ACADEMIA
Москва
Издательский центр «Академия»
2007

DUNCAN CRAMER

Advanced Quantitative Data Analysis

Open University Press
Maidenhead · Philadelphia

ДУНКАН КРАМЕР

Математическая обработка данных в социальных науках: современные методы

Перевод с английского

Рекомендовано Советом по психологии
УМО по классическому университетскому образованию
в качестве учебного пособия для студентов высших учебных заведений,
обучающихся по направлению и специальностям психологии

ACADEMIA

Москва
Издательский центр «Академия»
2007

УДК 519.221.25(075.8)
ББК 22.172я73
К777

Рецензенты:
доктор техн. наук *Л. С. Куравский*, зав. кафедрой «Прикладная информатика» Московского городского психолого-педагогического университета;
доктор филос. наук, канд. физ.-мат. наук *А. Н. Кричевец*, вед. науч. сотр. факультета психологии Московского государственного университета им. М. В. Ломоносова

Крамер Д.

К777 Математическая обработка данных в социальных науках : современные методы : учеб. пособие для студ. высших учеб. заведений / Дункан Крамер; пер. с англ. И. В. Тимофеева, Я. И. Киселевой; науч. ред. О. В. Митина. — М. : Издательский центр «Академия», 2007. — 288 с.

ISBN 978-5-7695-2878-1

В пособии американского психолога и математика Дункана Крамера рассматриваются методы статистической обработки данных, применяемые в современных социальных исследованиях. Не останавливаясь на базовых понятиях и критериях, известных из любого начального курса статистики, автор пытается доступным языком, без сложных формул и расчетов, объяснить принципы применения факторного и кластерного анализа, линейной и логистической регрессии, анализа путей, дисперсионного и ковариационного анализа, дискриминантного и, наконец, логлинейного анализа.

Пособие снабжено предисловием и комментариями научного редактора, списками рекомендуемой литературы на русском и английском языках, глоссарием статистических терминов и приложением.

Для студентов высших учебных заведений, изучающих методы математической обработки в социальных и гуманитарных науках, а также для преподавателей и исследователей-практиков.

УДК 519.221.25(075.8)
ББК 22.172я73

ISBN 978-5-7695-2878-1

DUNCAN CRAMER

Advanced Quantitative Data Analysis

Open University Press
Maidenhead · Philadelphia

ОГЛАВЛЕНИЕ

Часть VII

АНАЛИЗ ТАБЛИЦ ЧАСТОТ С ТРЕМЯ ИЛИ БОЛЕЕ КАЧЕСТВЕННЫМИ ПЕРЕМЕННЫМИ

ПРЕДИСЛОВИЕ НАУЧНОГО РЕДАКТОРА

Издание книги на русском языке, в которой описывается применение современных статистических методов анализа социальных и гуманитарных данных во всем их разнообразии, было и ожидаемо, и необходимо. Ни для кого не секрет, что Россия значительно отстает от стран Западной Европы и США в области приложения количественных методов. В этих странах подобных учебников, адресованных читателям самого различного уровня подготовки (от студентов колледжей, еще не имеющих даже степени бакалавра, до докторантов, избравших прикладную статистику и измерения в гуманитарных науках своей основной специализацией), издается великое множество, однако соответствующих переводов на русский язык не было с тех пор, как прекратилось издание серии «Библиотечка иностранных книг для экономистов и статистиков». За последнюю четверть XX в. за рубежом достигнуты большие успехи в области статистики с точки зрения как развития науки (появились новые методы статистического анализа), так и программного обеспечения (все больше методов реализуется в широко используемых статистических пакетах), а также в области методологии (понимания того, какие методы и в каких случаях использовать). Западные исследователи активно применяют статистические методы разной степени сложности, о которых российские ученые узнают либо из публикуемых статей, либо из выступлений на конференциях.

За последние годы было издано достаточно много монографий и учебных пособий, написанных отечественными авторами. Но они охватывают в основном простейшие и самые распространенные способы статистического анализа: описательную статистику, сопоставление двух выборок, а все выходящее за рамки этого обязательного минимума рассматривается только в ознакомительном формате (А. Д. Наследов, 2004; А. П. Кулаичев, 2006) либо посвящено отдельным методам (А. Н. Гусев, 2000; О. В. Митина, И. Б. Михайловская, 2001). В результате большое количество методов, уже широко известных и активно используемых коллегами за рубежом, позволяющих проверять более интересные и дифференцированные гипотезы, до настоящего времени известны узкому кругу исследователей, применяются крайне редко (например, дискриминантный анализ, анализ ковариаций) или не используются вообще (логлинейный анализ).

Структурное моделирование, оформившееся как методология работы с данными в США и странах Западной Европы в конце 70-х — начале 80-х годов XX в. и по сути органично включающее практически все линейные статистические методы — от определения простейших показателей до многомерного регрессионного и факторного анализа, получившего здесь естественное развитие и объединение (О. В. Митина, 2005), является, по нашему мнению, абсолютно необходимым в психологии и других гуманитарных дисциплинах, но только начинает применяться в России, поэтому учебно-методическое обеспечение особенно актуально.

Хотя статистический анализ имеет многовековую историю, по-настоящему оформление прикладной статистики как методологии работы с числовыми данными можно связывать с именем Ф. Гальтона, который в конце XIX в. впервые применил статистический анализ в биологии и психологии, ввел в психологию тесты и опросники (включая и сам термин «тест»), разработал близнецовый анализ. В 1888 г. ученый выступил с докладом на заседании Королевского общества «Корреляции и их измерение, преимущественно по антропометрическим данным».

Применение статистических методов во многом шло параллельно и взаимосвязанно с развитием метрических дисциплин как в естественных и инженерных (био-, хемометрика, метрология), так и в гуманитарных, социальных (измерения в психологии, социологии, экономике, истории) областях и способствовало их оформлению в науки с точки зрения требований строгости, доказательности, объективности, верифицируемости и пр. Статистика помогает доказывать различные гипотезы о проявлениях психологических свойств личности, целесообразности использования новых средств и методов лечения, причин и следствий тех или иных заболеваний и отклонений, предсказывать результаты политических выборов, проводить разработку месторождений полезных ископаемых, контролировать работу атомных станций и т.д. Как правило, проверяются гипотезы о количественных характеристиках различных параметров (переменных) и связи этих параметров друг с другом попарно или в более сложных конфигурациях. Затем возникает вопрос точности и обоснованности результатов. Насколько надежны результаты исследования? Достаточно ли большая выборка? Существуют ли подвыборки, в которых взаимосвязь между установленными переменными значимо различается? Насколько можно доверять полученным измерениям? На все эти вопросы можно ответить с помощью статистических методов. Широкое распространение компьютеров привело к тому, что возможность провести статистический анализ имеет практически каждый, однако, для того чтобы не сделать ошибочных выводов, необходимы соответствующие знания.

Одним из наиболее ранних примеров такого рода стал так называемый парадокс Симпсона (E. H. Simpson, 1951). Предположим, нам необходимо проверить эффективность определенного вида воздействия. Это могут быть какой-то новый вид психотерапевтического воздействия, медицинский препарат, обучение какому-то предмету в школе или в вузе по новой методике и т. д. На формальном уровне, для того чтобы зафиксировать наличие или отсутствие воздействия, вводится независимая дихотомическая переменная X, принимающая значения «0» при отсутствии воздействия и «1», если воздействие имело место. Необходимо оценить, влияет ли X на значения зависимой переменной Y, соответствующей успешности. Дихотомическая переменная $Y = 1$, если эффект был зафиксирован (успех достигнут), $Y = 0$ — при отсутствии эффекта (успеха). Согласно всем требованиям проведения подобного рода экспериментов, всех испытуемых разделяют на две равные группы: экспериментальную (подвергшуюся воздействию) и контрольную (воздействию не подвергавшуюся). В приведенной ниже таблице общее количественное распределение в контрольную и экспериментальную группы указано в столбцах, озаглавленных «Всего», строки содержат информацию о достигнутом результате (успехе). Ведь можно обучиться какому-либо предмету и по старым учебникам, решить свою психологическую проблему и без вмешательства психотерапевта и т. д.

Согласно данным, представленным в таблице, в эксперименте принимало участие 2000 испытуемых[1]. Они были поровну распределены в экспериментальную и контрольную группы в соответствии со стандартными требованиями экспериментального дизайна. Из 1000 человек, подвергшихся воздействию изучаемого фактора (попавших в экспериментальную группу), успех был зафиксиро-

Таблица. **Оценка успешности воздействия по всей выборке в целом и по каждому полу**

Успех	Количество испытуемых			Воздействие					
				Есть ($X=1$)			Нет ($X=0$)		
	Всего	Мужчины	Женщины	Всего	Мужчины	Женщины	Всего	Мужчины	Женщины
Да ($Y=1$)	1100	375	725	500	300	200	600	75	525
Нет ($Y=0$)	900	625	275	500	450	50	400	175	225
Всего	2000	1000	1000	1000	750	250	1000	250	750

[1] Все данные, содержащиеся в таблице, являются гипотетическими.

ван у 500 человек (у половины — 50 %), у испытуемых, входивших в контрольную группу, успех был зафиксирован у 600 человек (т. е. в 60 % случаях). Предположим теперь, что мы хотим проанализировать эффекты воздействия отдельно на подвыборках мужчин и женщин. И тех и других было поровну, а конкретные данные проведения эксперимента по каждой из этих подвыборок содержатся в столбцах, озаглавленных «мужчины» и «женщины» соответственно. Успех был зафиксирован у 300 из 750 мужчин, входивших в экспериментальную группу (40 %), и у 75 из 250 мужчин, входивших в контрольную группу (30 %). Аналогично достигли успеха 200 из 250 женщин, входивших в экспериментальную группу (80 %), и 525 из 750 женщин, входивших в контрольную группу (70 %). Таким образом, согласно полученным результатам, изучаемое воздействие оказывает положительный эффект на мужчин и женщин в отдельности, но не на выборку в целом. В итоге имеем парадокс[1].

Чтобы исключить ошибочные построения, научные сообщества вырабатывают стандарты представления результатов статистического анализа. И от рецензентов в научных журналах требуется оценивать предлагаемые публикации не только с точки зрения содержательной ценности представленных результатов, но и с точки зрения статистической грамотности их получения и обоснования.

В нашей стране при высоком уровне математической подготовки студентов и развитии математической науки в целом все, что связано с прикладными областями, исторически считалось чем-то малозначимым. Интересно в связи с этим процитировать письмо великого математика Н. Н. Лузина одному из наиболее талантливых своих учеников А. Н. Колмогорову, написанное в середине 20-х годов XX в.: «...Мое желание, чтобы Вы несколько удалились от работ по теории вероятностей. И вовсе не потому, что Ваш вклад в нее не фундаментален: я прекрасно знаю, что он оценивается всеми, как равноценный вкладу классиков. Но самое-то теория вероятностей не стоит Вас: ее источники сомнительные... и ее действие на работающих в ней не положительное. Вам дан высокий дух, и я хочу, чтобы Вы его силы берегли для вещей, которые под силу очень немногим». В то же время в традиции зарубежных ученых умение измерять и сопоставлять результаты полученных измерений еще в XIX в. рассматривалось как необходимый компонент общей грамотности и культуры. Приведем цитату из Карманной книги солдата того времени: «Подразумевается, что

---

[1] Было бы ошибкой думать, что подобного рода ситуации носят искусственный характер, чтобы их можно было встретить в реальном исследовании. Реальный пример, с которым столкнулись исследователи, выясняя, не отдается ли при отборе кандидатов в аспирантуру преимущество представителям одного пола перед другим, изложен в (P. I. Bickel, E. A. Hammel, I. W. O'Connell, 1985).

