VorlesungsScripte Statistik

H. Läuter
Universität Potsdam

1. Dezember 1997

Inhaltsverzeichnis

1 Vorbemerkungen **4**

2 Beschreibende Statistik **4**

2.1 Lageparameter 5

2.2 Streuungsmaße 7

2.3 Verteilung der Daten 8

3 Empirische Verteilungsfunktion und empirische Momente **11**

3.1 Gesetze der großen Zahlen 11

3.2 Empirische Verteilungsfunktionen 12

4 Statistische Modelle **18**

4.1 Binomialverteilungsannahme 18

4.2 Multinomialverteilung 20

4.3 Geometrische Verteilung 20

4.4 Normalverteilung 21

4.5 Lebensdauerverteilungen 22

5 Wichtige Verteilungen, die mit der multivariaten Normalverteilung zusammenhängen **25**

5.1 Mehrdimensionale Normalverteilung 26

5.2 χ^2-Verteilung 29

5.3 F-Verteilung 31

5.4 Studentsche t-Verteilung 32

6 Charakteristische Funktionen **32**

7 Entscheidungsfunktionen **43**

7.1 Entscheidungsprobleme 43

7.1.1 Testen von Hypothesen 43

7.1.2 Mehrentscheidungsverfahren 47

7.1.3 Punktschätzungen 47

7.2 Entscheidungsprinzipien 49

7.2.1 Maximum-Likelihood-Prinzip 49

7.2.2 Kleinste-Quadrate-Methode 53

7.2.3 Bayessches Prinzip 54

7.3 Konsistenz von Schätzern 55
7.4 Optimalität von Entscheidungsfunktionen 58

8 Bedingter Erwartungswert und suffiziente Statistiken 71

9 Regression und Varianzanalyse 84
9.1 Das lineare Regressionsmodell 84
9.1.1 Schätzung unter Restriktionen 86
9.1.2 Schätzen der Varianz 88
9.1.3 Testen von Hypothesen 90
9.1.4 Konfidenzbereiche 94
9.2 Varianzanalysemodelle 98
9.2.1 Kreuzklassifikationsmodelle 98
9.2.2 Hierarchische Effekte 111
9.2.3 Gemischte Klassifikationstypen 114
9.3 Multiple Mittelwertsvergleiche 115
9.3.1 1. Multipler t–Test 119
9.3.2 Der Scheffé–Test 119
9.3.3 Der Tukey–Test 121
9.3.4 Bonferroni-Tests 122

10 Literatur 124

1 Vorbemerkungen

Stochastik ist die Mathematik des Zufalls. Sie ist von größter Bedeutung für die Berufspraxis der Mathematiker. Die beiden Hauptgebiete der Stochastik sind Wahrscheinlichkeitstheorie und Statistik. In der Wahrscheinlichkeitstheorie untersucht man zufällige Prozesse mit festen als bekannt angenommenen steuernden Wahrscheinlichkeiten (theoretisch und praktisch von Interesse). In der Statistik werden aus Daten (Beobachtungen) Aussagen über unbekannte Wahrscheinlichkeiten abgeleitet. Stochastische Fragen treten auf in

- Strategien in Glücksspielen und anderen unsicheren Entscheidungsprozessen
- Wahrscheinlichkeitsaussagen für das Wachstum von Populationen und die Vererbung von Eigenschaften
- Vergleich von Behandlungen, von Heilmitteln oder von Produktionsverfahren mit vorgegebener Sicherheit
- Aussagen über die Genauigkeit von Messungen.

Zur Lösung dieser Problemstellungen muß eine mathematische Formulierung gefunden werden. Aufgabe der Statistik ist das Auffinden von Gesetzmäßigkeiten und Zusammenhängen. Die mathematische Statistik untersucht die Eigenschaften bzw. die Optimalität der Verfahren.

2 Beschreibende Statistik

Ausgangspunkt aller statistischen Analysen sind Daten

$$y_1, ..., y_n$$

Diese y_i können reelle Zahlen, natürliche Zahlen oder auch Bezeichnungen sein.
Beispiel: Bei n Patienten werden Blutwerte, Augenfarbe, Einstufung der Hörfähigkeit,...untersucht. y_i ist der Vektor aller gemessenen Werte beim i-ten Patienten:

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{ip} \end{pmatrix}.$$

Man unterscheidet
metrische Skalen
ordinale Skalen
nominale Skalen.

Dabei sind nominale Skalen derartige Werte, die Namen repräsentieren, ordinale Skalen können geordnet werden. Bei metrischen Skalen sind überdies die Abstände zwischen einzelnen Werten ausnutzbar.
Durch Methoden und Verfahren der beschreibenden Statistik werden die gewonnenen Daten geordnet und übersichtlich dargestellt. Dabei werden die Lage, die Streuung und Verteilung der Daten beschrieben.

2.1 Lageparameter

beschreiben die ungefähre Lage der Daten, z.B. durch Angabe des Mittelwertes. Der bekannteste Mittelwert ist

1. Das arithmetische Mittel

Seien $z_1, ..., z_n$ reelle Zahlen.

$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i$ heißt arithmetisches Mittel.

Kommen unter den $z_1, ..., z_n$ nur m verschiedene Werte $x_1, ..., x_m$ mit den Häufigkeiten $n_1, ..., n_m$ vor, so gilt

$$\bar{z} = \frac{1}{n} \sum_{j=0}^{m} n_j x_j = \sum_{j=1}^{m} \frac{n_j}{n} x_j = \frac{\sum_{j=1}^{m} n_j x_j}{\sum_{j=1}^{m} n_j}$$

$$= \text{ gewogenes Mittel der } x_1, ..., x_m.$$

z.B. Holzmann S.33/34
Das arithmetische Mittel ist nicht immer geeignet, alle Sachverhalte gut darzustellen. Deshalb werden auch andere Mittelwerte betrachtet.

