибернетики. Вып.58. - М.: Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1979. С.17-33.
20. Анализ нечисловой информации в социологических исследованиях. / Под ред. В.Г. Андреенкова, А.И.Орлова, Ю.Н. Толстовой. - М.: Наука, 1985. - 220 с.

21. Орлов А.И. Асимптотическое поведение статистик интегрального типа. / Доклады АН СССР. 1974. Т.219. № 4. С.808-811.
22. Орлов А.И. Асимптотическое поведение статистик интегрального типа. - В сб.: Вероятностные процессы и их приложения. Межвузовский сборник. - М.: МИЭМ, 1989. С.118-123.
23. Горский В.Г. Современные статистические методы обработки и планирования экспериментов в химической технологии. - В сб.: Инженерно-химическая наука для передовых технологий. Международная школа повышения квалификации Труды третьей сессии. 26-30 мая 1997, Казань, Россия / Под ред. В.А. Махлина. - М.: Научно-исследовательский физико-химический институт им. Карпова, 1997. С.261-293.
24. Орлов А.И. О перестройке статистической науки и её применений / Вестник статистики. 1990. № 1. С.65 - 71.
25. Плошко Б.Г., Елисеева И.И. История статистики: Учебное пособие. - М.: Финансы и статистика. 1990. - 295 с.
26. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В.Прохоров. – М.: Большая Российская энциклопедия, 1999. – 910 с.
27. Орлов А.И. Распространенная ошибка при использовании критериев Колмогорова и омега-квадрат. // Заводская лаборатория. 1985. Т.51. No.1. С.60-62.
28. Контроллинг в бизнесе. Методологические и практические основы построения контроллинга в организациях / А.М. Карминский, Н.И. Оленев, А.Г. Примак, С.Г.Фалько. - М.: Финансы и статистика, 1998. - 256 с.
29. Орлов А. И. Задачи оптимизации и нечеткие переменные. - М.: Знание, 1980.- 64 с.
30. The teaching of statistics / Studies in mathematics education. Vol.7. - Paris, UNESCO, 1989. - 258 pp.
31. Ермаков С.М. Метод Монте-Карло и смежные вопросы. - М.: Наука, 1975. - 471 с.
32. Ермаков С.М., Михайлов Г.А. Статистическое моделирование. - М.: Наука, 1982. - 296 с.
33. Иванова И.М. Случайные числа и их применения. - М.: Финансы и статистика, 1984. - 111 с.
34. Ермаков С.М. О датчиках случайных чисел. // Заводская лаборатория. 1993. Т.59. No.7. С.48-50.
35. Неуймин Я.Г. Модели в науке и технике. История, теория, практика. - Л.: Наука, 1984. - 190 с.
36. Моисеев Н.Н. Математические задачи системного анализа. - М.: Наука, 1981. - 488 с.
37. Нейлор Т. Машинные имитационные эксперименты с моделями экономических систем. - М.: Мир, 1975. - 500 с.
38. Орлов А.И. О реальных возможностях бутстрепа как статистического метода. // Заводская лаборатория. 1987. Т.53. No.10. С.82-85.
39. Бэстенс Д.Э., Берт В.М. ван дер, Вуд Д. Нейронные сети и финансовые рынки: принятие решений в торговых операциях. - М.: ТВП, 1998.
38. Эльясберг П.Е. Измерительная информация. Сколько ее нужно, как ее обрабатывать? - М.: Наука, 1983. - 208 с.
39. Крамер Г. Математические методы статистики. - М.: Мир, 1975. - 648 с.
40. Орлов А.И., Орловский И.В. О поправках на группировку. - В сб.: Прикладной многомерный статистический анализ. - М.: Наука, 1978. - С.339-342.
41. Орлов А.И. Поправка на группировку для коэффициента корреляции. / Экономика и математические методы. - 1980. - Т.XVI. - №4. - С.800-801.
42. Феллер В. Введение в теорию вероятностей и ее приложения. Т.2. - М.: Мир, 1984. - 751 с.
43. Боровков А.А. Теория вероятностей. - М.: Наука, 1976. - 352 с.
44. Бернштейн С.Н. Современное состояние теории вероятностей и ее приложений. - В сб.: Труды Всероссийского съезда математиков в Москве 27 апреля - 4 мая 1927 г. - М.-Л.: ГИЗ, 1928. С.50-63.
45. Каган А.М., Линник Ю.В., Рао С.Р. Характеризационные задачи математической статистики. - М.: Наука, 1972. - 656 с.
46. Ибрагимов И.А., Хасьминский Р.З. Асимптотическая теория оценивания. - М.: Наука, 1979. - 528 с.
47. Никитин Я.Ю. Асимптотическая эффективность непараметрических критериев. - М.: Наука, 1995. - 240 с.

48. Орлов А.И. О современных проблемах внедрения прикладной статистики и других статистических методов. / Заводская лаборатория. 1992. Т.58. № 1. С.67-74.
49. Орлов А.И. О перестройке статистической науки и её применений. / Вестник статистики. 1990. № 1. С.65 - 71.

Приложение 1

## Методологические вопросы прикладной статистики

При разработке и применении методов прикладной статистики необходимо опираться на четкие методологические принципы, разработанные поколениями специалистов. Рассмотрим некоторые из них.

**Задача – модель - метод – условия применимости**. Разработка и применение методов прикладной статистики предполагает последовательное осуществление трех этапов исследования. Первый - от исходной практической проблемы до теоретической чисто математической задачи. Второй – внутриматематическое изучение и решение этой задачи. Третий – переход от математических выводов обратно к практической проблеме.

В литературе вопросы методологии прикладной статистики обсуждаются явно недостаточно. Зато наблюдается поток публикаций, в которых постановки решаемых задач иногда выглядят весьма искусственно. Цель настоящего приложения - обосновать необходимость развития методологии прикладной статистики как самостоятельного научного направления, рассмотреть ряд проблем, относящихся к этому направлению.

В области моделирования задач прикладной статистики, как, впрочем, и в иных областях применения математики и кибернетики, целесообразно выделять четверки проблем:

ЗАДАЧА – МОДЕЛЬ - МЕТОД - УСЛОВИЯ ПРИМЕНИМОСТИ.