вы неплохо знаете арифметику; по крайней мере, две первые книги Евклида; планиметрию; алгебру вплоть до квадратных уравнений и фортификацию. Следует научиться с первого взгляда распознавать обычные разновидности растительного покрова, включая различные виды древесных пород. Для удобства измерения расстояний и т. п. каждому следует знать точную длину своего обычного шага и научиться точно отсчитывать шагами ярды; следует знать точную длину своей ступни, кисти, локтя, сабли, а также руки — расстояние от кончиков пальцев левой кисти до правого уха; следует знать высоту своего колена, талии и линии глаз, а также точное отношение объема своей питьевой кружки к пинте» (G. Wolseley, 1886).

Эти и ряд других обстоятельств привели к тому, что среди отечественных ученых сложилась традиция считать психологию сугубо гуманитарной наукой, а потому в определенном смысле даже несовместимой с количественным анализом, «приводящим к выхолащиванию принципов гуманности».

В последние годы российские психологи все интенсивнее сотрудничают с американскими и западноевропейскими коллегами, получают приглашения опубликовать результаты своих работ в иностранных журналах, но часто сталкиваются с проблемами, связанными с качеством выполнения и описания результатов количественного анализа данных. Для решения этих проблем, на наш взгляд, необходимо обеспечить доступ к определенной информации — удачно составленным лекционным курсам, качественным учебникам и монографиям; возможность присутствовать на докладах и мастер-классах, проводимых ведущими специалистами в этой области. Современная мобильность — возможность участия в зарубежных мероприятиях и Интернет-коммуникации — позволяет решить эти проблемы.

Наиболее образованными в области использования статистических методов оказались ученые-экономисты. И это неудивительно. Исторически сложилось так, что благодаря экономике у многих математиков появилась возможность стать лауреатами Нобелевской премии. Интегрирование математики в эту область науки способствовало интенсивному развитию эконометрики непосредственно (за счет собственных достижений) и косвенным образом (через повышение общего уровня математической грамотности ученых-экономистов). По мнению Нобелевского комитета, в настоящее время эконометрика применяется в качестве стандартного метода микроэкономики, изучающей все, начиная от расходов на ведение домашнего хозяйства и предпринимательских инвестиций и заканчивая организацией производств, рынков труда и эффектами государственной политики.

Россия в данном случае не является исключением. Статистическая и математическая грамотность эконометриков — достой-

ный пример для подражания. Госстандарт, утвержденный Министерством образования РФ для соответствующих специальностей, свидетельствует о возможности хорошей подготовки прикладных статистиков среди гуманитариев и социальных исследователей в нашей стране.

Отметим также, что, несмотря на лидерство статистики в области работы с количественными данными, это не единственный способ анализировать и репрезентировать результаты. Неправомерно сводить эконометрию, биометрию, клиометрию, психометрию и т.д. исключительно к статистике. В настоящее время методология гораздо шире и включает методы нелинейных динамических систем, нейронные сети, симуляционное моделирование и др. Однако даже в этом случае при построении моделей на основе функционального подхода исследователи используют данные предварительного статистического анализа, например аппроксимации или регрессии, прежде чем строить дифференциальную или разностную модель (О.В.Митина, В.Ф.Петренко, 2002). Более того, уместно напомнить, что даже математические физики, традиционно отдающие предпочтение аналитическим методам дифференциальных и интегральных уравнений, функционального анализа и т.п., в настоящее время предлагают учитывать флуктуационные эффекты, полноценно проанализированные с помощью статистических методов (В.И.Кляцкин, 2002). С учетом таких тенденций правомерно прогнозировать развитие количественных методов анализа и математического моделирования в противоположную сторону: от использования статистических методов к построению функциональных моделей. И в этом случае статистика будет играть роль не только иллюстратора результатов, но и своеобразного фундамента для построения дальнейших операциональных моделей, а значит, требования к уровню ее использования существенно возрастают.

Теория и практика анализа данных в той или иной науке, являясь одной из ее отраслей, занимают не какое-то обособленное место, а выполняют важную интегрирующую функцию. Использование сходного математического аппарата при решении исследовательских задач из разных сфер науки позволяет фиксировать их однотипность и тем самым помогает классифицировать исследовательские научные проблемы как интегральные, объединяющие в единый класс частные задачи, возникшие в различных отраслях, но по сути своей являющиеся проявлениями обобщенной латентной проблемы.

Здесь уместно вспомнить идею Л.С.Выготского (1982) о сравнении методологии со скелетом: внешним, наблюдаемым в простейших случаях, когда внутренности остаются мало дифференцированы и слабо детерминированы этим каркасом, и внутренним (являющимся опорой, костью каждого движения), и необхо-

димости различать низшие и высшие типы методологической организации. В настоящее время можно констатировать, что использование математического аппарата соответствует первому типу методологии, т.е. исследователь не имеет точного представления о том, какие методы, в каких случаях целесообразно применять, а исходит в первую очередь из того, что ему знакомо, привычно. Осознанное с этой точки зрения интегрирование математики в такой методологический каркас позволяет осуществить его преобразование из внешнего во внутренний. Можно предположить, что постановка проблемы и тип решаемой задачи существенным образом определяют выбор того или иного метода. Именно поэтому, по всей видимости, неправомерно ставить вопрос о создании учебника по всем методам анализа данных. Возможна лишь выработка основных стратегий, в рамках которой ученый должен проявить свой исследовательский талант, чувство «темы» и «данных». Однако для того чтобы такого рода творчество стало действительно доступным, необходимо в полном масштабе освоить технологию: совокупность приемов, алгоритмов и техник, используемых в той или иной науке на протяжении всей истории ее развития.

Итак, перейдем к анализу предлагаемой читателю книги. Собственный научный интерес автора связан с анализом психологических и педагогических данных. Преподавательскую деятельность он также ведет для студентов этих специальностей, чем и объясняется психологическое содержание рассматриваемых в книге примеров. Поэтому, безусловно, книга в первую очередь привлечет внимание ученых и аспирантов психолого-педагогического профиля; несмотря на то что требования к выполнению статистических процедур достаточно универсальны, тем не менее для каждой дисциплины существует своя специфика: степень строгости, допустимая погрешность, правила интерпретации и т.д. Однако и все остальные исследователи, чьи интересы связаны с анализом данных, найдут в ней много полезного для себя. Точно так же целесообразным является изучение психологами соответствующих книг для экономистов, медиков или биологов.

В каждой главе разбирается хотя и искусственный (сокращенный), но все же достаточно содержательный пример, что делает процедуру интерпретации достаточно осмысленной и стимулирует читателя не только проделать вслед за автором все вычисления, но, быть может, повторить их и на других переменных из этого примера, представляющих содержательный интерес. Все расчеты автор эксплицирует, и это методически очень полезно. Конечно, в настоящее время никто не выполняет вычисления ни на бумаге, ни даже с помощью калькулятора, однако при изучении того или иного метода целесообразно хоть раз проделать все выкладки «вручную», чтобы не относиться к компьютерной программе как к чер-

ному ящику (со страхом и недоверием или, наоборот, возлагая слишком большие надежды).

Необходимо заметить, что применение количественных методов сродни искусству и наивно было бы ожидать, что прочтя одну или даже несколько книг, вы обретете желаемую степень уверенности при работе. Главное — это большая практика. Именно тогда вы обнаруживаете «подводные камни», которые не указаны ни в одной книге — авторы даже порой не рефлексируют их. Поэтому не огорчайтесь, если, просчитав за автором весь пример, а потом выполнив все процедуры на компьютере, вы все еще не обрели желаемой уверенности. Главное — сохранить мотивацию продолжать освоение этой области.

Второй очень правильный методический прием (помимо включения в объяснения всех численных расчетов) — последовательное и подробное описание диалога работы с компьютером. Приведение в книге пошагово всех интерфейсных окон позволяет читателю без особых проблем воспроизвести все шаги самостоятельно и делает этот процесс воспроизводимым даже для читателей самого начального уровня компьютерной грамотности (не секрет, что среди психологов и педагогов таких немало).

Базовой программой, с помощью которой проанализировано большинство примеров, является SPSS. Это действительно наиболее популярная среди западных психологов и педагогов программа, и ее полезно освоить. При этом работать лучше с нерусифицированной версией. Недостатки русификации англоязычных статистических программ связаны с отсутствием устоявшейся статистической терминологии (в силу недостаточной развитости статистики в нашей стране). Поэтому переводчики порой проявляют фантастическую изобретательность, переводя то или иное слово, но при этом увеличивают хаос в сознании пользователя, не имеющего профессиональной математической и статистической подготовки. Отсюда возникает путаница при переводе, например, таких терминов: «part correlation» и «partial correlation». Или человек с удивлением осознает, что «ящик с усами» и «boxplot» — это одно и то же, или обнаруживает, что «индекс пригодности» обозначает надежность альфа-Кронбаха.

Одну «нишу» с SPSS (для исследователей и аспирантов, специализирующихся в гуманитарных и социальных науках) занимает программа STATISTICA. Это более молодая, но быстро развиваемая программа, имеющая множество дополнительных программных блоков, существенно расширяющих ее возможности. Большей частью они не востребованы отечественными учеными (например, алгоритмы нейронных сетей), но в перспективе владение программой STATISTICA может оказаться очень полезным.

Для читателей, имеющих начальный уровень компьютерного и статистического образования, полезно будет в качестве дополни-

тельной точки опоры использование программы STADIA, созданной в России (автор А. П. Кулаичев), а потому во многом учитывающей специфику отечественной аудитории (хорошо продуманный с методической точки зрения, дружественный интерфейс легко осваивается студентами-первокурсниками, встроенный Help достаточно информативен и понятен и содержит статистические рекомендации). Кроме того, автором STADIA недавно выпущен учебник (А. П. Кулаичев, 2006).

Для реализации конфирматорного факторного анализа и путевого анализа, являющихся частными случаями структурного моделирования, в книге используется еще одна программа — LISREL, узконаправленная для анализа структурных моделей. Эта программа не является распространенной в России, однако для примера, рассматриваемого в гл. 2, достаточно возможностей демонстрационной версии, бесплатно загружаемой прямо с сайта. На наш взгляд, кроме LISREL существуют более дружественные программы, с помощью которых выполнять структурное моделирование, в частности конфирматорный факторный анализ, оказывается намного легче. Примером такой программы является EQS (P. M. Bentler, 1996; О. В. Митина, 2005). Также у SPSS есть дополнительная надстройка AMOS, но она не входит в стандартный пакет и ее нужно приобретать дополнительно, в то время как в программе STATISTICA этот блок устанавливается по умолчанию (Л. Я. Дорфман, А. В. Огородников, 2005).

После каждой главы приведен список литературы, рекомендуемый автором, а также редактором перевода. Предлагаемые списки не претендуют на полноту, главное, что указанные в них книги еще не стали библиографической редкостью и их можно найти в книжных магазинах (год издания — не ранее 2000 г.). Правила применения той или иной процедуры можно найти и в справочной информации (Help) используемых программ SPSS, STATISTICA, STADIA, а также на их веб-сайтах.