2. Das geometrische Mittel

Beispiel 2.1 Kapital von 1000,–DM ist in Aktien angelegt. Im ersten Jahr vermehrt sich das Kapital überhaupt nicht, im zweiten um 10%, im dritten um 20%. Der Wert der Aktien entwickelt sich also wie folgt:

Wert am Ende des	1.Jahres	1000,–DM
	2.Jahres	1000,–DM +10%·1000DM = 1100,–DM
	3.Jahres	1100,–DM +20%·1100DM = 1320,–DM

Gesucht ist die durchschnittliche Wachstumsrate, d.h. diejenige Steigerungsrate, die den Wert der Aktien nach drei Jahren auch auf 1320,–DM hätte anwachsen lassen. Das arithmetische Mittel liefert:

$$\bar{z} = \frac{0 + 10 + 20}{3}\% = 10\%$$

Wert am Ende des	1.Jahres	$1000,–\text{DM} + 10\% \cdot 1000\text{DM} = 1100,–\text{DM}$
	2.Jahres	$1100,–\text{DM} + 10\% \cdot 1100\text{DM} = 1210,–\text{DM}$
	3.Jahres	$1210,–\text{DM} + 10\% \cdot 1210\text{DM} = 1331,–\text{DM}$

Das ist mehr als 1320,–DM. $\Rightarrow$ Das arithmetische Mittel ist hier also nicht dasjenige Mittel, das im üblichen Sprachgebrauch der Durchschnitt ist. Es ist

$$\begin{aligned} 1320 &= 1000(1+0\%)(1+10\%)(1+20\%) = 1000(1+0)(1+0,1)(1+0,2) \\ &= 1000(1+w)(1+w)(1+w) = (1+w)^3 \\ \Longrightarrow\ w &= \sqrt[3]{\frac{1320}{1000}} - 1 = 0,097 \text{ ist die mittlere Wachstumsrate.} \end{aligned}$$

Wenn w_i Wachstumsrate im Jahre i ist, so heißt $z_i = 1 + w_i$ Wachstumsfaktor. Der durchschnittliche Wachstumsfaktor ist in diesem Beispiel

$$\begin{aligned} 1+w &= \sqrt[3]{\frac{1320}{1000}} = 1,097 \\ &= \sqrt[3]{(1+0)(1+0,1)(1+0,2)} \end{aligned}$$

Allgemein sind die Wachstumsfaktoren gegebene Daten $z_1, ..., z_n\,(> 0)$.

Definition 2.1 *Den Wert* $\sqrt[n]{z_1 \cdot ... \cdot z_n}$ *nennt man geometrisches Mittel.*

3. Zentralwert oder Median

Seien jetzt $z_1, ..., z_n$ beliebige reelle Werte, die die Beobachtungen bezeichnen. Man ordnet die $z_1, ..., z_n$ der Größe nach:

$z_{[1]} \leq ... \leq z_{[n]}$ mit $\{z_{[i]}\}$ ist Permutation von $\{z_i\}$

und bestimmt den in der Mitte liegenden Wert.

Definition 2.2 *Für ungerades n heißt* $z_{[\frac{n+1}{2}]}$*, für gerades n jeder Wert aus* $\left[z_{[\frac{n}{2}]},\ z_{[\frac{n}{2}+1]}\right]$ *Zentralwert oder Median.*

Bezeichnung: $med(z_1, ..., z_n)$

Beispiel 2.2 Reihe von Einkommenswerten (n=11)

$$\begin{aligned} & 2450,\ 2500,\ 2500,\ 2600,\ 2600,\ 2650,\ 2700,\ 2720,\ 2750,\ 2750,\ 5240 \\ \Longrightarrow\quad & \bar{z} = \frac{1}{11} \cdot 31460 = 2860 \\ & med(z_1, ..., z_{11}) = 2650 \end{aligned}$$

Man erkennt: 10 der 11 Beobachtungen sind kleiner als $\bar{z}$.

Der Median ist unabhängig von Extremwerten, Ausreißern.

4. Modalwert

Definition 2.3 *Der Modalwert einer Beobachtungsreihe ist der am häufigsten auftretende Werte dieser Beobachtungsreihe.*

Bezeichnung: $mod(z_1, ..., z_n)$

Beispiel 2.3

$$6, 5, 4, 2, 1, 3, 3, 2, 2, 6, 5, 4$$
$$\Longrightarrow \quad mod(z_1, ..., z_{12}) = 2$$

Voraussetzungen für die Mittelwerte

Modalwerte	nominal, ordinal, metrisch
Zentralwert	ordinal, metrisch
geometrisches Mittel	metrisch
arithmetisches Mittel	metrisch

Mittelwerte als Lösung von Optimierungsaufgaben:

Seien $z_1, ..., z_n$ gegebene (reelle) Daten.

Satz 2.1 *Das arithmetische Mittel minimiert* $\sum_{i=1}^{n}(z_i - a)^2$, *d.h.*

$$\min_a \sum_{i=1}^{n}(z_i - a)^2 = \sum_{i=1}^{n}(z_i - \bar{z})^2.$$

Satz 2.2 *Der Zentralwert minimiert* $\sum_{i=1}^{n}|z_i - a|$, *d.h.*

$$\min_a \sum_{i=1}^{n}|z_i - a| = \sum_{i=1}^{n}|z_i - med(z_1, ..., z_n)|.$$

2.2 Streuungsmaße

Zur Beschreibung von Daten werden neben den Mittelwerten auch Streuungsmaße verwendet. Daß dies wichtig ist, soll folgendes Beispiel zeigen:

$1, 2, 2, 3, 4, 5, 9, 10, 12, 12 \quad \bar{z}=6$
$4, 4, 4, 6, 6, 7, 7, 7, 7, 8 \quad \bar{z}=6$

1. Spannweite $= max - min = z_{[n]} - z_{[1]}$ (sehr abhängig von Extremwerten)

2. Quartilsabstand

Unteres Quartil z_u ist ein derartiger Wert, daß ein Viertel aller Beobachtungen kleiner als z_u ist.
Oberes Quartil z_o ist ein derartiger Wert, daß ein Viertel aller Beobachtungen größer als z_o ist.