Обсудим каждую из только что выделенных составляющих.

Задача, как правило, порождена потребностями той или иной прикладной области. Вполне понятно, что при этом происходит одна из возможных математических формализаций реальной ситуации. Например, при изучении предпочтений потребителей у экономистов - маркетологов возникает вопрос: различаются ли мнения двух групп потребителей (см. главу 1.2). При математической формализации мнения потребителей в каждой группе обычно моделируются как независимые случайные выборки, т.е. как совокупности независимых одинаково распределенных случайных величин, а вопрос маркетологов переформулируется в рамках этой модели как вопрос о проверке той или иной статистической гипотезы однородности. Речь может идти об однородности характеристик, например, о проверке равенства математических ожиданий, или о полной (абсолютной однородности), т.е. о совпадении функций распределения, соответствующих двух совокупностям (см. главу 3.1).

Задача может быть порождена также обобщением потребностей ряда прикладных областей. Приведенный выше пример иллюстрирует эту ситуацию: к необходимости проверки гипотезы однородности приходят и медики при сравнении двух групп пациентов, и инженеры при сопоставлении результатов обработки деталей двумя способами, и т.д. Таким образом, одна и та же математическая модель может применяться для решения самых разных по своей прикладной сущности задач.

Важно подчеркнуть, что выделение перечня задач находится вне математики. Выражаясь инженерным языком, этот перечень является сутью технического задания, которое специалисты различных областей деятельности дают статистикам.

Метод, используемый в рамках определенной математической модели - это уже во многом, если не в основном, дело математиков. В моделях прикладной статистики речь идет, например, о методе оценивания, о методе проверки гипотезы, о методе доказательства той или иной теоремы, и т.д. В двух первых случаях алгоритмы разрабатываются и исследуются математиками, но используются прикладниками, в то время как метод доказательства касается лишь самих математиков.

Ясно, что для решения той или иной задачи в рамках одной и той же принятой исследователем модели может быть предложено много методов. Приведем примеры. Для специалистов по теории вероятностей и математической статистике наиболее хорошо известна история Центральной Предельной Теоремы теории вероятностей. Предельный нормальный закон был получен многими разными методами, из которых напомним теорему Муавра-Лапласа, метод моментов Чебышева, метод характеристических функций Ляпунова, завершающие эпопею методы, примененные Линдебергом и Феллером. В настоящее время для решения практически важных задач могут быть использованы современные информационные технологии на основе метода статистических испытаний и соответствующих датчиков псевдослучайных чисел. Они уже

заметно потеснили асимптотические методы математической статистики. В рассмотренной выше проблеме однородности для проверки одной и той же гипотезы совпадения функций распределения могут быть применены самые разные методы – Смирнова, Лемана - Розенблатта, Вилкоксона и др. (см. главу 3.1).

Наконец, рассмотрим последний элемент четверки - условия применимости. Он - полностью внутриматематический. С точки зрения математика замена условия (кусочной) дифференцируемости некоторой функции на условие ее непрерывности может представляться существенным научным достижением, в то время как прикладник оценить это достижение не сможет. Для него, как и во времена Ньютона и Лейбница, непрерывные функции мало отличаются от (кусочно) дифференцируемых функций. Точнее, они одинаково хорошо (или одинаково плохо) могут быть использованы для описания реальной действительности.

Точно также прикладник не сможет оценить внутриматематическое достижение, состоящее в переходе от условия конечности четвертого момента случайной величины к условию конечности дисперсии. Поскольку результаты реальных измерений получены с помощью некоторого прибора (средства измерения), шкала которого конечна, то прикладник априори уверен, что все результаты измерений заведомо лежат на некотором отрезке (т.е. финитны). Он с некоторым недоумением наблюдает за математиком, который рассуждает о конечности тех или иных моментов - для прикладника они заведомо конечны.

**Математики и прикладники.** Таким образом, в настоящее время наблюдается значительное расхождение интересов "типового" математика и "типового" прикладника. Конечно, мы рассуждаем здесь, строя гипотетические модели восприятия и поведения того и другого. Опишем эти модели более подробно.

Прикладник заинтересован в научно обоснованном решении стоящих перед ним реальных задач. При этом при формализации задач он готов принять достаточно сильные математические предположения. Например, с точки зрения прикладника случайные величины могут принимать конечное множество значений, или быть финитными, или иметь нужное математику число моментов, и т.д. Как говорил А.Н. Колмогоров, переход от дискретности к непрерывности для прикладника оправдан только тогда, когда этот переход облегчает выкладки и расчеты, как в математическом анализе переход от сумм к интегралам облегчает рассуждения и вычисления. Если же при переходе к непрерывности возникают сложности типа необходимости доказательства измеримости тех или иных величин относительно тех или иных сигма-алгебр, то прикладник готов вернуться к постановке задачи с конечным вероятностным пространством. Здесь уместно напомнить, что один из выдающихся вероятностников XX в. В. Феллер выпустил свой учебник по теории вероятностей в двух книгах, посвятив первую дискретным вероятностным пространствам, а вторую - непрерывным.

Другой пример - задачи оптимизации. Если оптимизация проводится по конечному множеству, то оптимум всегда достигается (хотя может быть не единственным). Если же множество параметров бесконечно, то задача оптимизации может и не иметь решения. Поэтому у прикладника есть стимул ограничиться математическими моделями с конечным множеством параметров. Напомним в связи с этим, что основные задачи прикладной статистики допускают оптимизационную постановку, а статистика объектов нечисловой природы как целое построена на решении оптимизационных задач (а не на суммировании тех или иных выражений, поскольку в пространствах объектов нечисловой природы нет операции сложения).

Модель поведения типового математика совершенно иная. Он, как правило, не обдумывает реальные задачи, поскольку не вникает в конкретные прикладные области. (Если же вникает, то является уже не только математиком, но и прикладником, и его поведение промоделировано в предыдущих абзацах.) Математик берет те задачи, которые уже ранее рассматривались, и старается получить для них математически интересные результаты. Зачастую это означает борьбу за ослабление математических условий, при которых были получены предыдущие результаты. При этом математика абсолютно не волнует, имеют ли какое-либо реальное содержание доказанные им теоремы, могут ли они принести какую-либо пользу прикладнику. Его интересует реакция математической общественности, а не реакция прикладников.