О. В. Митина

ПРЕДИСЛОВИЕ РЕДАКТОРА СЕРИИ

Серия «Постижение методов социальных исследований» ставит своей задачей помочь студентам познакомиться с тем, как проводятся исследования в области общественных наук, и разобраться в различных проблемах методологии исследований в данной области. Книги этой серии адресованы студентам, аспирантам и начинающим исследователям, в программы обучения которых входят разделы, посвященные методам исследований в науках о человеке и обществе, будущим социологам, исследователям социальной политики, психологам, культурологам, демографам, политологам, криминалистам, а также тем, кто специализируется в области социального взаимодействия, организационных исследований и т. п. Эти книги будут полезны при проведении исследований в рамках курсовых работ, дипломных проектов, диссертаций.

Книги серии помогут читателям «постичь» проблемы и методы исследований в области общественных наук, что означает развитие умения ценить те радости и огорчения, с которыми связано любое исследование в этой области, выработать навыки использования определенных методов обработки данных и знаний основных проблемных моментов в обсуждаемых областях. Относительный акцент на том или ином из этих аспектов меняется от книги к книге, но цель каждой из них — осветить конкретный метод или проблему с позиций практического исследователя, а не представить просто сборник «пошаговых» инструкций. Чтобы достичь этой цели, серия включает освещение основных методов социальных исследований и рассмотрение большого числа проблем и спорных вопросов. Каждая книга серии написана практическим исследователем, имеющим опыт использования рассматриваемых методов или решения обсуждаемых проблем и вопросов. Таким образом, авторы опираются на свои практические знания и личный опыт.

Несмотря на то что существует множество книг, посвященных основам статистического анализа данных, сравнительно небольшое их количество в доступной форме рассматривает более изысканные виды и аспекты анализа. Именно это удалось сделать Дункану Крамеру в предлагаемой вниманию читателя книге. Он опирается на опыт чтения лекций и проведения семинаров, а также

написания книг, освещающих основные методы статистического анализа данных. Его подход состоит в том, чтобы на тщательно разобранных примерах познакомить начинающего студента или исследователя с рассматриваемыми методами. Более того, он связывает изложение статистических методов с объяснением того, какие шаги должны предпринять читатели, чтобы реализовать эти методы, используя компьютерные программы. Насколько это возможно, Д. Крамер уделяет внимание применению SPSS — наиболее широко используемого психологами пакета программ для статистического анализа. Кроме того, проведен разбор большей части результатов, выдаваемых рассматриваемыми в книге компьютерными программами, и указаны моменты, на которые необходимо обращать внимание и как интерпретировать полученные результаты.

Дункан Крамер подробно знакомит читателя с такими широко используемыми в западной экспериментальной психологии методами, как множественная регрессия, логлинейный анализ, логистическая регрессия и дисперсионный анализ. Знание этих методов принципиально важно, если исследователь хочет выйти за рамки простых манипуляций с данными и традиционной описательной статистики. Более того, изучение тех или иных методов анализа требует также знаний того, как их применять на практике, а в наше время это во многом сводится к умению использовать соответствующее программное обеспечение для реализации представленных методов. Именно в этом особая ценность книги Д.Крамера, уделяющего большое внимание изучению компьютерных программ. В то же время исследователь должен знать, как интерпретировать полученные в ходе статистической обработки данных результаты, и в этом отношении книга окажет неоценимую помощь в объяснении того, как правильно читать и понимать файлы, содержащие результаты работы компьютерных программ.

Эта книга очень своевременна, учитывая тот факт, что студентов всячески поощряют приобретать навыки, которые можно передать другим, а высшие учебные заведения, в свою очередь, призваны формировать такие умения у студентов. Знание того, как осуществлять более сложные и тонкие виды статистического анализа данных и как использовать программное обеспечение в связи с таким анализом, является важным компонентом внедрения подобных навыков в практику. В этом качестве данная книга окажет неоценимую помощь студентам, исследователям и преподавателям высших учебных заведений.

Алан Брайман

ПРЕДИСЛОВИЕ

Цель этой книги состоит в том, чтобы служить введением в некоторые из основных статистических методов, которые, однако, нельзя назвать абсолютно общеизвестными и повсеместно используемыми в социальных науках и психологии для анализа количественных данных, и сделать это насколько возможно проще и доступнее, без привлечения технически сложного математического аппарата. Развитие компьютерных программ количественного анализа данных, относительно простых в применении, привело не только к широкому распространению этих методов, но и к стремлению, когда это возможно, использовать данные методы для анализа собственных результатов. Чтобы понимать результаты количественного анализа, представленные в публикуемых статьях, и быть в состоянии критически оценить их, необходимо знать сами методы количественного анализа, ибо авторы статей обычно предполагают такое понимание со стороны читателя само собой разумеющимся, описывая только результаты своей работы. Несмотря на то что использование количественных методов имеет достаточно долгую историю, книг, в которых эти методы излагаются относительно просто и доступно для не очень подготовленного с точки зрения математики читателя, мало. Автор выражает надежду, что эта книга поможет восполнить данный пробел.

В книге показано конкретное использование определенных методов статистического анализа. Каждая глава книги начинается с общего описания назначения того или иного метода, а затем приводится простой пример-иллюстрация с небольшим набором данных. В целом используется восемь различных наборов данных. Они не велики по объему, включают от 9 до 15 наблюдений и от двух до девяти переменных. Поэтому с ними сравнительно легко работать. Вначале рассматриваются наиболее важные для понимания разбираемого метода аспекты статистики. Хотя в данной книге анализируются более изысканные статистические методы, все статистические термины объясняются. Автор надеется, что это поможет читателям, недостаточно хорошо знакомым со статистикой.

Чтобы не перегружать читателя запоминанием множества символов, вместо них использованы термины. По возможности, про-

водятся необходимые статистические расчеты, чтобы показать, как получены численные величины, а сами вычисления, чтобы отделить их от основного текста, размещены в таблицах. Также, где только возможно, приводится краткое изложение результатов каждого рассматриваемого примера. Разные издательства требуют различного стиля изложения полученных статистических результатов. В данной книге все примеры изложены в одном стиле. Но изменить их в соответствии с требованиями редактора не составляет труда. В конце каждой главы приведена рекомендуемая литература, включающая наименее технически сложные источники, и тем не менее их понимание требует более высокого уровня подготовки, чем тот, на который ориентирована данная книга.

Несмотря на то что более глубокое понимание статистического метода часто достигается путем проведения некоторых сопутствующих расчетов, мы не предполагаем и даже не рекомендуем читателям анализировать данные, непосредственно проводя такие вычисления. Эти вычисления эффективнее и приятнее выполнить, используя статистические компьютерные программы. В настоящее время в наличии имеется несколько различных широко распространенных программных продуктов. Автор использовал для выполнения вычислений в рассматриваемых примерах везде, где только возможно, пакет SPSS, полагая именно его наиболее популярным и доступным программным средством. Последней версией, выпущенной к моменту завершения рукописи, была версия пакета с номером 11. Эта версия подобна трем предыдущим версиям программы. Таким образом, данная часть книги также доступна тем, кто использует более ранние версии пакета. Пакет SPSS не содержит модуля структурного моделирования[1]. Для выполнения этих расчетов был выбран LISREL, поскольку он является одной из наиболее широко используемых программ для подобных вычислений[2]. Применялась последняя версия этой программы, которой на тот момент являлась LISREL 8.51. Рассматриваемый в книге пример может быть выполнен с помощью свободно распространяемой студенческой или ограниченной версий этой программы, которые можно загрузить со следующего вебсайта: http://www.ssicentral.com/other/entry.htm. Инструкции о том, как загрузить данные программы, доступны на этом сайте. Чтобы обозначить тот факт, что используемые термины и величины относятся к пакетам SPSS и LISREL или являются частью их выходных файлов, они выделены в тексте жирным шрифтом. Автор старался

---

[1] Здесь автор не прав, ибо специальный модуль в рамках SPSS существует и называется AMOS (здесь и далее примечания научного редактора).

[2] На наш взгляд, использовать LISREL достаточно сложно, с этой точки зрения заслуживает внимания программа EQS с более дружественным интерфейсом.

по возможности сократить описания программ, с тем чтобы книга в максимальной степени была полезна читателям, которые пользуются другим программным обеспечением.

Данные для разбираемых примеров были подобраны таким образом, чтобы проиллюстрировать важные статистические моменты, и не претендуют на получение результатов, типичных для исследовательской литературы по данной теме. При составлении примеров мы также старались упростить их. Если эти примеры покажутся вам недостаточно интересными или содержательными, то можно по аналогии создать свои собственные. Самостоятельное построение и анализ примеров полезны для проверки общего понимания рассматриваемых методов. Это понимание можно углубить, изучая опубликованные отчеты с примерами подобного анализа в своей области исследований. Анализ количественных данных — умение, которое только выигрывает от должным образом осмысленной практики. Автор надеется, что данная книга поможет читателю развить этот навык.

Хотелось бы выразить свою благодарность Алану Брайману, Тиму Ляо и Аманде Сакер за их отзывы и комментарии к первому проекту рукописи.

Дункан Крамер,
Университет Лохборо

ВВЕДЕНИЕ[1]

Одна из главных задач социальных наук и психологии состоит в объяснении различных аспектов человеческого поведения. Например, нас может интересовать объяснение того, почему одни люди более агрессивны, чем другие. Один из способов определения адекватности или действительности объяснений состоит в том, чтобы собрать данные, характеризующие изучаемые характеристики людей, и найти, до какой степени эти данные совместимы с предлагаемыми объяснениями. Данные, согласующиеся с объяснением, поддерживают его в той степени, в какой они противоречат другим объяснениям. Данные, противоречащие объяснению, являются доказательными лишь настолько, насколько корректно операционализированы признаки, имеющие отношение к рассматриваемому явлению. Операционализация включает управление признаками или их измерение.

Качественные и количественные переменные

Как следует из названия, переменная должна представлять собой изменяющуюся особенность или характеристику изучаемого объекта или явления. Если характеристика не изменяется, ее называют *постоянной*. Для психологии и общественных наук представляет интерес объяснение того, почему изменяются те или иные признаки. Самый простой тип переменной — *бинарный*, в котором данное качество либо присутствует, либо отсутствует. Такая характеристика, как пол, представляет собой бинарную переменную, значения которой в каждом конкретном случае определяются полом объекта — женским или неженским (мужским). Аналогично, бинарной является переменная «быть в разводе». Обе категории, составляющие бинарную переменную, могут быть представлены (закодированы) любой парой чисел, таких, как 1 и 2

[1] Во введении уделяется внимание переменным и измерениям, с помощью которых исследователи получают данные для своего анализа. Специфика того или иного способа сбора данных определяет, какие математические и статистические действия с ними можно выполнять.

или 23 и 71. Например, «быть женщиной» может быть закодировано как 1, а «быть мужчиной» как 2.