$z_o - z_u$ heißt Quartilsabstand.

Im allgemeinen liegen 50% aller Beobachtungen in $[z_u, z_o]$.

3. Mittlere lineare Abweichung

Der Wert $\frac{1}{n}\sum_{i=1}^{n}|z_i - z|$ heißt mittlere lineare Abweichung bezogen auf z. Im allgemeinen sollte $z = med(z_1, ..., z_n)$ gewählt werden.

4. Mittlere quadratische Abweichung

Der Wert $\frac{1}{n}\sum_{i=1}^{n}(z_i - z)^2$ heißt mittlere quadratische Abweichung bezogen auf z. Im allgemeinen sollte $z = \bar{z}$ gewählt werden.

Der Wert $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$ heißt Varianz von $z_1, ..., z_n$. Der Wert s heißt Standardabweichung.

2.3 Verteilung der Daten

Die Funktion $H_n(z) = \frac{1}{n}\#\{z_j : z_j \leq z,\ 1 \leq j \leq n\}$ heißt empirische Verteilungsfunktion (auch Häufigkeitsverteilung).

$\Longrightarrow$ $H_n(z) \in [0, 1]$. H_n ist

1. monoton nichtfallend,
2. rechtsseitig stetig,
3. Stufenfunktion (Stufenhöhe ist Vielfaches von $\frac{1}{n}$.)

Ist $H_n(z)$ gegeben, so können alle auftretenden Werte z_i vollständig erkannt werden. Ferner gilt:

$H_n(z_u) = 0,25$ oder z_u ist der kleinste Wert mit $H_n(z) > 0,25$
$H_n(z_o) = 0,75$ oder z_o ist der größte Wert mit $H_n(z) < 0,75$
$H_n(med(z_1, ..., z_n)) = \frac{1}{2}$ oder $med(z_1, ..., z_n)$ ist der größte Wert mit $H_n(z) < \frac{1}{2}$.

Das arithmetische Mittel, das auch erstes (empirisches) Moment genannt wird, kann in der Form

$$\bar{z} = \int z\, H_n(\mathrm{d}z)$$

geschrieben werden. Für die mittlere quadratische Abweichung ergibt sich

$$\frac{n-1}{n}s^2 = \int (z - \bar{z})^2\, H_n(\mathrm{d}z)$$

Allgemeiner definiert man

Definition 2.4

$$m_i = \int (z-\bar{z})^i \, H_n(\,\mathrm{d}z) \qquad \textit{heißt i–tes zentrales (empirisches) Moment,}$$

$$\int (z-a)^i \, H_n(\,\mathrm{d}z) \qquad \textit{heißt i–tes (empirisches) Moment bezüglich a.}$$

$$\Longrightarrow m_i \;=\; \tfrac{1}{n}\sum_{j=1}^{n}(z_j-\bar{z})^i \;=\; m_i(H_n)$$

Spezielle charakteristische Größen:

$\dfrac{m_3}{m_2^{\frac{3}{2}}}$ heißt Schiefe

$\dfrac{m_4}{m_2^2}-3$ heißt Exzeß.

Der Exzeß charakterisiert die Steilheit:

positive Werte bedeuten eine größere Steilheit als bei der Normalverteilung,
negative Werte bedeuten eine kleinere Steilheit als bei der Normalverteilung.

Mit der empirischen Verteilungsfunktion werden verschiedene weitere Kenngrößen definiert, z.B. Säulendiagramme, Quantile.

Säulendiagramme

Es seien $b_1, ..., b_l$ gegebene Zahlen mit $b_1 < ... < b_l$, $z_j \in (\,b_1, b_l]$ für alle $j = 1, ..., n$. Dann ist

$$\begin{aligned} H_n(b_{i+1}) - H_n(b_i) &= \frac{1}{n}\#\{b_i < z_j \le b_{i+1}\} \\ &= \text{ relative Häufigkeit von } (\,b_i, b_{i+1}]. \end{aligned}$$

Wir definieren

$$\varrho(x) := \begin{cases} \frac{H_n(b_{i+1})-H_n(b_i)}{b_{i+1}-b_i} & \text{für } x \in (\,b_i, b_{i+1}] \\ 0 & \text{sonst} \end{cases}$$

Es gilt

$$\begin{aligned} \int \varrho(x)\,\mathrm{d}x &= \sum_{i=1}^{l-1} \int_{b_i}^{b_{i+1}} \varrho(x)\,\mathrm{d}x \;=\; \sum_{i=1}^{l-1}(H_n(b_{i+1}) - H_n(b_i)) \\ &= H_n(b_l) \;=\; 1. \end{aligned}$$

Definition 2.5 *Die Funktion $\varrho(x)$ heißt Säulendiagramm.*

Bei der Herleitung des Säulendiagramms sind $b_1, ..., b_l$ fest gewählte Zahlen, und man bestimmt dann die Anzahl der Beobachtungen, die in jedem Intervall $(b_i, b_{i+1}]$ liegen. Meist werden gleichlange Intervalle gewählt:

$$b_{i+1} - b_i = \frac{\max_j z_j - \min_j z_j}{l} \quad \forall i$$

Eine Faustregel für die Wahl von l ist $l = 1 + 3,32 \cdot \lg n$.