**Сколько реально используется чисел?** Для демонстрации разрыва между математиками и прикладниками обратим внимание на два парадокса.

Все реальные результаты наблюдений записываются рациональными числами (обычно десятичными числами с небольшим - от 2 до 5 - числом значащих цифр). Как известно, в

математике множество рациональных чисел счетно, а потому вероятность попадания значения непрерывной случайной величины в него равно 0. Следовательно, все рассуждения, связанные с моделированием непрерывными случайными величинами реальных результатов наблюдений - это рассуждения о том, что происходит внутри множества меры 0. Первый парадокс состоит в том, что множествами меры 0 в теории вероятностей принято пренебрегать. Другими словами, в точки зрения теории вероятностей всеми реальными данными можно пренебречь, поскольку они входят в одно фиксированное множество меры 0.

Глубже проанализируем ситуацию. Сколько всего чисел используется для записи реальных результатов наблюдений? Речь идет о типовых результатах наблюдений, измерений, испытаний, опытов, анализов. Они используются в технических, естественнонаучных, экономических, социологических, медицинских и иных исследованиях. Анализ практики показывает, что эти числа имеют вид $(a,bcde)10^k$. Здесь $a$ принимает значения от 1 до 9, а стоящие после запятой $b, c, d, e$ - от 0 до 9. В то же время показатель степени $k$ меняется от (-100) до +100. Ясно, что общее количество возможных чисел равно $9x10^4x201=18090000$, т.е. меньше 20 миллионов.

Итак, второй парадокс, усиливающий первый, состоит в том, что для описания реальных результатов наблюдений вполне достаточно 20 миллионов отдельных символов. Бесконечность натурального ряда и континуум числовой прямой - это математические абстракции, надстроенные над дискретной и состоящей из конечного числа элементов реальностью. (При изменении числа значащих цифр, используемых для описания результатов наблюдений, принципиальный вывод не меняется.) Таким образом, реальные данные лежат не только во множестве меры 0, но и в конечном множестве, причем число элементов в этом множестве вполне обозримо.

**Практические следствия методологии прикладной статистики.** Из сказанного вытекают некоторые вполне определенные выводы, в том числе касающиеся преподавания и научных исследований.

Например, преподавание теории вероятностей может быть сосредоточено на случае конечного вероятностного пространства. Бесконечные вероятностные пространства могут при этом рассматриваться как удобные математические схемы. Их роль – давать возможность более легко и быстро получать полезные утверждения для конечных вероятностных пространств. Из сказанного вытекает, в частности, что различные параметрические семейства распределений (семейства нормальных, логарифмически нормальных, экспоненциальных, Коши, Вейбулла-Гнеденко, гамма-распределений) приобретают статус не более чем удобных приближений для распределений на конечных вероятностных пространствах. При таком подходе теряет свою парадоксальность тот эмпирически не раз проверенный факт, что распределение погрешностей измерений, как правило, не является гауссовым (см. главу 2.1).

В качестве другого примера рассмотрим методы оценивания параметров. По традиции много внимания в учебных курсах уделяется оценкам максимального правдоподобия (ОМП). Однако столь же хорошие асимптотические свойства имеют т.н. одношаговые оценки, гораздо более простые с вычислительной точки зрения (см. главу 2.2). Целесообразно их включить в учебные курсы, а ОМП исключить.

Целесообразно уделять внимание (репрезентативной) теории измерений, в частности, концепции шкал измерения. Необходимо знакомство с определениями и основными свойствами шкал наименований, порядковой, интервалов, отношений, разностей, абсолютной. Установлено, какими алгоритмами статистического анализа данных можно пользоваться в той или иной шкале, в частности, для усреднения результатов наблюдений. Так, для данных, измеренных в порядковой шкале, некорректно вычислять среднее арифметическое. В качестве средних величин для таких данных можно использовать порядковые статистики, в частности, медиану.

Статистические методы исследования часто опираются на использование современных информационных технологий. В частности, распределение статистики можно находить методами асимптотической математической статистики, а можно и путем статистического моделирования (метод Монте-Карло, он же - метод статистических испытаний).

Методологический анализ - первый этап моделирования задач принятия решений, да и вообще любого исследования. Он определяет исходные постановки для теоретической проработки, а потому во многом и успех всего исследования.

Методологический анализ - первый этап статистического исследования. Он определяет исходные постановки для теоретической проработки, а потому во многом и успех всего исследования [1]. Этот этап - один из наиболее важных [2]. Подчеркнем, что анализ динамики

развития методов прикладной статистики выделить наиболее перспективные методы. В частности, в работе [3] установлено, что в настоящее время наиболее перспективными являются методы нечисловой статистики. Именно поэтому им уделено большое внимание в настоящем учебнике.

## Литература

1. Комаров Д.М., Орлов А.И. Роль методологических исследований в разработке методоориентированных экспертных систем (на примере оптимизационных и статистических методов). – В сб.: Вопросы применения экспертных систем. - Минск: НПО «Центрсистем», 1988. С.151-160.
2. Орлов А.И. О развитии методологии статистических методов. - В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. – Пермь: Изд-во Пермского государственного университета, 2001. – С.118-131.
3. Горский В.Г., Орлов А.И. Математические методы исследования: итоги и перспективы. - Журнал «Заводская лаборатория». 2002. Т.68. № 1. С.108-112.

Приложение 2

## Глазами американцев: российская дискуссия о прикладной статистике

Развитие прикладной статистики в нашей стране сопровождалось бурными дискуссиями. Объективный анализ их начального этапа был дан на страницах органа Американской статистической ассоциации [1]. Статья Сэмюеля Котца и Кэтлин Смит «Пространство Хаусдорфа и прикладная статистика: точка зрения ученых СССР» описывает различные взгляды, имеющие распространение и в XXI веке. Представляется поучительным привести изложение их работы (использован перевод статьи, выполненный М.В. Ильинской).