Различают два типа переменных. Переменные первого типа называют по-разному: качественными, категориальными, номинальными, или частотными. Примером качественной переменной может служить семейное положение, которое включает пять категорий: (1) холост/не замужем; (2) женат/замужем; (3) проживает отдельно; (4) разведен(а); (5) вдовец/вдова. Эти пять категорий могут быть представлены или закодированы любым набором из пяти чисел: 1, 2, 3, 4 и 5 или 32, 12, 15, 25 и 31. Данные числа просто используются для обозначения различных категорий. Можно подсчитать только число, или частоту, объектов, относящихся к каждой из данных категорий, поэтому такие переменные часто называют *частотными переменными*. Например, из 100 человек 30, возможно, никогда не были женаты, 40 могут состоять в браке, 8 — проживать раздельно, 12 — быть в разводе и у 10 человек супруги скончались. Частота случаев в категории может быть выражена как доля или процент от полной частоты случаев. Так, доля людей, состоящих в браке, равна 0,40 (40/100 = 0,40). Эта же величина, выраженная в процентах, равна 40 ($40/100 \times 100\,\% = 40\,\%$). Данные, состоящие из качественных переменных, являются количественными в том смысле, что частота, доля или процент случаев могут быть определены количественно. Категории качественной переменной могут рассматриваться как бинарные переменные. Например, разведенные могут представлять одну категорию бинарной переменной, а оставшиеся четыре группы из никогда не состоявших в браке, состоящих в браке, живущих раздельно и вдовых — другую категорию. К качественным переменным можно отнести также вид потребляемой пищи, страну происхождения, характер заболевания и метод получаемого лечения.

Другой тип переменных называют *количественными переменными*. Числовые значения в этом случае используются, чтобы отобразить и (или) упорядочить уровни увеличения значений этих переменных. Самый простейший пример количественной переменной — бинарная переменная, такая как пол, где одна из категорий интерпретируется как представляющая большее количество данного качества по сравнению с другой. Например, если женщины закодированы как 1, а мужчины как 2, то эта переменная может рассматриваться как отражение маскулинности, и большее значение указывает на бóльшую выраженность данного качества. Следующий простой пример: переменная социального класса, которая включает три уровня значений — высший, средний и низший. Высшее сословие может быть закодировано как 1, средний класс как 2 и низший класс как 3; в данном случае меньшие значения соответствуют более

высокому социальному статусу. Эти числа можно рассматривать как величины, принадлежащие шкале отношений. Значение 1 соответствует в два раза более высокому рангу по сравнению с 2, что дает отношение 1 к 2[1]. Как правило, количественные переменные включают более трех категорий, к их числу относятся возраст, доход или суммарный балл по шкале какого-нибудь опросника. Например, для оценки агрессивности человека может использоваться опросник, в который входят десять вопросов. Для каждого вопроса предусмотрено два варианта ответа «Да» или «Нет». Значение 1 можно присвоить ответам, которые указывают на агрессивность, в то время как значение 0 можно приписать ответам, которые показывают отсутствие агрессивности. Сложив вместе баллы для каждого из десяти вопросов, получим суммарный балл, изменяющийся в пределах от 0 (минимальное значение) до 10 баллов (максимальное значение). Более высокие значения будут означать бóльшую агрессивность. Правомерно считать, что эти числа представляют собой шкалу отношений. Испытуемый, набравший 10 баллов, имеет в два раза больший балл по сравнению с тем, кто набрал 5 баллов по данной шкале (10 : 5 = 2 : 1).

При необходимости всегда можно объединить смежные категории, чтобы сформировать меньшее число групп значений, однако количество вновь образованных категорий не должно быть меньше двух. Например, 11 категорий только что упомянутого полного набора значений агрессивности могут быть повторно сгруппированы в три новые категории, объединяющие значения от 0 до 1, от 2 до 5 и от 6 до 10. Вновь образованные категории не обязаны включать равное количество значений, и именно такая ситуация имеет место в нашем примере, где первая категория состоит из двух значений (0 и 1), вторая включает четыре значения (2, 3, 4 и 5), а третья объединяет пять значений (6, 7, 8, 9 и 10). Эти три новые категории будут теперь иметь новые числовые обозначения: 1 — для первой группы, 2 — для второй группы и 3 — для третьей группы. Однако такая перегруппировка должна быть обоснована. В рассмотренном случае смысл новых категорий менее ясен, чем исходных, а диапазон значений меньше.

В социальных науках и психологии нас обычно интересует вопрос: связана ли переменная с одной или несколькими другими переменными? Чем сильнее зависимость между переменными, тем больше между ними общего. Двумерный анализ исследует отношения между двумя переменными, в то время как много-

---

[1] На наш взгляд, это не совсем верно, ибо крайне сложно определить, как один статус может быть в ДВА раза выше, чем другой. Система показателей оценки таких характеристик в социальных науках является приближенной.

мерный[1] анализ исследует отношения между тремя или более переменными одновременно. В книге рассматривается только один случай анализа связи между двумя переменными — однофакторный дисперсионный анализ (см. гл. 9). Все другие примеры касаются трех или более переменных одновременно. В однофакторном дисперсионном анализе рассматриваются отношения между качественной переменной «семейное положение» и количественной переменной «уровень выраженности депрессии», измеряемой в баллах по специальной шкале. Однофакторный анализ ковариаций исследует эту зависимость, контролируя уровень значений второй количественной переменной в зависимости от значения первой. Другими словами, он включает три переменные, одна из которых качественная (семейное положение), а две другие — выраженность депрессии и взаимосвязь между семейным положением и депрессией — количественные. Следовательно, этот тип анализа также является многопеременным. Любой аспект поведения человека находится под влиянием нескольких различных факторов. Поэтому анализ, учитывающий их одновременно, будет способствовать лучшему пониманию исследуемых феноменов.

Статистический вывод

Очень часто перед исследователем возникает проблема: как определить, можно ли факт определенной взаимосвязи переменных, обнаруженный на данных, полученных из выборки, распространить на всю генеральную совокупность, из которой осуществлялась выборка. В этом контексте понятие генеральной совокупности относится к значениям переменных, а не к людям или другим организмам. Выборка представляет собой подмножество таких значений. Чтобы установить, можно ли факт, полученный на выборке, распространить на генеральную совокупность, необходимо вычислить вероятность его обнаружения в ситуации, обусловленной случайными событиями. Если вероятность такого обнаружения $\leq 0{,}05$ (1 шанс на 20 или меньше), то можно считать, что данный факт закономерен: присутствует в генеральной сово-

[1] В английском языке используется термин «multivariate», который по сложившейся традиции переводится термином «многомерный», хотя, на наш взгляд, лучше использовать именно термин «многопеременный», оставив термин «мерность» для обозначения размерности структуры анализируемых данных. Плоская таблица является двумерной независимо от того, сколько переменных она содержит, а данные, имеющие структуру куба (в случае, когда каждый респондент заполняет двумерную матрицу, или лонгитюдные наблюдения, требующие от одних и тех же респондентов неоднократного ответа по нескольким пунктам фиксированного опросника), являются трехмерными.

купности, из которой осуществлялась выборка. Другими словами, данная взаимосвязь вряд ли случайна. Такая взаимосвязь называется *статистически значимой*. Однако возможно, что мы как раз столкнулись с редким случаем (вероятность его наступления < 0,05) а именно, связь есть, но она обусловлена случайными событиями. В этой ситуации вывод о существовании закономерности связи будет сделан, в то время как на самом деле ее нет. Такая ошибка называется *ошибкой первого рода*[1]. Хотя за общепринятый уровень статистической значимости принимается величина 0,05 или менее, следует помнить о произвольности подобного соглашения.

Если вероятность установления связи в ситуации, обусловленной случайными причинами, превосходит 0,05, считаем, что на генеральной совокупности, из которой осуществлялась выборка, исследуемой закономерности нет. Такая взаимосвязь называется *статистически незначимой*. Однако возможно, что исследуемая взаимосвязь, вероятность случайного обнаружения которой превышает 0,05, носит все же неслучайный характер и на самом деле закономерна. В этом случае, отвергая имеющуюся в действительности закономерность на основании того, что она с достаточно большой вероятностью (более 0,05) может быть обусловлена случайными событиями, мы также совершаем ошибку. Такая ошибка известна как *ошибка второго рода*[2].

Заметим, что чем больше объем выборки, тем с большей вероятностью можно утверждать, что установленная связь обусловлена неслучайными событиями. Другими словами, вероятность совершить ошибку первого рода (утверждать существование взаимосвязи, когда на самом деле ее нет) увеличивается с ростом выборки. Вероятность обнаружения статистически незначимой взаимосвязи тем больше, чем меньше объем выборки. Другими словами, вероятность совершить ошибку второго рода (утверждать отсутствие взаимосвязи, когда на самом деле она существует) увеличивается с уменьшением выборки. Более сильные взаимосвязи имеют бóльшую вероятность оказаться статистически значимыми. Следовательно, при интерпретации статистической значимости необходимо принять во внимание и величину взаимосвязи, и объем

---

[1] Говорить, что ошибка первого рода — это неправильно сделанный вывод о том, что исследуемая связь закономерна, в то время как она обусловлена случайными событиями, было бы не совсем точно. В ситуации статистического вывода проверяется так называемая гипотеза H_0, которая может утверждать закономерность существования связи между какими-то данными или, наоборот, закономерность отсутствия этой связи и т.д. В дополнение к гипотезе H_0 формулируется альтернативная гипотеза H_1, заключающаяся в отрицании H_0. В терминах гипотез ошибка первого рода — ошибочное отвергание нулевой гипотезы (H_0), когда она на самом деле верна.

[2] В терминах гипотез ошибка второго рода означает принятие нулевой гипотезы в то время, когда она ошибочна.

выборки. Из всех статистических методов, описанных в этой книге, процедура вычисления статистической значимости отсутствует только в двух случаях: кластерном и эксплораторном факторном анализе[1].

Зависимые и независимые переменные

Для многих из описываемых статистических методов необходимо различать зависимые и независимые переменные. В некоторых случаях — множественной регрессии, дисперсионном факторном анализе — зависимую переменную можно назвать откликом, или переменной-откликом, а независимую переменную — предиктором, или переменной-предиктором. Зависимая переменная, или переменная-отклик, — переменная, поведение которой мы стремимся объяснить в терминах независимых переменных, или переменных-предикторов. Зависимая переменная называется так, потому что мы предполагаем, что она находится под влиянием, или зависит от независимых переменных. Предполагается, что независимые переменные не испытывают влияния, т.е. независимы, от других переменных.

В анализе путей (иногда мы будем употреблять синонимический термин «путевой анализ»), обсуждаемом в гл. 6 и 7, будет рассмотрена такая последовательность переменных, когда первая переменная влияет на вторую, вторая — на третью и т.д. Переменную, с которой начинается последовательность, иногда называют *внешней*, или экзогенной[2], переменной, потому что она является внешней по отношению к модели, рассматриваемой в путевом анализе. Переменные следующих за внешними переменными уровней называют *внутренними*, или эндогенными[3], переменными, потому что модель путевого анализа должна их объяснить. В то время как экзогенная переменная является независимой переменной, эндогенные переменные выступают в роли как независимых, так и зависимых переменных. Они зависят от предшествующей и влияют на следующую за ними переменную. Также возможно, что две или более переменных влияют друг на друга. В этом случае их зависимость называют взаимной. Данная тема не освещается в настоящей книге.

Следует ясно понимать, что, используя статистический анализ, можно только установить, связаны ли переменные друг с другом, но нельзя определить, влияет ли одна переменная на

---

[1] Собственно эта ситуация верна не только для выборки примеров, включенных в книгу, но и отражает общую картину статистических методов.

[2] От англ. exogenous — внешний.

[3] От англ. endogenous — внутренний.