Merke: Hierbei erfolgt eine Einteilung auf der Beobachtungsachse.

Quantile

Die inverse Aufgabenstellung zu dem eben betrachteten Problem lautet: Zu gegebenem $\alpha \in (0,1)$ ist ein kleinster Wert q_α gesucht, so daß $\alpha \leq H_n(q_\alpha)$ gilt.

Definition 2.6 *q_α heißt empirisches α–Quantil, falls $\alpha \cdot n$ Beobachtungen im Intervall $(-\infty, q_\alpha]$ liegen.*

Meist wird q_α minimal mit dieser Eigenschaft gewählt. Am häufigsten werden α–Quantile genutzt, für die α Vielfache von $\frac{1}{4}$ oder $\frac{1}{10}$ sind

$q_{\frac{1}{4}}, q_{\frac{2}{4}}, q_{\frac{3}{4}}$ Quartile

$q_{\frac{1}{10}}, ..., q_{\frac{9}{10}}$ Dezile .

Das heißt:

Definition 2.7 *Zu gegebenem k bestimmt man $\tilde{q}_1, ..., \tilde{q}_{k-1}$ durch $\tilde{q}_j = q_{[\frac{n}{k}]\frac{j}{n}}, \tilde{q}_j,\ j = 1, ..., k-1$. Für $k = 4$ heißen $\tilde{q}_j$ Quartile, für $k = 10$ Dezile.*

Für jedes n und $k = 2$ kann $\tilde{q}_1$ als Median gewählt werden.

Bemerkung: Falls mehrere Meßwerte gleich sind, müssen die $\tilde{q}_i$ nicht voneinander verschieden sein. Wenn alle $\tilde{q}_1, ..., \tilde{q}_{k-1}$ voneinander verschieden sind, so liegen in den Intervallen

$$(-\infty, \tilde{q}_1], (\tilde{q}_1, \tilde{q}_2], ..., (\tilde{q}_{k-2}, \tilde{q}_{k-1}]$$

jeweils $[\frac{n}{k}]$ Werte.

Die hier betrachteten Quantile können als Lösungen von Gleichungen mit der empirischen Verteilungsfunktion dargestellt werden.
q_α ist Lösung von $H_n(q-0) \leq \alpha,\ H_n(q) \geq \alpha$.

Merke: Hier erfolgt eine Einteilung der H_n–Achse.

3 Eigenschaften der empirischen Verteilungsfunktion und der empirischen Momente

3.1 Gesetze der großen Zahlen

Definition 3.1 *Sei* $\bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$ *, wobei* Y_i *unabhängige, reellwertige Zufallsgrößen sind. Wenn eine Folge* $\{a_n\}$ *reeller Zahlen existiert, so daß* $\bar{Y}_n - a_n$ *fast sicher gegen Null konvergiert, so erfüllt die Folge* $\{Y_n\}$ *das starke Gesetz der großen Zahlen, konvergiert* $Y_n - a_n$ *nur in Wahrscheinlichkeit gegen Null, so erfüllt* $\{Y_n\}$ *das schwache Gesetz der großen Zahlen.*

Satz 3.1 (1.Kolmogorowscher Satz) *Seien* $Y_1, Y_2, ...$ *unabhängige Zufallsgrößen mit* $\mathsf{E}Y_i = \mu_i$ *und* $\mathsf{Var}\, Y_i = \sigma_i^2 < \infty$.

$$\textit{Wenn} \sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty \,,$$

dann gilt für $\bar{Y}_n$ *das starke Gesetz der großen Zahlen, d.h.*

$$\bar{Y}_n - \bar{\mu}_n \xrightarrow{f.s.} 0 \qquad , \textit{ wobei } \bar{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \mu_i$$

Beweis: Wesentliches Hilfsmittel ist die Ungleichung von Hajek-Rényi (1955): Seien $X_1, X_2, ...$ unabhängige Zufallsgrößen mit $\mathsf{E}X_i = 0$, $\mathsf{Var}\, X_i = \sigma_i^2 < \infty$. Wenn $c_1, c_2, ...$ eine positive, nicht wachsende Folge reeller Zahlen ist, so gilt für beliebige ganze Zahlen m, n, $(m < n)$ und beliebige $\varepsilon > 0$:

$$\mathsf{P}\left(\max_{m \le k \le n} c_k |X_1 + ... + X_k| \ge \varepsilon \right) \le \frac{1}{\varepsilon^2} \left(c_m^2 \sum_{i=1}^{m} \sigma_i^2 + \sum_{i=m+1}^{n} c_i^2 \sigma_i^2 \right)$$

(Gnedenko: Lehrbuch der Wahrscheinlichkeitsrechnung).

Die Anwendung dieser Ungleichung für $X_i = Y_i - \mu_i$, $c_i = \frac{1}{i}$ liefert

$$\mathsf{P}\left(\max_{m \le k \le n} |\bar{X}_k| \ge \varepsilon \right) \le \frac{1}{\varepsilon^2} \left(\frac{1}{m^2} \sum_{i=1}^{m} \sigma_i^2 + \sum_{i=m+1}^{n} \frac{1}{i^2} \sigma_i^2 \right) .$$

Es folgt

$$\mathsf{P}\left(\sup_{m \le k} |\bar{X}_k| \ge \varepsilon \right) \le \frac{1}{\varepsilon^2} \left(\frac{1}{m^2} \sum_{i=1}^{m} \sigma_i^2 + \sum_{i=m+1}^{\infty} \frac{1}{i^2} \sigma_i^2 \right) \underset{m \to \infty}{\longrightarrow} 0 \,,$$

das heißt die starke Konvergenz der Folge $\bar{Y}_n - \bar{\mu}_n$ gegen Null ist gezeigt.