Несколько слов об авторах. На момент написания статьи профессор Сэмюель Котц, специалист в области статистических наук, работал в Колледже коммерческой деятельности и управления Мэрилендского университета. Кэтлин Смит была аспиранткой отделения по изучению России Калифорнийского университета (Беркли). В 1987 г. они сотрудничали с Кеннанским институтом повышения квалификации специалистов, изучающих Россию.

Статья Сэмюеля Котца и Кэтлин Смит посвящена дискуссии, развернувшейся на страницах советского статистического журнала "Вестник статистики" по вопросам существования и релевантности (уместности) прикладной статистики как самостоятельной научной дисциплины. В ней анализируется содержание четырех писем редактору и редакционных комментариев к ним, которые были опубликованы в этом журнале в период с октября 1985 г. по июнь 1987 г.

Основная задача статьи состоит в том, чтобы осветить длительную (продолжающуюся по крайней мере 40 лет) полемику в советской (и российской – А.О.[1]) статистике между "идеологическими пуристами" и "прагматиками", которая в 1980-е годы значительно усилилась. Существование разногласий, безусловно, не является новым явлением среди статистиков и в определенной степени оно носит здоровый характер, способствуя выработке критического отношения к предмету. На данном этапе полемика в СССР затрагивает суть предмета в отличие от более ранних этапов, когда она отличалась идеологической направленностью. В 1950 - 1960 годах, в период хрущевской оттепели, когда в СССР более свободно начали публиковать статистические данные, в журнале "The American Statistician" ("Американский статистик" – орган Американской статистической ассоциации – А.О.) было опубликовано несколько статей, посвященных различным аспектам советской статистики, как организационным, так и затрагивающим существо предмета. Мы надеемся, что движение в этом направлении будет более энергичным. Это несомненно приведет к лучшему пониманию как глобальных, так и специфических региональных проблем, связанных с беспрецедентным развитием статистики, свидетелями которого мы являемся.

*Советская статистика: 1917 – 1964.* Вопросы развития статистики в СССР с 1917 по 1964 г. были довольно подробно освещены первым автором в статьях [2, 3]. В указанных статьях рассматривалась история статистики в СССР в данный период. В них автор в общих чертах рассказал о появлении двух противоположных мнений по вопросу о роли и содержании статистической науки в СССР. Между официальными статистиками Центрального статистического управления (ЦСУ) и статистиками - экономистами математической направленности во главе с В.С. Немчиновым (1890 - 1964) возникли разногласия.

Официальные статистики считали, что статистика представляет собой описательную науку, в задачи которой входит сбор данных по плановой экономике, и что в условиях коммунизма статистику в конечном счете заменит простая бухгалтерия. Противоположных взглядов придерживались практики и статистики теоретической направленности. Они считали, что статистика и теория вероятностей важны в любой области. В 1954 г. на Всесоюзной конференции по статистике, в работе которой приняли участие ведущие ученые, известный советский математик А.Н.Колмогоров (1903-1987) помог представителям этих двух противостоящих школ прийти к прагматическому компромиссу. В своих статьях Котц коснулся роли Колмогорова в

---

[1] Примечания автора настоящего учебника здесь и далее подписаны инициалами А.О.

урегулировании этих разногласий [2, 3]. На конференции 1954 г. было заявлено, что статистика является самостоятельной общественной наукой и что "она изучает количественный аспект массовых социальных явлений в неразрывном единстве с их качественным аспектом " (см. Котц, [3, с.136]). Был сделан вывод, что советскую статистику от "буржуазной" статистики отличает акцент на качественном аспекте явлений. Для "буржуазной" статистики, согласно официальной оценке в Советском Союзе, характерен формальный, чисто математический подход к изучению социальных явлений, при котором количественный показатель рассматривается отдельно от качественной основы.

В соответствии с постановлением Верховного Совета СССР от 29 июля 1987 г. ЦСУ было реорганизовано и возведено в ранг Государственного комитета СССР по статистике (Госкомстат СССР).

*Разногласия на современном этапе.* Появление статьи Сэмюеля Котца и Кэтлин Смит связано с началом публикации в 1965 году полупериодического журнала "Ученые записки по статистике" под редакцией Немчинова (т.е. серии сборников статей, выпускавшихся издательствами "Наука", "Статистика", "Финансы и статистика"). Предполагалось, что этот журнал, который начали публиковать после конференции 1954 г., будет выходить наряду с американским журналом "Journal of the American Statistical Association" ("Журнал Американской статистической ассоциации") и английским журналом " Journal of the Royal Statistical Society, Series C - Applied Statistics" ("Журнал Королевского статистического общества, Серия С - Прикладная статистика").

В 1986 году вышел юбилейный 50-й выпуск сборника "Ученые записки по статистике". В нем опубликовали свои статьи статистики математической ориентации. Многие из них - выпускники и кандидаты наук престижной школы теории вероятностей и математической статистики при МГУ, которую первоначально возглавлял А.Н. Колмогоров, и такой же школы при Ленинградском университете, во главе которой некоторое время стоял Ю.В. Линник. Эти ученые работали в больших городах, в различных институтах, занимающихся вопросами применения прикладной статистики. Ученые выполняли ориентированные на практическое применение работы в теории управления запасами, прикладном многомерном анализе и т.д., однако создается впечатление, что они испытывали желание заниматься вопросами, носящими более математический характер. Эта тенденция нашла свое отражение на страницах сборника "Ученые записки по статистике", в котором постепенно, но постоянно начали публиковать статьи математического и абстрактного характера, что вызвало недовольство среди статистиков различных научно-исследовательских институтов, связанных с ЦСУ.

В 1983 году в издательстве «Наука» вышел в свет 45-й выпуск "Ученых записок по статистике" под редакцией А.И. Орлова и С.А. Айвазяна, который был скромно озаглавлен "Прикладная статистика", и разразился скандал. В данной статье мы (Сэмюель Котц и Кэтлин Смит – А.О.) опишем ход развития полемики, проанализировав содержание четырех писем редактору, которые были опубликованы с октября 1985 г. по июль 1987 г. в ежемесячном журнале "Вестник статистики" - органе ЦСУ.