другую. Например, можно найти зависимость между уровнем безработицы и преступности, при которой те, кто не имеет оплачиваемой работы, с большей вероятностью вовлечены в преступную деятельность или осуждены за нее. Это соотношение, однако, не означает, что безработица ведет к преступной деятельности. Одинаково правдоподобно и то, что те, кто вовлечен в преступную деятельность, меньше интересуются поиском оплачиваемой работы. Также возможно, что обе переменные влияют друг на друга. Определение причинной природы или характера (направления) зависимости двух переменных в социальных науках и психологии обычно представляет собой сложную проблему, которая предусматривает необходимость веских оснований для хорошо аргументированной точки зрения. Роль статистического анализа для выработки подобной аргументации состоит в том, чтобы предложить показатель величины любой наблюдаемой взаимосвязи и дать оценку вероятности случайного появления такой зависимости.

Выбор адекватного статистического метода

Здесь будет дан краткий обзор статистических методов, рассматриваемых в настоящей книге, чтобы помочь читателям, которые не очень хорошо представляют, какой из методов наиболее адекватен для анализа их данных, и, в первую очередь, заинтересованы в том, чтобы получить информацию только о таком методе. Методы в книге были упорядочены таким образом, чтобы стало ясным, как анализировать потенциально большой набор количественных переменных и выделять статистические идеи, используемые в дальнейшем. Например, множественная регрессия рассматривается перед дисперсионным анализом потому, что в ходе ее выполнения решаются в том числе и задачи анализа вариаций (дисперсий), т. е. дисперсионного анализа. При выборе оптимального метода анализа тех или иных данных необходимо учитывать тип переменных, к которым этот метод может быть применен.

В настоящей книге представлен только один метод, позволяющий одновременно рассматривать три или более качественные переменные. Это — логлинейный анализ, описанный в гл. 13. Например, нас может интересовать соотношение между психическим расстройством (указанным по классификации тревога, депрессия, тревога и депрессия одновременно), религиозной ориентацией (отсутствует, протестант, католик) и детским семейным статусом (жил с обоими родителями, только с матерью, только с отцом). Логлинейный анализ используется, чтобы ответить на два связанных между собой типа вопросов. Первый вопрос: отличается

ли статистически значимо частота интересующих исследователя наблюдений от случайной, в терминах взаимодействия между тремя или более качественными переменными? Другими словами, должны ли мы рассматривать более двух переменных для объяснения распределения интересующих нас наблюдений в зависимости от этих переменных? Второй вопрос: какие из переменных и/или их взаимодействия необходимы для объяснения распределения наблюдений? Этот вопрос отличается от первого тем, что в данном случае наряду с влиянием отдельных переменных и их двусторонних взаимодействий с другими переменными рассматриваются и взаимодействия более высокого порядка.

Если одну из качественных переменных необходимо рассматривать как зависимую (например, классифицированные психические расстройства), а другие качественные переменные как независимые, то логистическая регрессия является наиболее адекватным методом, поскольку она рассматривает только взаимосвязи между зависимой переменной и независимыми переменными. Другими словами, она исключает из рассмотрения взаимосвязи независимых переменных между собой (например, отношения между религиозной ориентацией и детским семейным статусом). Единственным типом логистической регрессии, рассматриваемым в настоящей книге, является бинарная логистическая регрессия, где зависимая переменная состоит из двух категорий; она описана в гл. 8. Множественная логистическая, или логит-регрессия, используется, чтобы определить, какие качественные и количественные переменные и их взаимодействия наиболее сильно связаны с вероятностью осуществления определенной категории зависимой переменной, при этом учитываются их связи с другими независимыми переменными, участвующими в анализе. Качественные переменные должны быть преобразованы в фиктивные двоичные переменные. Эта процедура описана в гл. 9—11 для дисперсионного анализа.

Существует три метода введения предикторов в логистическую регрессию. В стандартном, или прямом, методе все предикторы вводятся одновременно, хотя некоторые из этих независимых переменных могут играть незначительную роль в максимизации вероятности реализации категории. В иерархическом, или последовательном, методе независимые переменные вводятся в заранее определенном порядке с тем, чтобы можно было выяснить, какой вклад в максимизацию вероятности реализации рассматриваемой категории они вносят. Например, демографические переменные, такие, как возраст, пол, социальный статус, могут вводиться в первую очередь для того, чтобы их влияние можно было зафиксировать на следующем этапе. В статистическом, или пошаговом, методе среди предикторов выбираются те переменные, которые вносят наибольший вклад в максимизацию веро-

ятности реализации категории. Если два предиктора связаны друг с другом и имеют очень близкие вклады в максимизацию вероятности осуществления рассматриваемой категории, то будет выбран предиктор с большей величиной вклада в максимизацию категории, даже если различие в величине коэффициентов максимизации у этих двух предикторов минимально.

Дискриминантный анализ, описанный в гл. 12, может использоваться для определения той количественной переменной, которая лучше всего предсказывает, в какую категорию попадет объект, при условии, что данные отвечают следующим требованиям. Число объектов в категориях зависимой переменной не должно сильно различаться. Независимые переменные должны иметь нормальное распределение, а внутригрупповые дисперсии — быть одинаковыми[1]. Независимые переменные порождают новую составную переменную, называемую дискриминантной функцией. Максимальное число рассматриваемых дискриминантных функций не должно превышать число предикторов, с одной стороны, и быть, как минимум, на единицу меньше числа групп — с другой. Как и в случае с логистической регрессией, существует три метода включения предикторов в дискриминантную функцию. В стандартном, или прямом, методе все предикторы вводятся одновременно, хотя некоторые из этих независимых переменных, возможно, и не позволяют выявить различия между группами. В иерархическом, или последовательном, методе предикторы группами вводятся в заранее определенном порядке с тем, чтобы можно было выяснить, какой вклад в дискриминацию (различение между группами) они вносят. В статистическом, или пошаговом, методе среди предикторов по одному выбираются те переменные, которые вносят наибольший вклад в дискриминантную функцию. Если два предиктора связаны друг с другом и их вклады в дискриминантную функцию почти равны, то будет выбран предиктор, у которого этот вклад больше, даже если различие минимально.

Другие статистические методы, описанные в этой книге, предназначены для тех случаев, когда либо зависимая переменная, либо все переменные вместе являются количественными. Множественная регрессия используется, чтобы определить, какие количественные и качественные независимые переменные и их взаимодействия наиболее сильно связаны с количественной переменной-откликом. Качественные переменные необходимо рассматривать как фиктивные переменные, как это описано в гл. 9—11 для дисперсионного анализа. Так же, как и в случае с логистической регрессией и дискриминантным анализом, существует три метода

---

[1] Это предположение называется предположением об однородности дисперсий (или их гомогенности — от англ. homogeneous).

включения предикторов в множественную регрессию. В стандартном, или прямом, методе все предикторы вводятся одновременно, хотя некоторые из этих независимых переменных могут и не быть связаны с переменной-откликом. В статистическом, или пошаговом, методе в качестве предикторов выбираются те независимые переменные, которые вносят наибольший вклад в дисперсию зависимой переменной. В том случае, когда два предиктора связаны (коррелируют) друг с другом и имеют очень близкие коэффициенты связи с переменной-откликом, выбирают предиктор, имеющий больший коэффициент связи с зависимой переменной, даже если различие между соответствующими сравниваемыми коэффициентами минимально. Этот метод описан в гл. 4. В иерархическом, или последовательном, методе независимые переменные вводятся в заранее определенном порядке с тем, чтобы можно было выяснить, какой вклад они вносят. Этот метод описан в гл. 5. Иерархическая множественная регрессия также используется в основном варианте путевого анализа, в дисперсионном и ковариационном анализе.

Анализ путей применяют, чтобы определить силу связи в гипотетической последовательности, или ряду, количественных эндогенных (экзогенных) переменных и ту степень, в которой выбранные пути обеспечивают удовлетворительное описание или величину критерия согласия между всеми переменными. Качественные переменные могут быть включены в анализ как экзогенные переменные, когда они преобразованы в фиктивные двоичные переменные, как это описано в гл. 9—11. В самом простом случае рассмотрения трех переменных анализ путей может использоваться, чтобы оценить, в какой степени одна из переменных является прямой функцией двух других и косвенной функцией одной из них. Самая простая форма анализа путей рассматривается в гл. 6. Более сложная форма анализа путей, учитывающая надежность переменных, описана в гл. 7.

Дисперсионный анализ используется, чтобы определить, связаны ли значимо одна или несколько качественных независимых переменных и их взаимодействия с количественным откликом или зависимой переменной. Если качественная переменная состоит только из двух категорий (является бинарной), значимая связь означает, что средние значения откликов в двух группах, принадлежность к которым определяется по значению качественной переменной, значимо различаются. Если качественная переменная включает более двух категорий, значимая связь подразумевает, что средние значения откликов двух или более групп (соответствующих различным категориям независимой качественной переменной) значимо различаются. Если имеются веские основания для того, чтобы прогнозировать, какие из этих средних различаются, значимость этих различий может быть опреде-

лена с использованием одностороннего критерия Стъюдента. Если никаких различий не прогнозировалось или не было веских причин предполагать какие-либо различия, то значимость различий необходимо анализировать с помощью апостеорных критериев (post-hoc, т.е. возникающих на основании работы с данными, а не в результате выдвинутых предварительных теоретических соображений, например критерий Шеффе). Дисперсионный анализ с одной качественной переменной описан в гл. 9, а дисперсионный анализ с двумя качественными переменными — в гл. 11. Ковариационный анализ позволяет фиксировать влияние независимых количественных переменных, связанных с количественной переменной-откликом. В гл. 10 рассматривается анализ ковариаций, в котором участвуют две независимые переменные (одна качественная и одна количественная) и одна зависимая количественная переменная.

Наконец, статистические методы, описываемые в гл. 1—3, позволяют определить возможности группировки связанных количественных переменных в меньшее число объемлющих их факторов или кластеров. Например, нас может интересовать, можно ли сгруппировать пункты опросника, измеряющие тревогу и депрессию, соответственно, в два фактора или кластера, представляющие эти два типа вопросов. Разновидность эксплораторного[1] (разведочного) факторного анализа, называемая методом главных компонент, описана в гл. 1. Такой анализ является разведочным в том смысле, что способ возможной группировки переменных не предопределен заранее, как это имеет место в случае конфирматорного[2] (подтверждающего) факторного анализа, описанного в гл. 2. Конфирматорный факторный анализ дает статистическую меру для определения того, насколько удовлетворительно заранее определенная структура группировки наблюдаемых признаков объясняет реально существующие связи (вычисленные по экспериментальным данным) между ними. Другим методом группировки переменных является кластерный анализ. Разновидность кластерного анализа, называемая иерархической агломеративной кластеризацией, описана в гл. 3.

Следует отметить, что одни методы, описанные в этой книге, встречаются в литературе по социальным наукам и психологии чаще, другие — реже. К наименее популярным методам относятся кластерный анализ, логлинейный анализ и дискриминантный анализ. Следовательно, читатели с меньшей вероятностью встретятся с этими методами при чтении публикаций с использованием статистического анализа данных. Поэтому, чтобы лучше познакомиться с более редкими способами примене-

[1] От англ. explore — исследовать.
[2] От англ. confirm — подтверждать.

ния данных методов, имеет смысл попытаться найти примеры использования этих методов, осуществляя специальный поиск в электронных библиографических базах данных, относящихся к области ваших собственных научных интересов[1].

Дополнительная литература, рекомендуемая научным редактором

Кричевец А. Н. Математика для психологов / А. Н. Кричевец, Е. В. Шикин, А. Г. Дьячков. — М.: Флинта: Московский психолого-социальный институт, 2003.

Сидоренко Е. В. Методы математической обработки данных в психологии. — СПб. : Речь, 2003.