$\Longrightarrow \bar{Y}_n - \bar{\mu}_n \xrightarrow{f.s.} 0.$

□

Folgerung: Wenn $Y_1, Y_2, ...$ unabhängig mit $\mathsf{E}Y_i = \mu_i$, $\mathsf{Var}\, Y_i \leq c < \infty$, so genügt $\bar{Y}_n$ einem starken Gesetz der großen Zahlen.

Satz 3.2 (2.Kolmogorowscher Satz) *Seien $Y_1, Y_2, ...$ unabhängig, identisch verteilte Zufallsgrößen. Dann gilt:*
$\bar{Y}_n$ konvergiert mit Wahrscheinlichkeit 1 gegen μ genau dann, wenn $\mathsf{E}Y_i$ existiert und $\mathsf{E}Y_i = \mu$ gilt.

ohne BEWEIS

3.2 Empirische Verteilungsfunktionen

Es seien $Y_1, ..., Y_n$ unabhängige, identisch verteilte Zufallsgrößen mit der Verteilungsfunktion F. Im Unterschied zur empirischen Verteilungsfunktion H_n definieren wir als empirische Verteilungsfunktion der $Y_1, ..., Y_n$

$$\hat{F}_n(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(Y_i)$$

Definition 3.2 *Die durch $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(Y_i)$ definierte Funktion heißt die empirische Verteilungsfunktion von $Y_1, ..., Y_n$, Als meßbare Funktion der Y_i ist $\hat{F}_n(x)$ für alle x eine Zufallsgröße.*

Aus den Gesetzen der großen Zahlen folgen wichtige Eigenschaften der empirischen Verteilungsfunktion.

Satz 3.3 *Für alle (festen) x gilt $\hat{F}_n(x) \xrightarrow{f.s.} F(x)$, d.h.*

$$\mathsf{P}\left(\sup_{n\geq N}\left\{\omega : \left|\frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(Y_i(\omega)) - F(x)\right| < \varepsilon\right\}\right) \to 1 \text{ für } N \to \infty \;.$$

BEWEIS: Da $Y_1, ..., Y_n$ unabhängige und identisch verteilte Zufallsgrößen sind, sind auch $I_{(-\infty,x]}(Y_i)\,,\ i = 1, ..., n$ unabhängig und identisch verteilt.

Es gilt $\quad \mathsf{E}I_{(-\infty,x]}(Y_i) = \mathsf{P}(Y_i \leq x) = F(x)$ und
$\mathsf{Var}\, I_{(-\infty,x]}(Y_i) = F(x)(1 - F(x))$

Nach Satz 3.2 folgt damit die Behauptung.

□

Bemerkung:
Die empirische Verteilungsfunktion $\hat{F}_n$ ist ein Schätzer der Verteilungsfunktion F. Aus Satz 3.2 und dessen Beweis folgt, daß $\hat{F}_n$ erwartungstreu und stark konsistent ist, d.h. $\mathsf{E}\hat{F}_n(x) = F(x)$ und $\hat{F}_n(x) \xrightarrow{f.s} F(x)$ für alle $x \in \mathbb{R}$. Da $\hat{F}_n(x)$ eine meßbare Funktion über $(\Omega, \mathfrak{A})$ ist, ist $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(Y_i)$ eine Schätzvariable und deren Realisierungen sind Schätzungen für $F(x)$.
Darüber hinaus kann gezeigt werden, daß diese Konvergenz nicht nur punktweise, sondern gleichmäßig bezüglich x gilt:

Satz 3.4 (Glivenko-Cantelli) *Für alle $\varepsilon > 0$ gilt*

$$\mathsf{P}\left(\sup_{n\geq N}\sup_{x}|\hat{F}_n(x) - F(x)| \geq \varepsilon\right) \longrightarrow 0 \text{ für } N \to \infty \ , \ d.h.$$

$\hat{F}_n(x)$ konvergiert gleichmäßig mit Wahrscheinlichkeit 1 gegen $F(x)$

$$\sup_{x}|\hat{F}_n(x) - F(x)| \xrightarrow{f.s} 0 \ .$$

BEWEIS: Wir nehmen an, F sei stetig. (Dafür beweisen wir den Satz, er gilt aber allgemein.) Sei $\varepsilon > 0$ und k beliebig, mit $\varepsilon > \frac{1}{k}$. Wegen der Stetigkeit von F existieren Werte a_i, $a_1 < ... < a_k$ mit $F(a_i) = \frac{i}{k}$.
Für beliebiges $x \in (a_i, a_{i+1}]$ gilt

$$\begin{aligned} \hat{F}_n(a_i) \leq \ \hat{F}_n(x)) \ &\leq \hat{F}_n(a_{i+1}) \\ F(a_i) \leq \ F(x) \ &\leq F(a_{i+1}) \end{aligned}$$

und $F(a_{i+1}) - F(a_i) = \dfrac{1}{k}$. Hieraus folgt für $x \in (a_i, a_{i+1}]$

$$\begin{aligned} \hat{F}_n(a_i) - F(a_{i+1}) \leq \ \hat{F}_n(x) - F(x) \ &\leq \hat{F}_n(a_{i+1}) - F(a_i) \\ \| & \\ \hat{F}_n(a_i) - F(a_i) - \frac{1}{k} \leq \ \hat{F}_n(x) - F(x) \ &\leq \hat{F}_n(a_{i+1}) - F(a_{i+1}) + \frac{1}{k} \ . \end{aligned}$$

Damit gilt für alle x

$$\begin{aligned} |\hat{F}_n(x) - F(x)| \ &\leq \ \max_i |\hat{F}_n(a_i) - F(a_i)| + \frac{1}{k}, \\ \text{und} & \\ \sup_{n\geq N}\sup_{x}|\hat{F}_n(x) - F(x)| \ &\leq \ \sup_{n\geq N}\max_i |\hat{F}_n(a_i) - F(a_i)| + \frac{1}{k}, \\ \Longrightarrow \mathsf{P}\left(\sup_{n\geq N}\sup_{x}|\hat{F}_n(x) - F(x)| \geq \varepsilon\right) \ &\leq \ \mathsf{P}\left(\sup_{n\geq N}\max_i |\hat{F}_n(a_i) - F(a_i)| \geq \varepsilon - \frac{1}{k}\right). \end{aligned}$$

Aus dem vorigen Satz folgt: $\mathsf{P}\left(\sup_{n\geq N}\max_i |\hat{F}_n(a_i) - F(a_i)| > \varepsilon - \frac{1}{k}\right) \to 0$ für $N \to \infty$, womit der Satz bewiesen ist.