Основными сторонниками становления и развития прикладной статистики как центра статистической науки являются в СССР А.И. Орлов и С.А. Айвазян. Орлов - плодовитый исследователь, тяготеющий к абстракции. В 1975 г. он написал кандидатскую диссертацию в МГУ, под руководством Л.Н. Большева. Диссертация была связана с асимптотическим распределением статистик типа Смирнова, но он еще ранее сотрудничал с Ю.Н. Тюриным по проблемам анализа нечисловой информации и нечетких множеств и опубликовал большое количество работ по широкому кругу вопросов, от чисто математических до экономических. Айвазян, доктор физико-математических наук, работал в области прикладной статистики. В журнале "Экономика и математические методы" за 1966 г. мы обнаружили одну из его ранних работ - обзорную статью по математической статистике с уклоном в прикладную статистику. Он продолжал публиковать работы по вопросам многомерного статистического анализа, и проводил конференции и семинары по прикладной статистике. В сотрудничестве с И.С. Енюковым и Л.Д. Мешалкиным он написал и отредактировал два справочника по прикладной статистике, которые были изданы в 1983 г. и 1985 г. Из предисловий к этим книгам ясно, что участие в подготовке этих изданий принимал Орлов.

В ответ на публикацию в сборнике "Ученые записки по статистике" многочисленных математических статей абстрактного характера К. Тимофеев (псевдоним – А.О.) написал сердитое

письмо под заголовком "Что же такое прикладная статистика?" [4]. Он утверждал, что термин "прикладная статистика" является абсурдным, так как то, что она якобы описывает, является одной из областей статистической науки, а не новым направлением. Тимофеев заявил: "Из содержания представленных в 45-м томе статей становится совершенно очевидным: название "Прикладная статистика" использовано для того, чтобы в "Ученых записках по статистике" опубликовать материалы, которые к ней не имеют ни прямого, ни даже косвенного отношения" [4, с.66]. Кроме этого, он выразил несогласие с рядом приведенных в сборнике математических формул и абстрактных концепций. В частности, он привел цитату из статьи, в которой говорится, что статья посвящена "измеримым отображениям произвольного вероятностного пространства в множество непустых компактов плоскости, снабженное метрикой Хаусдорфа" (метрика Хаусдорфа – одно из расстояний между множествами; статья была озаглавлена «Статистика случайных множеств» – А.О.). Тимофеев не только не захотел перенестись «в другое измерение», он подверг автора критике за то, что он в своей статье сослался на работы зарубежных ученых, а не на работы классиков марксизма-ленинизма и советские статистические источники, а также за то, что он написал работу, не связанную с реальной жизнью. Он с неодобрением указал, что авторы статей, публикуемых в «Ученых записках по статистике», часто ссылаются на свои собственные работы. Он написал: "Создается впечатление, что книга "Прикладная статистика" использована не только для публикации не относящихся к статистике материалов, но и для рекламы и саморекламы некоторых математиков, решивших снискать себе славу в области экономики и статистики" [4, с.67]. Тимофеев признал, что эти статьи могут представлять определенный интерес для математиков, однако он полагал, что они вряд ли будут полезны в практической работе тем специалистам, на службе у которых, по его мнению, должна быть статистическая наука, а именно статистикам, экономистам и социологам.

Через десять месяцев журнал "Вестник статистики" опубликовал ответ [5] на выступление Тимофеева. Один из авторов, которых критиковал Тимофеев, А. Орлов, написал ответ в таком же резком тоне, и он был опубликован в официальном органе ЦСУ. В своей статье, перед которой было напечатано вступление от редакции, Орлов пункт за пунктом критиковал позицию Тимофеева. Орлов представил себя, как современного статистика. В начале и в конце своего длинного письма он привел выдержки из речей М.С. Горбачева (в 1985-1991 гг. – генеральный секретарь ЦК КПСС – А.О.), который говорил о необходимости стимулировать развитие современных подходов к решению социально-экономических и научно-технических проблем, о том, что "на задачи науки необходимо взглянуть по-новому" [5, с.56]. Он написал, что Тимофеев запутался и не знаком с переменами, которые произошли в статистике. Он отметил, что термин "прикладная статистика" не является ни новым, ни редко употребляемым. Он используется специалистами различных учреждений по всей стране. Он провел грань между математической статистикой и прикладной статистикой, добавив, что прикладная математическая статистика является "неотъемлемой частью" прикладной статистики, а прикладная математическая статистика наряду с аналитической статистикой составляют математическую статистику, которая является одной из областей математики. Однако Орлов подчеркнул, что прикладная статистика включает и нематематические области, такие, как:

> методология организации и проведения прикладного статистического исследования и применения его результатов: как планировать исследование, как выбирать вероятностно-статистическую модель, как собирать данные, как подготавливать их к обработке, как представлять результаты обработки и т.д. (Орлов, 1985, стр.54).

Далее он сказал, что использование в прикладной статистике программирования на вычислительных машинах свидетельствует о том, что в действительности ее можно рассматривать как часть кибернетики.

Орлов привел много примеров использования прикладной статистики в народном хозяйстве, сделав акцент на планировании эксперимента и контроле качества. Он отметил, что благодаря прикладной статистике была получена большая экономия финансовых средств: "Высокая эффективность прикладной статистики естественна - она родилась из практических нужд" [5, с.54]. Он охарактеризовал большой вклад в практическую работу, который внесли многие из тех статей, которые Тимофеев высмеял за абстрактные заголовки. В заключение статьи он привел таблицу, из которой видно, что ученые, публикующие свои работы в "Ученых записках по статистике", чаще ссылаются на работы советских авторов, чем зарубежных, и он подчеркнул, что эти авторы опираются на опыт своей практической работы, а не повторяют ранее

опубликованный материал. Он составил эту таблицу на основе советского реферативного журнала «Математика», в котором "советские публикации составляют 1/6 мировых публикаций по прикладной статистике, реферируемых за год" [5, с.56].