Гусев А. Н. Измерение в психологии / А. Н. Гусев, Ч. А. Измайлов, М. Б. Михалевская. — М.: изд-во «Смысл», 2000.

Тюрин Ю. Н. Анализ данных на компьютере / Ю. Н. Тюрин, А. А. Макаров. — М.: Инфра-М, 2003.

[1] На наш взгляд, кластерный анализ в отечественной литературе встречается несравнимо чаще. Пример дискриминантного анализа см.: О. В. Митина, В. Ф. Петренко, 2002, а вот примеры использования логлинейного анализа нам не удалось найти вообще.

ЧАСТЬ I
ГРУППИРОВКА КОЛИЧЕСТВЕННЫХ ПЕРЕМЕННЫХ

Глава 1
ЭКСПЛОРАТОРНЫЙ (РАЗВЕДОЧНЫЙ) ФАКТОРНЫЙ АНАЛИЗ

Предисловие научного редактора

Статистические методы, описываемые в ч. I (гл. 1—3), позволяют определить возможности группировки связанных количественных переменных в меньшее число объемлющих их факторов или кластеров.

В гл. 1 описывается метод главных компонент. Этот метод часто рассматривают как одну из наиболее распространенных (установленных в большинстве компьютерных программ по умолчанию) разновидностей эксплораторного (исследовательского, разведочного) факторного анализа. Хотя с математической точки зрения эти два метода существенно отличаются друг от друга: в задачу первого входит объединение исходных признаков (шкал) в минимально возможное число классов с сохранением максимально возможной информации, задаваемой этими переменными, а для второго целью является максимально приближенное воспроизведение матрицы взаимосвязей между переменными, с точки зрения содержательной (для предметной интерпретации) большой роли это не играет — оба метода используются для группировки первичных переменных с установлением весовых коэффициентов, определяющих степень включенности каждой из них в ту или иную группу. Выделить главные компоненты проще (меньше требований к эмпирическим соотношениям, определяемым первичными переменными). Решение, получаемое в результате выполнения именно факторного анализа, надежнее и устойчивее, однако условия, накладываемые на переменные, более жесткие. Поэтому для построения эвристических моделей чаще всего используют главные компоненты, в то время как для построения опросников — инструментов, связанных с практическим применением для получения психодиагностических показателей, предпочтительнее использовать более надежный и математически более корректный факторный анализ.

Факторный анализ представляет собой совокупность методов, призванных определить, насколько связанные (коррелирующие) переменные могут быть сгруппированы так, чтобы

каждую группу можно было рассматривать как одну составную переменную, или фактор, а не как ряд отдельных переменных. Возможно, наиболее распространенное применение факторного анализа в социальных науках и психологии состоит в том, чтобы определить, можно ли объединить совокупность пунктов опросника, оценивающих конкретную характеристику, таким образом, чтобы получить общий индикатор данной характеристики.

Например, нас может интересовать оценка восприятия людьми собственной тревожности. Мы могли бы просто спросить некоторых людей, насколько тревожными они обычно бывают. Однако существуют три главные проблемы при попытке измерить социальную характеристику, индивидное качество, личностную черту и т.д. с помощью единственного вопроса или пункта опросника. Во-первых, потенциальная чувствительность такого показателя будет существенно ограничена. Например, если мы ограничим возможные ответы на задаваемый вопрос только вариантами «Да» или «Нет», то сможем распределить наших респондентов лишь по двум категориям. Чем больше вопросов о тревожности мы зададим, тем больше возможных категорий респондентов в зависимости от данных ими ответов на все вопросы в совокупности можно получить. Так, на основании ответов на два вопроса можно составить четыре категории (вариантов ответов)[1], три вопроса — шесть категорий и т.д.

Вторая проблема с измерением, основанным на ответе на единственный вопрос, состоит в том, что в этой ситуации невозможно определить, насколько надежен показатель. Так, может оказаться, что люди, которых мы спрашиваем, не знают, что значит «испытывать тревожность», и отвечают «Да» или «Нет» в большей степени случайным образом без четкого понимания смысла вопроса. Чтобы определить надежность этого вопроса, можно задать его два или более раз в рамках одного и того же интервью.

Если бы данный вопрос являлся надежным показателем, мы могли бы предполагать, что испытуемые каждый раз дадут на него один и тот же ответ. Поскольку попытка задать один и тот же вопрос несколько раз в течение короткого промежутка времени может быть воспринята как факт недостаточной организованности интервьюеров или их недоверия к интервьюируемым, предпочтительнее задавать вопросы, которые различаются по форме, но сохраняют свою содержательную направленность. Например, можно спросить респондентов: напряжены ли они обычно?

---

[1] Для двух вопросов варианты ответов могут быть следующие: Да-Да, Нет-Да, Да-Нет, Нет-Нет. Читателю может быть полезно выписать все возможные варианты ответов на три вопроса.

Третья проблема, связанная с измерением, построенном на одном пункте опросника, состоит в том, что оно не позволяет исследовать различные аспекты данной характеристики. Например, «тревожиться», «быть в напряжении», «нервничать» или «легко пугаться» может описывать несколько различные аспекты состояния, которое мы называем тревогой. Если это действительно различные проявления тревоги, то можно ожидать, что те, кто описал себя как тревожащихся, также опишут себя как испытывающих напряжение, нервничающих или пугливых. Аналогично, можно предполагать, что те, кто отвечал, что не испытывает тревогу, также скажут, что не испытывают напряжение, мало нервничают и не боятся. Другими словами, мы предполагаем, что ответы на эти четыре вопроса связаны друг с другом и образуют единый фактор. Если дело обстоит именно таким образом, мы могли бы объединить вместе ответы на эти четыре вопроса и создать единый суммарный показатель, а не рассматривать их как четыре различных отдельных показателя, дающих сходную информацию.

Однако интерпретация результатов факторного анализа, опирающаяся только на соображения о том, что какие-то пункты вопросника образуют единый фактор, который можно было бы назвать тревожностью, проблематична. Если, например, обнаружено, что ответы на четыре пункта вопросника, связанных с тревогой, группируются и образуют единый фактор, то без дополнительной информации нельзя узнать, в самом ли деле этот фактор специфичен для тревоги и отражает ее уровень или представляет более общий фактор, который можно было бы назвать склонностью к жалобам. Или же, если эти четыре пункта вопросника объединяются в два отдельных фактора, например, в фактор тревоги (напряженности), с одной стороны, и нервозности (пугливости) — с другой, нельзя узнать, расщепился ли общий фактор тревожности на два более специфичных подфактора. Следовательно, проводя факторный анализ, полезно включать ответы на те пункты, которые, как предполагается, не относятся к исследуемой (проверяемой) характеристике (состоянию). Например, если мы считаем, что депрессия и тревога представляют собой отдельные состояния, то, чтобы лучше интерпретировать результаты, можно включить в анализ ответы на вопросы, связанные с депрессией. Если бы тревога и депрессия группировались в единый фактор, это означало бы, что респонденты, испытывающие тревогу, также испытывают депрессию, и эти характеристики нельзя развести, по крайней мере, на основании самоотчета. Если же пункты, связанные с тревогой, группировались бы в один фактор, а пункты, связанные с депрессией, — в другой, мы могли бы быть более уверены в том, что наши вопросы, связанные с тревогой, являются не просто мерой общей склонности респондентов чувствовать себя несчастными.

Проиллюстрируем интерпретацию и некоторые вычисления, связанные с факторным анализом, пытаясь установить, можно ли выделить тревогу и депрессию, получаемые по результатам самоотчета, в отдельные факторы. Чтобы не усложнять пример, ограничимся тремя короткими вопросами по тревоге (Anxiety, A1 — A3) и тремя — по депрессии (Depression, D1 — D3), соответственно, хотя в большинстве случаев использования факторного анализа оперируют бóльшим числом переменных:

A1 Я испытываю тревогу
A2 Я становлюсь напряженным
A3 Я спокоен
D1Я подавлен
D2Я чувствую себя бесполезным
D3Я счастлив

Ответы на каждый из этих вопросов даются по 5-балльной шкале, где 1 означает «никогда», 2 — «иногда», 3 — «часто», 4 — «бóльшую часть времени» и 5 — «всегда».

Рекомендуемый минимальный объем выборки, которую можно использовать для проведения факторного анализа, по-разному определяется разными авторами, но общепринято, что он должен быть больше числа переменных. Например, R. L. Gorsuch (1983) полагает, что для проведения факторного анализа необходимо не менее 100 испытуемых (наблюдений) и, как минимум, 5 наблюдений в расчете на переменную. Наблюдение представляет собой единицу анализа, которой в нашем случае является респондент. Однако это может быть школа, коммерческая организация, город и т. п. Чтобы упростить процедуру ввода данных для анализа, мы использовали вымышленные ответы только девяти испытуемых, приведенные в табл. 1.1[1]. В строке, соответствующей первому испытуемому, видим, что он иногда испытывает тревогу, никогда не напряжен и часто спокоен[2].

Корреляционная матрица

Первым шагом при осуществлении факторного анализа является создание корреляционной матрицы, в которой содержатся коэффициенты корреляции всех переменных друг с другом. Корреляционная матрица для данных табл. 1.1 представлена в табл. 1.2.

---

[1] В соответствии с идеей R. L. Gorsuch (1983) наблюдений должно быть существенно больше.

[2] Данные, содержащиеся в табл. 1.1, называют сырыми, или первичными.

Таблица 1.1. **Ответы девяти испытуемых по шести переменным (вопросам)**

Наблюдения	A1 Тревожный	A2 Напряженный	A3 Спокойный	D1 Подавленный	D2 Бесполезный	D3 Счастливый
1	2	1	3	1	2	5
2	1	2	3	4	3	3
3	3	3	4	2	1	4
4	4	4	3	3	2	3
5	5	5	2	3	4	4
6	4	5	2	4	3	1
7	4	3	2	5	4	1
8	3	3	4	4	4	3
9	3	5	3	3	4	1

Таблица 1.2. **Нижний треугольник корреляционной матрицы для шести переменных**

Переменные	A1 Тревожный	A2 Напряженный	A3 Спокойный	D1 Подавленный	D2 Бесполезный	D3 Счастливый
A1 Тревожный	1,00					
A2 Напряженный	0,74	1,00				
A3 Спокойный	–0,50	–0,40	1,00			
D1 Подавленный	0,22	0,30	–0,37	1,00		
D2 Бесполезный	0,28	0,39	–0,43	0,65	1,00	
D3 Счастливый	–0,25	–0,54	0,41	–0,74	–0,53	1,00

Такая матрица называется *нижнетреугольной* из-за своей формы, при которой корреляции между каждой парой переменных показаны лишь один раз.[1]

[1] Полная матрица получается в результате симметричного отображения чисел нижнего треугольника относительно главной диагонали.