□

Der Satz von Glivenko-Cantelli zeigt, daß die Kenntnis der empirischen Verteilungsfunktion bereits wesentliche Informationen über die Verteilung einer Zufallsgröße liefert. Die Bedeutung der empirischen Verteilungsfunktion für die Statistik kommt darüber hinaus in folgenden Ansatz zum Ausdruck:
Seien $Y_1, ..., Y_n$ unabhängige, identisch verteilte Zufallsgrößen mit der Verteilungsfunktion F. Der Erwartungswert von Y_i ist definiert durch $\int z \, \mathrm{d}F(z)$, ein sinnvoller Schätzer für $\mathsf{E}Y_1$ ist $\frac{1}{n}\sum_{i=1}^{n} Y_i$, der in der Form $\int z \, \mathrm{d}\hat{F}_n(z)$ geschrieben werden kann, d.h. beim Schätzen von $\int z \, \mathrm{d}F(z)$ wird F durch $\hat{F}_n$ ersetzt. Das legt nahe, Momente $m_j(F) = \int z^j \, \mathrm{d}F(z)$ durch $m_j(\hat{F}_n)$ zu schätzen. Ist das gerechtfertigt ?

Satz 3.5 *$Y_1, ..., Y_n$ seien unabhängig identisch verteilte Zufallsgrößen. Für die nichtzentralen empirischen Momente*

$$M_{j,n} = \frac{1}{n}\sum_{i=1}^{n} Y_i^j \;=\; \int z^j \, \mathrm{d}\hat{F}_n(z) \quad \textit{gilt:}$$

Falls das j–te Moment von Y_i existiert und gleich μ_j ist, so gilt $M_{j,n} \xrightarrow{f.s} \mu_j$.

BEWEIS: Die Behauptung folgt unmittelbar aus Satz 3.2.

□

Um eine entsprechende Aussage über die zentralen empirischen Momente herleiten zu können, benötigen wir folgenden Hilfssatz.

Hilfssatz 3.6 *Seien* $X_1, ..., X_n, ...$
$Y_1, ..., Y_n, ...$
Zufallsgrößen mit

$$\mathsf{P}\left(\sup_{n\geq N} |X_n - \mu| > \varepsilon\right) \longrightarrow 0$$
$$\mathsf{P}\left(\sup_{n\geq N} |Y_n - \nu| > \varepsilon\right) \longrightarrow 0 \ .$$

Dann gilt

$$\mathsf{P}\left(\sup_{n\geq N} |X_nY_n - \mu\nu| > \varepsilon\right) \longrightarrow 0$$

BEWEIS:

$$\begin{aligned} X_nY_n - \mu\nu &= X_nY_n - \frac{1}{2}X_n\nu + \frac{1}{2}X_n\nu - \frac{1}{2}\mu Y_n + \frac{1}{2}\mu Y_n - \mu\nu \\ &= \frac{1}{2}X_n(Y_n - \nu) + \frac{1}{2}(X_n - \mu)Y_n + \frac{1}{2}(X_n - \mu)\nu + \frac{1}{2}\mu(Y_n - \nu) \ , \end{aligned}$$

$$\begin{aligned}
\Rightarrow |X_nY_n - \mu\nu| &\leq \frac{1}{2}|X_n(Y_n-\nu)| + \frac{1}{2}|(X_n-\mu)Y_n| \\
&+ \frac{1}{2}|(X_n-\mu)\nu| + \frac{1}{2}|\mu(Y_n-\nu)|
\end{aligned}$$

$$\begin{aligned}
\Longrightarrow \mathsf{P}\left(\sup_{n\geq N}|X_nY_n-\mu\nu|>\eta\right) &\leq \mathsf{P}\left(\sup_{n\geq N}\frac{1}{2}|X_n(Y_n-\nu)|>\frac{\eta}{4}\right) \\
&+ \mathsf{P}\left(\sup_{n\geq N}\frac{1}{2}|(X_n-\mu)Y_n|>\frac{\eta}{4}\right) \\
&+ \mathsf{P}\left(\sup_{n\geq N}\frac{1}{2}|\mu(Y_n-\nu)|>\frac{\eta}{4}\right) \\
&+ \mathsf{P}\left(\sup_{n\geq N}\frac{1}{2}|(X_n-\mu)\nu|>\frac{\eta}{4}\right)
\end{aligned}$$

Für die letzten beiden Wahrscheinlichkeiten gilt: $\lim_{n\to\infty}\mathsf{P}(..) = 0$. Weiter folgt für ein beliebiges $k > 0$:

$$\begin{aligned}
\mathsf{P}\left(\sup_{n\geq N}|X_n(Y_n-\nu)|>\frac{\eta}{2}\right) &= \mathsf{P}\left(\sup_{n\geq N}|X_n(Y_n-\nu)|>\frac{\eta}{2}, \sup_{n\geq N}|Y_n-\nu|\leq\frac{\eta}{2k}\right) \\
&+ \mathsf{P}\left(\sup_{n\geq N}|X_n(Y_n-\nu)|>\frac{\eta}{2}, \sup_{n\geq N}|Y_n-\nu|>\frac{\eta}{2k}\right) \\
&\leq \mathsf{P}\left(\sup_{n\geq N}|X_n|>k\right) + \mathsf{P}\left(\sup_{n\geq N}|Y_n-\nu|>\frac{\eta}{2k}\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \downarrow \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad 0
\end{aligned}$$

Es gilt auch

$$\mathsf{P}\left(\sup_{n\geq N}|X_n-\mu|>\varepsilon\right)\to 0 \quad\Longrightarrow\quad \mathsf{P}\left(\sup_{n\geq N}|X_n|>k\right)\to 0 \quad \text{für ein } k$$

und somit folgt die Behauptung.