Однако, по-видимому, редакторов журнала "Вестник статистики" не убедили доводы Орлова. В дополнение к его письму они напечатали свое заявление о том, что письмо Тимофеева было опубликовано для того, чтобы показать, что сборник "Ученые записки по статистике" перестал отвечать своей цели и превратился в математический журнал и что содержание статей в "Прикладной статистике" (вып. 45 "Ученых записок по статистике") не отвечает названию сборника. Более того, редакторы добавили, что находят доводы Тимофеева убедительными. Выступив с критикой письма Орлова, они упрекнули его за то, что он пытается "опровергнуть содержание письма К. Тимофеева, а заодно изобразить его автора как человека, не сведущего в делах, которыми занимается А. Орлов, а с ним и ряд других математиков" (стр.57). Они продолжали утверждать, что многие леммы и теоремы, которыми оперирует Орлов и его коллеги, не используются в практической работе. В частности, они проявили упорное желание узнать, "каков экономический эффект (в миллионах рублей), который удалось извлечь из шума при помощи измеримых отображений "произвольного вероятностного пространства в множество непустых компактов плоскости, снабженное метрикой Хаусдорфа" [5, с.57]. Касаясь ссылок на работы зарубежных авторов, редакторы отметили, что из таблицы Орлова видно, что ученые действительно ссылаются на зарубежные источники, и таким образом они приходят к выводу, что их утверждение верно. [Обширные политизированные тексты «редакторов», весьма враждебные, но не подписанные, демонстрируют уместность – в данном случае - ссылок на выступления М.С. Горбачева в моем «письме» – А.О.]

Подтверждением того, что спорные вопросы еще не решены, по крайней мере в умах читателей, явилась публикация третьего письма, написанного Н. Шереметом [6]. Шеремет, преподаватель Московского института железнодорожного транспорта, придерживается умеренных взглядов по вопросу об определении прикладной статистики и ее роли. [Шеремет часто публикует свои статьи в журнале "Вестник статистики", в частности, он рецензировал новый учебник по теории статистики известного экономиста-статистика А.Я. Боярского (1906-1985) и более молодого, но также крупного экономиста-статистика Г.Л. Громыко.] В начале своего письма он отметил, что Тимофеев не ответил на свой собственный вопрос: "Что же такое прикладная статистика?" По мнению Шеремета, прикладные науки являются связующим звеном между чисто техническими работами и научными исследованиями или чистой наукой. Он выступил в защиту необходимости стадии "корректировки" или "подстройки" между стадиями научных изысканий и применением научных теорий на практике. Затем он привел хорошо известное мнение Большева о том, что вся статистика является прикладной (Л.Н. Большев высказал это мнение в личной беседе с А.И.Орловым, цитата была включена в статью [5] – А.О.), но не поддержал это утверждение, так как оно является слишком широким обобщением. Затем Шеремет проанализировал точку зрения, что каждая наука имеет свою собственную статистику (например, физическая статистика и биологическая статистика), но отверг ее, так как она противоречит мнению Ф.Энгельса, высказанному при подобных обстоятельствах в связи с механикой, физикой и химией. Шеремет критиковал Орлова за примеры из области экономики, так как эти примеры могли привести к ошибочному предположению, что прикладная статистика является универсальной наукой. (Шеремет также подчеркнул эту мысль в своей рецензии на учебник Боярского – Громыко; он упорно придерживается неизменного официального определения статистики как общественной науки, хотя в вопросах, относящихся к другим областям статистики, его позиция характеризуется определенной гибкостью.)

Шеремет настаивал на определении статистики как общественной науки, однако признает возможность использования прикладной статистики в своей собственной области. Шеремет написал в свойственных ему неопределенных выражениях:

> "Можно предположить, что предметом данной научной дисциплины являются "статистические данные"... Здесь уже не важно, от какого реального явления отвлечены данные абстрактные понятия... Математическая идеализация "статистических данных" и операций над ними дает возможность сводить известное разнообразие связей и закономерностей конкретной практической области к их определенному классу, производить необходимые расчеты" [6, с.69].

Он заявил, что прикладная статистика пока еще не является четко определенной областью, и в заключение написал, что "прикладной статистике" в большей степени присущи черты междисциплинарных исследований, чем исследований, проводимых в рамках самостоятельной дисциплины [6, с.71].

В письме Шеремет допустил несколько неточностей, граничащих с дезинформацией. Он, кажется, не знает, что с 1973 г. журнал «Анналы статистики» ("The Annals of Statistics" – основной западный статистический журнал – А.О.) является непосредственным продолжением журнала "Анналы математической статистики" ("Annals of Mathematical Statistics") и не делает разницы между узким техническим термином "статистика" (как функция от результатов наблюдений) и термином "статистика" (как наука и методология). Ссылка на элементарный учебник Вайнберга и Шумахера 1969 г. [7] как на образцовую современную монографию по прикладной статистике в лучшем случае вызывает сомнение.

Показательным является сам факт публикации подобного письма без редакционного комментария в советском консервативном журнале по статистическим наукам - в журнале, который со времени своего возрождения в 1949 году стал выразителем позиций официальных статистиков (многие из них строго придерживаются марксистко-ленинской ориентации), рассматривающих статистику только как описательную науку.