Корреляция отражает направление и величину линейной зависимости между двумя переменными. Коэффициент корреляции принимает значения в диапазоне от –1,00 до +1,00. Значение коэффициента корреляции, равное –1,00, указывает на максимальную обратную зависимость между переменными, при которой наибольшее значение одной переменной (например, 5 — для «Тревожный») связано с наименьшим значением другой переменной (1 — для «Спокойный»), следующее по величине значение первой переменной (4 — для «Тревожный») связано со следующим из наименьших значений второй переменной (2 — для «Спокойный»), и т.д. Коэффициент корреляции, равный +1,00, указывает на максимальную прямую зависимость между двумя переменными, при которой максимальное значение одной переменной (например, 5 — для «Тревожный») связано с максимальным значением другой переменной (например, 5 — для «Напряженный»), следующее по величине наибольшее значение первой переменной (4 — для «Тревожный») связано со следующим по величине наибольшим значением второй переменной (4 — для «Напряженный»), и т.д. Значение коэффициента корреляции, равное 0,00, указывает на отсутствие линейной зависимости между двумя переменными. Обычно максимальные по модулю значения коэффициента корреляции +1,00 или –1,00 очень редко встречаются. Коэффициенты корреляции, стоящие на диагонали матрицы, приведенной в табл. 1.2, представляют собой просто корреляции переменных с самими собой и по своей сути всегда равны 1,00, а потому не представляют никакого интереса и никакой роли не играют.

Чем больше по абсолютному значению коэффициент корреляции, независимо от знака, тем сильнее линейная взаимосвязь между двумя переменными. Наибольшие по абсолютному значению коэффициенты корреляции в табл. 1.2 равны 0,74 (корреляция между «Тревожный» и «Напряженный») и –0,74 (корреляция между «Подавленный» и «Счастливый»). Следующая по абсолютному значению корреляция между «Подавленный» и «Бесполезный» равна 0,65. Наименьший по абсолютному значению коэффициент корреляции между «Тревожный» и «Подавленный» равен 0,22.

Величина совместной дисперсии для двух переменных есть квадрат коэффициента корреляции этих переменных. Так, величина совместной дисперсии, объясняемой тем общим, что присутствует и в ощущении тревожности и в ощущении напряженности, равна $0{,}74^2$, или приближенно 0,55, в то время как величина общей дисперсии между тревожностью и подавленностью равна $0{,}22^2$, или около 0,05. Другими словами, величина совместной дисперсии между тревожностью и напряженностью в 11 раз больше величины совместной дисперсии между тревожностью и подавленностью. Максимальное значение величины совместной дис-

персии равно 1,00 ($\pm 1{,}00^2 = 1{,}00$), а минимальное — 0,00 ($0{,}00^2 = 0{,}00$).

Если посмотреть на значения коэффициентов корреляции в табл. 1.2, можно увидеть, что имеется некоторая тенденция, проявляющаяся в том, что пункты вопросника, связанные с тревогой, сильнее коррелируют друг с другом, чем с пунктами, связанными с депрессией, а последние, в свою очередь, сильнее коррелируют между собой, чем с пунктами, связанными с тревогой. Таким образом, можно предположить существование двух отдельных групп для пунктов, связанных с тревогой и депрессией. Например, коэффициент корреляции между ощущением тревожности и напряженностью равен 0,74, в то время как коэффициент корреляции между ощущением тревожности и подавленностью равен всего лишь 0,22. Однако картина не совсем ясна, поскольку абсолютное значение коэффициента корреляции между «Напряженный» и «Счастливый», входящих в группу депрессивных пунктов, выше (–0,54), чем абсолютное значение коэффициента корреляции между «Напряженный» и «Спокойный» (–0,40). Обычно невозможно сказать, просто глядя на корреляционную матрицу, на сколько групп, или факторов, разобьются переменные; чем больше переменных, тем труднее это сделать. Следовательно, чтобы определить, сколько групп получится, надо использовать более строгий формальный метод. Таковым является эксплораторный факторный анализ.

Метод главных компонент

Существует много различных видов факторного анализа, но, возможно, самым простым и наиболее часто используемым среди них является метод главных компонент. Компоненты — еще один термин для обозначения факторов, и компоненты в методе главных компонент часто называются факторами. В этой книге будем использовать данные термины как взаимозаменяемые[1]. В методе главных компонент величина объясняемой дисперсии равна числу переменных, поскольку дисперсия или общность каждой пере-

[1] Нам кажется, следующий комментарий может облегчить понимание сути эксплораторного факторного анализа. Каждая переменная имеет определенную дисперсию (общность). В методе главных компонент переменные стандартизируются и нормируются, поэтому предполагается, что дисперсия каждой из них равна 1. Исходя из их независимости можно предположить, что суммарная дисперсия всех переменных равна числу переменных. Каждая главная компонента является в определенном смысле переменной величиной, а потому также имеет дисперсию, называемую объясняемой дисперсией. Дисперсия отсутствует только у постоянных величин. Суммарная объясняемая дисперсия всех главных компонент должна быть равна суммарной дисперсии всех переменных, т.е. их общему числу.

менной принимается равной 1,00. Таким образом, в случае шести переменных суммарная (общая) объясняемая дисперсия равна 6,00. Число образованных, или извлекаемых, компонент математически всегда равно числу переменных, участвующих в анализе[1]. Таким образом, в случае шести переменных будет извлечено шесть факторов. Факторы всегда упорядочиваются в соответствии с величиной их дисперсии. Первый фактор всегда объясняет наибольшую долю общей дисперсии, второй фактор — следующую по величине долю общей дисперсии, которая не была объяснена первым фактором, и т.д., последний фактор объясняет наименьшую долю общей дисперсии. Значения коэффициентов корреляции переменной с каждым фактором дают величины нагрузок данной переменной по этим факторам[2]. Так как первый фактор объясняет наибольшую долю общей дисперсии, значения коэффициентов корреляции, или нагрузки, всех переменных по этому фактору будут, в среднем, самыми высокими, следующими по величине будут нагрузки на второй фактор и т.д. Чтобы вычислить долю общей дисперсии, объясняемой каждым фактором, надо суммировать квадраты нагрузок по данному фактору, в результате получаем собственное, или характеристическое, значение этого фактора, которое делим на число переменных.

В табл. 1.3 приведены шесть главных компонент для набора данных из шести переменных, представленных в табл. 1.1. Как можно видеть, все шесть переменных сильнее всего коррелируют с первым фактором, за исключением ощущения тревожности, которое чуть сильнее коррелирует со вторым фактором. В табл. 1.4 приведены доли общей дисперсии, объясненной этими шестью главными компонентами, для чего были вычислены квадраты нагрузок, затем суммированием этих квадратов были получены собственные значения и, наконец, путем деления на число переменных — доли общей дисперсии. Так, нагрузка тревожности по первому фактору, равная 0,66, при возведении в квадрат и округлении до двух десятичных знаков дает 0,43. Суммирование квадратов нагрузок переменных по первому фактору дает собственное значение, рав-

---

[1] Только так будет соблюдено условие о том, что сумма суммарных дисперсий факторов в точности равна сумме дисперсий первичных переменных.

[2] Такая интерпретация коэффициентов корреляции переменных с факторами правомерна, если полагать каждый фактор как некоторую гипотетическую переменную, которую нельзя реально измерить. Эти переменные называются латентными, подробно о них будет рассказано далее, здесь следует упомянуть, что факторы, будучи переменными, хотя и гипотетическими, могут коррелировать с другими реальными переменными, и именно эти гипотетические коэффициенты корреляции и называются факторными нагрузками. Сумма квадратов факторных нагрузок всех переменных по фактору (сумма совместных дисперсий) равна объясняемой этим фактором дисперсии, умноженной на число первичных переменных.

Таблица 1.3. **Начальные главные компоненты**

	1	2	3	4	5	6
A1 Тревожный	0,66	0,67	0,03	0,11	0,29	–0,14
A2 Напряженный	0,76	0,46	0,37	0,02	–0,20	0,18
A3 Спокойный	–0,69	–0,20	0,64	0,25	0,10	–0,05
D1 Подавленный	0,76	–0,53	0,05	–0,04	0,35	0,14
D2 Бесполезный	0,75	–0,34	–0,15	0,52	–0,17	–0,06
D3 Счастливый	–0,80	0,34	–0,27	0,35	0,14	0,17

Таблица 1.4. **Доля общей дисперсии, объясняемая начальными главными компонентами**

	1	2	3	4	5	6	Общности
A1 Тревожный	0,43	0,45	0,00	0,01	0,08	0,02	1,00
A2 Напряженный	0,58	0,21	0,14	0,00	0,04	0,03	1,00
A3 Спокойный	0,48	0,04	0,41	0,06	0,01	0,00	1,00
D1 Подавленный	0,57	0,28	0,00	0,00	0,12	0,02	1,00
D2 Бесполезный	0,56	0,11	0,02	0,27	0,03	0,00	1,00
D3 Счастливый	0,64	0,12	0,07	0,02	0,02	0,03	1,00
Собственные значения	3,26	1,21	0,64	0,47	0,31	0,11	6,00
Доля[1]	0,54	0,20	0,11	0,08	0,05	0,02	

ное 3,26, которое при делении на число переменных, равное шести, дает, с учетом округления, долю общей дисперсии, равную 0,54(3,26/6 = 0,543). Другими словами, первая главная компонента (фактор) объясняет 54 % общей дисперсии шести переменных, второй фактор — еще 20 % общей дисперсии и т.д.[2].

Определение числа главных компонент, оставляемых для дальнейшего анализа

Поскольку число выделяемых компонент (факторов) равно числу переменных, нам необходим некоторый критерий отбора и решения вопроса о том, сколькими меньшими факторами можно

[1] Объясняемой дисперсии.

[2] Читатель может попробовать подсчитать эту величину для того, чтобы убедиться, что основные принципы усвоены правильно.

пренебречь, учитывая незначительную долю объясняемой ими общей дисперсии. Одним из главных критериев, используемых для решения этого вопроса, является критерий Кайзера, или Кайзера — Гутмана, согласно которому не следует рассматривать факторы с собственными значениями, меньшими или равными единице. Обосновать это можно следующим образом. Наибольшая величина дисперсии, объясняемая одной переменной, равна единице, поэтому факторы с собственными значениями, меньшими единицы, могут хорошо объяснить, самое большее, дисперсию одной-единственной переменной. Как следует из табл. 1.4, первые два фактора имеют собственные значения, превосходящие единицу, в то время как собственные значения остальных четырех факторов меньше единицы. Таким образом, если мы примем этот широко используемый критерий, нам следует обращать внимание только на первые два фактора и игнорировать четыре оставшихся меньших фактора.

R. B. Cattell (1966) указал на то, что критерий Кайзера может сохранять слишком много факторов в случае большого числа переменных и слишком мало факторов в случае малого числа переменных. Он предложил альтернативный критерий, критерий «каменистой осыпи»[1], связанный, главным образом, с нахождением отчетливой границы между первыми бóльшими факторами, объясняющими значительную часть общей дисперсии, и следующими за ними меньшими факторами, объясняющими примерно равные по величине малые доли общей дисперсии. Чтобы определить, где проходит эта граница, собственные значения каждого фактора изображают точками на плоскости, при этом по вертикали откладывают собственные значения, а по горизонтали располагают номера факторов в порядке уменьшения величин соответствующих собственных значений. На рис. 1.1 таким образом изображены собственные значения шести главных компонент (факторов), приведенных в табл. 1.4.

«Каменистая осыпь» представляет собой геологический термин для обозначения обломков горных пород, скапливающихся у подножия крутого склона и скрывающих настоящее основание самого склона. Факторы, образующие сам склон, считаются важными для рассмотрения, а факторы, представляющие «каменистую осыпь», — нет. Последние можно определить путем проведения прямой линии либо через точки, представляющие соответствующие собственные значения, либо в непосредственной близости от них[2]. Это не всегда легко сделать, как на рис. 1.1, где не совсем ясно, начинается ли «осыпь» со второго или с третьего фактора.