□

Analog gilt auch

$$X_n + Y_n \xrightarrow{f.s.} \mu + \nu \ .$$

Sei $m_j(F)$ das zentrale Moment von Y:

$$m_j(F) = \mathsf{E}(Y - \mathsf{E}Y)^j$$

Dann gilt

Satz 3.7 *Falls das j–te Moment von Y_i existiert, so konvergiert $m_j(\hat{F}_n)$ fast sicher gegen das j–te zentrale Moment $m_j(F)$.*

$$m_j(\hat{F}_n) \xrightarrow{f.s.} \sum_{k=0}^{j} \binom{j}{k} (-\mu_1)^{j-k} \mu_k \;=\; m_j(F) \;,$$

wobei $\mu_j = \mathsf{E}Y_i^j$ ist.

BEWEIS: Durch algebraische Umformungen erhält man

$$\begin{aligned} m_j(\hat{F}_n) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^j \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{j}\binom{j}{k} Y_i^k (-\bar{Y}_n)^{j-k} \\ &= \sum_{k=0}^{j}\binom{j}{k}(-1)^{j-k}\bar{Y}_n^{j-k}\frac{1}{n}\sum_{i=1}^{n} Y_i^k \;. \end{aligned}$$

Nach Satz 3.5 gilt: $\frac{1}{n}\sum_{i=1}^{n} Y_i^k \to \mu_k$ f.s., die Anwendung des Hilfssatzes 3.6 liefert

zunächst

$$\bar{Y}_n^l \xrightarrow{f.s.} \mu_1^l \quad \text{und}$$

$$\bar{Y}_n^{j-k}\overline{Y_n^k} \xrightarrow{f.s.} \mu_1^{j-k}\mu_k \;,$$

schließlich

$$m_j(\hat{F}_n) \xrightarrow{f.s.} \sum_{k=0}^{j}\binom{j}{k}(-1)^{j-k}\mu_1^{j-k}\mu_k \;.$$

Da

$$m_j(F) \;=\; \int (x-\mu_1)^j \,\mathrm{d}F(x) \;=\; \sum_{k=0}^{j}\binom{j}{k}(-1)^{j-k}\mu_1^{j-k}\mu_k$$

folgt damit die Behauptung.

□

$\Longrightarrow$ Die empirischen Momente konvergieren stark gegen die Momente.

Wir berechnen die Erwartungswerte der empirischen Momente:

$$m_j(\hat{F}_n) = \sum_{k=0}^{j}\binom{j}{k}(-1)^{j-k}\bar{Y}_n^{j-k}\frac{1}{n}\sum_{i=1}^{n} Y_i^k$$

Voraussetzung: Y_i unabhängig, identisch verteilt.

$\mathsf{E}Y_i^k = \mu_k \qquad k \le j.$

Satz 3.8 *Im allgemeinen gilt* $\mathsf{E}m_j(\hat{F}_n) \neq m_j(F)$.

BEWEIS: Wir wählen $j = 2$

$$\begin{aligned}
m_2(F) &= \int (y-\mu_1)^2 \,\mathrm{d}F(y) = \mu_2 - \mu_1^2 \\
m_2(\hat{F}_n) &= \frac{1}{n}\sum_{i=1}^{n}((Y_i - \bar{Y}_n)^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \underbrace{\frac{2}{n}\sum_{i=1}^{n} Y_i\bar{Y}_n}_{2\bar{Y}_n^2} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_n^2}_{\bar{Y}_n^2} \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}_n^2 \\
\mathsf{E}m_2(\hat{F}_n) &= \frac{1}{n}n\mu_2 - \mathsf{E}\bar{Y}_n^2 = \mu_2 - \mathsf{E}\bar{Y}_n^2 \\
\mathsf{E}\bar{Y}_n^2 &= \frac{1}{n^2}\mathsf{E}\left(Y_1^2 + ... + Y_n^2 + \sum_{i\neq l} Y_iY_l\right) \\
&= \frac{1}{n}\mu_2 + \frac{n(n-1)}{n^2}\mu_1^2 = \frac{1}{n}\mu_2 - \frac{n-1}{n}\mu_1^2 \\
\Longrightarrow \mathsf{E}m_2(\hat{F}_n) &= \frac{n-1}{n}\mu_2 - \frac{n-1}{n}\mu_1^2 = \frac{n-1}{n}\left(\mu_2 - \mu_1^2\right)
\end{aligned}$$

□

Zusammenfassung
Für die empirische Verteilungsfunktion

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(Y_i)$$

gilt

1. Satz von Glivenko-Cantelli

$$\sup_x \left|\hat{F}_n(x) - F(x)\right| \longrightarrow 0 \qquad \text{mit Wahrscheinlichkeit 1}$$

2. Für feste Punkte $b_1, ..., b_l$ aus $\mathcal{Y}$ mit $b_1 < b_2 < ... < b_l$, ist die Funktion $\varrho(x)$ mit

$$\varrho(x) = \frac{\hat{F}_n(b_{i+1}) - \hat{F}_n(b_i)}{b_{i+1} - b_i} \qquad \text{für } x \in (b_i, b_{i+1}]$$

das Säulendiagramm. Es approximiert $F'(x)$, falls $F'(x)$ existiert.