На страницах "Вестника статистики" письмо Шеремета было не единственным откликом на полемику между Тимофеевым и Орловым. По всей видимости, независимо от письма Шеремета в июле 1987 года "Вестник статистики" опубликовал письмо И. Манделя [8], сотрудника института Народного хозяйства в Алма-Ате (Казахстан). В качестве комментария на письма Тимофеева и Орлова Мандель составил развернутую схему, отражающую взаимосвязь теории статистики, прикладной статистики и математической статистики. Эта схема была представлена наряду с шестью другими методологическими приемами, чтобы показать, какое влияние оказывают теория статистики, прикладная статистика и математическая статистика на методы исследования массовых явлений. Главным в его доводах является положение о том, что в то время, как "теория статистики" в основном отражает "социальную сферу" массовых процессов, прикладная статистика должна быть направлена на отражение массовых явлений любого характера. Таким образом, прикладная статистика должна являться своего рода "буферной наукой", которая переводит результаты математической статистики на язык, понятный исследователям в различных областях науки и практики. Он высказал сожаление по поводу существующих расхождений во взглядах между чистыми математическими статистиками и чистыми "прикладными" и обратил внимание на многочисленные примеры неправильного использования статистической методологии. Он приветствовал усилия математиков (в СССР и за рубежом), направленные на ликвидацию разрыва между математикой и реальным миром. В заключение он посоветовал называть прикладную статистику в значении "буферной науки" "прикладной математической статистикой". На конкретный вопрос о том, является ли сборник "Ученые записки по статистике" подходящим изданием для публикации статей по прикладной (математической?) статистике, он дает категорический отрицательный ответ, полностью совпадающий с мнением Тимофеева по этому вопросу. Мандель составил таблицу, согласно которой в 4 выпусках "Записок" (1978-1985), подготовленных прикладными статистиками, опубликовано 85 статей (1092 стр.). Из них 62 статьи (787 стр.), т.е. почти три четверти, по его мнению, по своему содержанию больше подходили для публикации в известном советском журнале "Теория вероятностей и ее применения", так как были посвящены чисто математическим результатам и написаны в виде теорем и доказательств. [Мандель, увы, не знал, что к тому времени авторы журнала "Теория вероятностей и ее применения" уже полностью оторвались от практики анализа статистических данных – А.О.] По мнению Манделя, отличительной чертой прикладной статистики является отсутствие доказательств; для нее характерны только ссылки на теоремы и обсуждение вопросов «истинно» прикладного характера.

Обсуждение было продолжено в феврале 1988 г., когда в очередном выпуске "Вестника статистики" было опубликовано письмо болгарского профессора, специалиста по статистике, В. Цонева [9]. Он предлагает коренным образом изменить терминологию, связанную со всей статистической наукой.

*Перестройка в области статистики*. В настоящее время в СССР мы обнаруживаем все признаки перестройки в области статистики. Они проявляются не только в публикации новых статистических данных по промышленному травматизму, алкоголизму, преступности и т.д., но

также и в координации работы многочисленных учреждений, занимающихся обработкой статистических данных. Недавняя реорганизация ЦСУ является еще одним свидетельством озабоченности правительства недостатками в данной области. К примеру, статистические данные, связанные с производством черных металлов, собираются и обрабатываются тремя учреждениями - Госпланом, ЦСУ и Институтом экономики министерства черной металлургии. На Всесоюзной конференции статистиков в мае 1985 г. выяснилось, что данные по прокату черных металлов, поступающие из этих трех источников, "совершенно разные" [10]. В феврале 1987 г. литературно-художественный журнал "Новый мир" выступил с открытой и резкой критикой отсутствия достоверных статистических данных. Несколько статистиков, среди них – Н. Шеремет и Т. Козлов, заведующий кафедрой статистики Московского института инженеров железнодорожного транспорта - выступили с резким опровержением. В настоящее время продолжается спор по этому вопросу.

Перестройка в СССР, вероятно, приведет к значительным переменам в отношении к статистическим наукам. Реорганизация всесоюзного ЦСУ в июле 1987 г. (наряду с подобными реорганизациями республиканских статистических учреждений) и еще более радикальные перемены в Госплане свидетельствуют, что новое советское руководство озабочено такими проблемами, как точность, достоверность, масштаб государственной статистики и анализ статистических данных.

В любом случае, разногласия между учеными, о которых говорилось в статье, характерны не только для Советского Союза. Американские и другие западные статистики также сталкиваются с проблемой определения роли прикладной статистики и, в более широком плане, с проблемой определения статистики как науки.

*Что было позже* (от автора настоящего учебника). В марте 1989 г. в Центральном экономико-математическом институте АН СССР состоялся Всесоюзный круглый стол «Статистика и перестройка», на котором собрались представители различных направлений в статистике – впервые в отечественной истории! Выступления были опубликованы в виде 55-го тома «Ученых записок по статистике» [11].

Высшей точкой было создание в 1990 г. Всесоюзной статистической ассоциации (ВСА), объединившей статистиков всех направлений – специалистов по прикладной и математической статистике, по надежности (в основном представителей оборонных отраслей промышленности), преподавателей экономико-статистических дисциплин, работников Госкомстата [12, 13]. Ведущую роль в создании ВСА сыграли работники Всесоюзного центра статистических методов и информатики. Наша платформа была изложена в статье [14], опубликованной, несмотря на ее весьма резкую форму, в «Вестнике статистики». Устав ВСА, решения Учредительного съезда и Пленума правления ВСА предусматривали различные формы работы [15].

Однако в 1991 г. СССР прекратил свое существование. ВСА, как и другие союзные организации, перестала действовать. И наметившееся единство статистиков распалось. Госкомстат РФ полностью «закрылся» от статистической науки, перестал даже отвечать на обращения профессиональных статистических организаций. Произошел окончательный отрыв специалистов математической статистики от практики. В настоящее время журнал "Теория вероятностей и ее применения" не представляет никакого интереса для тех, кто обрабатывает конкретные данные.

Работы по прикладной продолжались в рамках Российской ассоциации статистических методов (созданной на базе одноименной секции ВСА) и Российской академии статистических методов. Основным местом публикации отечественных работ по прикладной статистике является секция «Математические методы исследования» журнала «Заводская лаборатория», созданная в 1961 г. Б.В. Гнеденко и В.В. Налимовым. В ней за более чем 40 лет помещено около 1000 статей по различным направлениям прикладной статистики, прежде всего по статистическому анализу числовых величин, статистике нечисловых данных, многомерному статистическому анализу, планированию эксперимента, опыту применения статистических методов при решении конкретных прикладных задач.

## Литература

1. Kotz S., Smith K. The Hausdorff Space and Applied Statistics: A View from USSR // The American Statistician. 1988. Vol. 42. № 4. P. 241-244.