---

[1] От англ. scree test. Часто этот критерий называют критерием следа.

[2] В результате число незначимых факторов определяется количеством точек, лежащих вблизи этой прямой линии (каменистой осыпи).

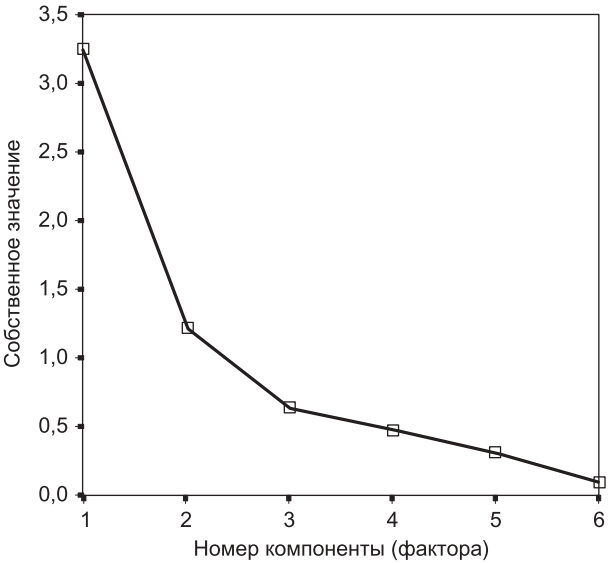

Рис. 1.1. График собственных значений для шести главных компонент

В этом случае может оказаться полезным сравнение переменных, коррелирующих как с первыми двумя, так и с первыми тремя факторами (после соответствующего вращения факторов), с целью определить, какое же из двух решений — двухфакторное или трехфакторное — имеет больший смысл[1]. Причины для осуществления вращения факторов будут обсуждаться в следующем подразделе. Если с помощью проведения прямых линий можно выделить несколько «осыпей», то заслуживающей максимального доверия считается самая верхняя из них.

Вращение факторов

Математическая процедура, позволяющая прояснить содержательный смысл выделенных на предыдущем этапе начальных глав-

[1] Выбор числа значимых факторов, основанный на сравнении решений с точки зрения их осмысленности, имеет существенный недостаток, связанный с отсутствием формальных статистических критериев, т.е. в этом случае исследователь должен выбрать модель исходя из своих субъективных предпочтений. Однако здесь стоит отметить, что даже в тех случаях, когда статистические критерии существуют, субъективность при выборе того или иного решения (гипотезы, модели) также присутствует. Например, выбор границы уровня значимости, т.е. допускаемой вероятности совершить ошибку первого рода, осуществляется исследователем и может быть 0,05; 0,1; 0,01 (а также любое другое число в диапазоне от 0 до 1).

Таблица 1.5. **Первые две главные компоненты после вращения по методу варимакс**

	1	2
A1 Тревожный	0,05	0,94
A2 Напряженный	0,27	0,85
A3 Спокойный	–0,39	–0,60
D1 Подавленный	0,92	0,10
D2 Бесполезный	0,79	0,24
D3 Счастливый	–0,83	–0,27

ных компонент, объясняющих бóльшую часть общей дисперсии переменных, называется *вращением*. Факторные нагрузки преобразуются по специальным формулам[1]. Существует множество методов таких преобразований (вращений). Мы обсудим два из них. Наиболее распространенным методом вращения является *варимакс*[2], при котором факторы остаются независимыми или ортогональными по отношению друг к другу, так что баллы испытуемых по одному из факторов не коррелируют с баллами по другим факторам. Метод варимакс пытается максимизировать дисперсию, объясняемую значимыми (отобранными на предыдущем этапе с помощью критерия Кайзера, следа или другим способом) факторами, путем еще большего увеличения коэффициентов корреляции (факторных нагрузок), высоко коррелирующих с этим фактором переменных, и уменьшения коэффициентов корреляции, низко коррелирующих с этим фактором переменных. В табл. 1.5 представлены результаты вращения двух выделенных главных компонент по методу варимакс.

Сравнение величин коэффициентов корреляции (факторных нагрузок) в табл. 1.5 и 1.3 показывает, что некоторые из коэффициентов корреляции увеличились, в то время как другие, наоборот, уменьшились, а факторная структура компонент после вращения стала более понятной и легче интерпретируемой. Первый из факторов после вращения варимакс можно интерпретировать как фактор депрессии, поскольку все три соответствующих пункта имеют по этому фактору нагрузки ±0,79 и выше, в то время как

[1] Эти преобразования факторов производятся по формулам таким образом, что если сопоставить старые и новые полученные после пересчета факторных нагрузок факторы в многомерном пространстве, координатами которого являются значения факторных нагрузок у этих факторов по всем первичным переменным, то геометрически это будет выглядеть как поворот вокруг начала координат.

[2] Стоит отметить, что вращение варимакс используется, как минимум, в 80 % случаев применения эксплораторного факторного анализа для анализа данных в конкретных эмпирических исследованиях.

Таблица 1.6. **Доли общей дисперсии, объясняемые первыми двумя компонентами после вращения по методу варимакс**

	1	2
A1 Тревожный	0,00	0,88
A2 Напряженный	0,07	0,72
A3 Спокойный	0,15	0,36
D1 Подавленный	0,84	0,01
D2 Бесполезный	0,62	0,06
D3 Счастливый	0,69	0,07
Собственные значения	2,37	2,10
Доля[1]	0,40	0,35

все три вопроса, касающиеся тревожности, имеют нагрузки ±0,39 и ниже. Второй фактор после вращения варимакс можно интерпретировать как фактор тревожности, поскольку пункты, соответствующие тревоге, имеют по этому фактору нагрузки ±0,60 и выше, в то время как пункты, связанные с депрессией, — ±0,27 и ниже. При анализе большого количества переменных полезно расположить все пункты в порядке убывания факторных нагрузок с тем, чтобы видеть, какие из переменных имеют самые высокие нагрузки по рассматриваемым факторам.

Доля общей дисперсии, объясняемая каждым из двух выделенных факторов, после вращения по методу варимакс равна их собственным значениям, или сумме квадратов факторных нагрузок для каждого фактора, деленной на число переменных. Эти данные для двух главных компонент после вращения варимакс представлены в табл. 1.6. Доля общей дисперсии, объясняемой первым фактором, равна 0,40, вторым фактором — 0,35. Эти доли объясняемой дисперсии отличаются от соответствующих величин для начальных, не подвергавшихся вращению, главных компонент, поскольку в результате вращения изменились нагрузки переменных на эти факторы.

Другой метод вращения называется *прямой облимин* (direct oblimin[2]), здесь факторы могут коррелировать, т.е. быть неортогональными по отношению друг к другу. Существует два способа представления результатов косоугольного вращения. Первый — это так называемая матрица факторного отображения[3], которая обычно приводится в выходных файлах, содержащих результаты

[1] Объясняемой дисперсии.

[2] От англ. oblique — в геометрии острый или тупой, наклонный, т.е. неортогональный (прямой).

[3] В SPSS — pattern matrix.

Таблица 1.7. **Матрицы факторного отображения и структурная для первых двух компонент после вращения по методу прямой облимин**

	Матрица факторного отображения		Структурная матрица	
	1	2	1	2
A1 Тревожный	–0,16	0,99	0,26	0,93
A2 Напряженный	0,09	0,85	0,45	0,88
A3 Спокойный	–0,28	–0,55	–0,51	–0,67
D1 Подавленный	0,97	–0,11	0,92	0,29
D2 Бесполезный	0,79	0,07	0,82	0,40
D3 Счастливый	–0,83	–0,09	–0,87	–0,44
Доля	0,39	0,34	0,47	0,42

анализа[1]. Матрица факторного отображения показывает, какой вклад каждая переменная вносит в уникальную часть того или иного фактора, и не учитывает вклада, делаемого переменной в совместную с другими факторами часть. Структурная матрица отражает общий вклад каждой переменной в соответствующий фактор[2]. Если факторы не коррелируют между собой, то эти матрицы совпадают, и в этом случае проще и корректнее осуществить вращение по методу варимакс. Если же факторы коррелируют между собой, не имеет смысла приводить доли общей дисперсии, объясняемой каждым фактором, поскольку матрица факторного отображения будет давать заниженную, а структурная матрица — завышенную оценки.

В табл. 1.7 приведены матрицы факторного отображения и структурная для двух главных компонент из табл. 1.4 после вращения по методу прямой облимин. Также приведены доли общей дисперсии, объясняемой этими факторами после вращения, с тем чтобы показать их различие, хотя обычно подобные показатели не приводятся.

---

[1] Поскольку в результате косоугольного вращения факторы коррелируют, то можно говорить о том, что дисперсия каждого фактора объясняется частично только этим фактором (уникальная часть), а также совместными корреляциями этого фактора с другими, т.е. каждый фактор можно рассматривать состоящим из двух слагаемых. Во-первых, это уникальная, или специфическая, часть фактора, во-вторых, та часть, которая совместна, т.е. входит в другие факторы тоже. В случае, когда вторая составляющая нулевая, корреляция фактора со всеми остальными факторами равна нулю.

[2] В структурной матрице учитывается не только непосредственная взаимосвязь переменной с фактором (как в матрице факторного отображения), но и опосредованная взаимосвязь факторов друг с другом. В результате получаются значения — коэффициенты корреляции переменных с факторами.

Коэффициент корреляции между двумя выделенными косоугольными факторами равен примерно 0,42. Так как данные факторы коррелируют, величины нагрузок в матрицах факторного отображения и структурной различаются между собой. Нагрузки из матрицы факторного отображения легче интерпретировать, чем нагрузки из структурной матрицы, хотя результаты в целом сходны. Первый фактор, подвергнутый вращению по методу прямого облимина, можно интерпретировать как фактор депрессии, поскольку в матрице факторного отображения все три пункта, связанные с депрессией, имеют по этому фактору нагрузки ±0,79 или выше, в то время как все три элемента, связанные с тревогой, — ±0,28 или ниже. Второй фактор после вращения по методу прямой облимин, видимо, представляет тревогу, так как все три элемента, связанные с тревогой, имеют по этому фактору в матрице факторного отображения нагрузки ±0,55 или выше, в то время как все три пункта, связанные с депрессией, — ±0,11 или меньше. В данном отношении результаты вращения по методу прямого облимина в основном совпадают с результатами вращения по методу варимакс.

Отчет о результатах

Форма отчета о проведении анализа по методу главных компонент в определенной степени зависит от целей его проведения. Один из кратких способов описать результаты примера, использованного в данной главе, состоит в следующем: «Корреляционная матрица шести переменных была подвергнута процедуре анализа по методу главных компонент. Было извлечено два фактора с собственными значениями больше единицы. Эти факторы подвергались вращению по методу варимакс и прямой облимин. Были получены существенно сходные результаты. Первый фактор можно интерпретировать как фактор, отражающий депрессию, так как все три пункта, связанные с депрессией, имеют по нему самые высокие нагрузки. Второй фактор можно интерпретировать как фактор, отражающий тревогу, поскольку все три пункта, соответствующие тревоге, имеют по нему высокие факторные нагрузки. Факторы, полученные в результате вращения по методу варимакс, объясняют 40 и 35 % совокупной (общей) дисперсии соответственно. Коэффициент корреляции между факторами, полученными в результате вращения direct oblimin (прямой облимин), составил 0,42».

Реализация процедуры в программе SPSS для Windows

Приведем алгоритм получения основной информации, связанной с анализом данных из табл. 1.1 по методу главных компонент.