2. Kotz S. Statistical Terminology - Russian Vs. English - in the Light of the Development of Statistics in the USSR // The American Statistician, 1965. Vol. 19, №. 3, P. 16-22.
3. Kotz S. Statistics in the USSR // Survey, 1965. Vol. 57, October, P. 132-141.
4. Тимофеев К. Что же такое прикладная статистика? // Вестник статистики. 1985. № 10. С.66-67.
5. Орлов А. Что дает прикладная статистика народному хозяйству? // Вестник статистики. 1986. № 8. С.52-57.
6. Шеремет Н. О так называемой прикладной статистике // Вестник статистики, 1987. № 2. С.67-71.
7. Weinberg J.H., Schumaker J. Statistics: An Intuitive Approach (2-nd ed.). - Belmont, CA: Brooks-Cole. 1969.
8. Мандель И. Теория статистики и прикладная статистика // Вестник статистики. 1987. № 7. С.76-79.
9. Цонев В. К дискуссии по вопросу: что же такое прикладная статистика // Вестник статистики, 1988, № 2. С.67-68.
10. Маркович М. Хроника и информация // Вестник статистики. 1986. № 11. С.62-64.
11. Статистика и перестройка: Ученые записки по статистике. Т.55. – М.: Наука, 1991. – 280 с.
12. Орлов А.И. Создана единая статистическая ассоциация // Вестник Академии наук СССР. 1991. № 7. С.152-153.
13. Орлов А.И. Всесоюзная статистическая ассоциация - гарантия успешного внедрения современных статистических методов // Надежность и контроль качества. 1991. № 6. С.54-55.
14. Орлов А.И. О перестройке статистической науки и её применений // Вестник статистики. 1990. № 1. С.65 – 71.
15. Устав Всесоюзной статистической ассоциации (ВСА). 1-й Пленум Правления ВСА // Вестник статистики. 1991. № 2. С.71-76.

Приложение 3

## Об авторе этой книги

Орлов Александр Иванович, 1949 г.р., профессор (1995 г. – по кафедре математической экономики), доктор технических наук (1992 г. - по применению математических методов), кандидат физико-математических наук (1976 г. – по теории вероятностей и математической статистике). Профессор кафедры "Экономика и организация производства" факультета "Инженерный бизнес и менеджмент" Московского государственного технического университета им. Н.Э. Баумана, руководитель секции "Эконометрика", директор Института высоких статистических технологий и эконометрики. Профессор кафедры "Экология и право" Московского государственного института электроники и математики (технического университета), научный руководитель Лаборатории эконометрических исследований. Профессор кафедры «Анализ стохастических процессов в экономике» Российской экономической академии им. Г.В. Плеханова. Визит-профессор (в 2002-2003 г.г.) Академии народного хозяйства при Правительстве Российской Федерации, Международного университета (в Москве), Всероссийского государственного института кинематографии (ВГИК), Московского государственного университета прикладной биотехнологии, Международного юридического института при Министерстве юстиции Российской Федерации. Член редколлегий журналов «Заводская лаборатория», «Контроллинг», «Социология: методология, методы, математические модели». Главный редактор электронного еженедельника «Эконометрика». Академик Российской Академии статистических методов, член-корреспондент МО "СовАсК" (Международная организация «Советская Ассоциация Качества»). Вице-президент Всесоюзной Статистической Ассоциации, президент Российской ассоциации статистических методов..

Основные направления научной и педагогической деятельности: прикладная статистика и другие статистические методы, эконометрика, теория принятия решений, экономико-математические методы, экспертные оценки, менеджмент, экономика предприятия, макроэкономика, экология. Автор более 500 публикаций в России и за рубежом, в том числе более 20 книг.

## Основные книги проф. А.И.Орлова

1. Устойчивость в социально-экономических моделях. - М.: Наука,1979. - 296 с.
2. Задачи оптимизации и нечеткие переменные. - М.: Знание, 1980. - 64 с.
3. Анализ нечисловой информации (препринт) (совместно с Ю.Н. Тюриным, Б.Г. Литваком, Г.А. Сатаровым, Д.С. Шмерлингом). - М.: Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1981. - 80 с.
4. Внеклассная работа по математике в 6-8 классах (совместно с В.А. Гусевым, А.Л. Розенталем). - М.: Просвещение, 1977. - 288 с. - Второе издание, исправленное и дополненное (М.: Просвещение, 1984). Переводы на казахский, литовский, молдавский языки.
5. Пакет программ анализа данных "ППАНД". Учебное пособие (совместно с И.Л. Легостаевой, О.М. Черномордиком и др.). - М.: Сотрудничающий центр Всемирной организации здравоохранения по профессиональной гигиене, 1990. - 93 с.
6. Математическое моделирование процессов налогообложения (подходы к проблеме) (совместно с В. Г. Кольцовым, Н.Ю. Ивановой и др.). - М.: Изд-во ЦЭО Министерства общего и профессионального образования РФ, 1997. – 232 с.
7. Экология. Учебное пособие (совместно с С.А. Боголюбовым и др.). - М.: Знание, 1999. - 288 с.
8. Менеджмент. Учебное пособие (совместно с С.А. Боголюбовым, Ж.В.Прокофьевой и др.). - М.: Знание, 2000. - 288 с.
9. Управление качеством окружающей среды. Учебник. Т.1 (совместно с С.А.Боголюбовым и др.). - М.: МГИЭМ(ту), 2000. – 283 с.
10. Системы экологического управления: Учебник (совместно с С.А.Боголюбовым и др.). - М.: «Европейский центр по качеству», 2002. – 224 с.
11. Эконометрика. Учебник. – М.: Экзамен, 2002 (1-е изд.), 2003 (2-е изд.). – 576 с.

12. Управление промышленной и экологической безопасностью: Учебное пособие (совместно с В.Н. Федосеевым, В.Г. Ларионовым, А.Ф. Козьяковым). - М.: УРАО, 2002 (1-е изд.), 2003 (2-е изд.). – 220 с.
13. Менеджмент в техносфере. Учебное пособие (совместно с В.Н. Федосеевым). – М.: Академия, 2003. – 384 с.
14. Теория принятия решений. – М.: Издательство «Экзамен», 2004 (в печати).