3. Die Quantile bzw. empirischen Quantile geben an, wieviele Beobachtungen in den Intervallen $(-\infty, q_\alpha]$ liegen.

4. Die empirischen zentralen bzw. nichtzentralen Momente konvergieren fast sicher gegen die entsprechenden zentralen bzw. nichtzentralen Momente

   $$m_j(\hat{F}_n) \longrightarrow m_j(F) \quad \text{mit Wahrscheinlichkeit 1.}$$

   Sie sind im allgemeinen nicht erwartungstreu, d.h.

   $$\mathsf{E}m_j(\hat{F}_n) \neq m_j(F).$$

Literatur:

Gnedenko: Lehrbuch der Wahrscheinlichkeitstheorie
Krengel : Einführung in die Wahrscheinlichkeitstheorie und Statistik

4 Statistische Modelle

In der Statistik soll aus Beobachtungen eine Aussage über unbekannte Verteilungen gemacht werden. Dabei muß im vorhinein geklärt werden, welche Klasse von Verteilungen in Frage kommt. Man muß dazu eine Vorstellung haben, zu welchen Vorgängen bestimmte Verteilungen „passen “.
In der Wahrscheinlichkeitstheorie wurden Wahrscheinlichkeiten für Ereignisse berechnet, wenn bekannte Verteilungen für die Ergebnisse vorausgesetzt werden konnten (Würfel, Lotto, Geburtstagszwillinge,...). Das ist wegen der vollständigen Kenntnis der Verteilungen kein Gegenstand für die Statistik.

4.1 Binomialverteilungsannahme

Ein Experiment (Versuch) besteht aus der n-fachen Wiederholung eines Einzelexperiments. Jedes Einzelexperiment habe genau zwei verschiedene Ergebnisse:

Erfolg	–	Mißerfolg
gut	–	schlecht
1	–	0

Bezeichnet man die Ausgänge mit $0, 1$, so ist $y_i = \begin{cases} 0 \\ 1 \end{cases}$
und $\mathcal{Y} = \{y \mid y = (y_1, ..., y_n)\} = \{0, 1\}^n$.
ϑ ist die Wahrscheinlichkeit, daß $Y_i = 1$ ist (Erfolgwahrscheinlichkeit). Wir schreiben $\vartheta = \mathsf{P}(Y_i = 1)$. Es gilt $\mathsf{P}(y) = \vartheta^k(1-\vartheta)^{n-k}$, falls genau k mal die 1 in y ist. Ein solches

Experiment heißt Bernoulli-Experiment. P hängt von ϑ ab. Wenn also ein solcher Verlauf betrachtet wird, ist es sinnvoll, die Klasse

$$\mathcal{P} = \{\mathsf{P}_\vartheta \mid \vartheta \in [0,1]\} \quad \text{mit } \mathsf{P}_\vartheta(y) = \vartheta^k(1-\vartheta)^{n-k} \quad \text{für } k \text{ mal 1 in } y$$

zu betrachten. ϑ ist unbekannt. Hier tritt schon ein wesentliches Problem auf. Gegeben ist $(\Omega, \mathfrak{A}, \mathsf{P})$. Y ist eine zufällige Variable
$(\Omega, \mathfrak{A}, \mathsf{P}) \stackrel{Y}{\to} (\mathcal{Y}, \mathfrak{C}, \mathsf{P}_\vartheta)$
Woher kommt auf einmal das ϑ?
Zunächst ist wichtig, daß die Meßbarkeit von Y unabhängig von P, P_ϑ ist, d.h. Zufallsgröße zu sein, ist durch $\mathfrak{A}$, $\mathfrak{C}$ bestimmt. Man kann nun zwei Denkmodelle finden:

1. Y hängt von einem Parameter ϑ ab, wird aber nicht explizit ausgewiesen.
   $(\Omega, \mathfrak{A}, \mathsf{P}) \stackrel{Y}{\to} (\mathcal{Y}, \mathfrak{C}, \mathsf{P}_\vartheta)$.

2. Y ist von ϑ unabhängig, betrachtet wird eine Folge von Wahrscheinlichkeitsgrundräumen:
   $(\Omega, \mathfrak{A}, \mathsf{Q}_\vartheta) \stackrel{Y_\vartheta}{\to} (\mathcal{Y}, \mathfrak{C}, \mathsf{P}_\vartheta)$.

In beiden Fällen ist das Problem, Aussagen über P_ϑ abzuleiten. Das ist im allgemeinen der Raum, der die Aufgabenstellung bestimmt. Wir bevorzugen die 1. Darstellung. So wie bei einer Funktion das Argument nicht angegeben wird, wird auch hier das ϑ nicht bei Y explizit geführt. Falls ϑ gegeben ist, so ist z.B. die Wahrscheinlichkeit für

$$\mathcal{E}_k = \left\{ y \mid y \in \mathcal{Y}, \ \sum\nolimits_{i=1}^{n} y_i = k \right\} \quad \text{gerade } \mathsf{P}_\vartheta(\mathcal{E}_k) = \binom{n}{k} \vartheta^k (1-\vartheta)^{n-k} = b_{n,\vartheta}(k).$$

Ist n groß, ϑ klein, so kann zur Berechnung von $\mathsf{P}_\vartheta(\mathcal{E}_k)$ eine Approximation durch die Poissonverteilung genutzt werden. Derartige Situationen treten auf bei der Berechnung der Wahrscheinlichkeit für

- die zufällige Anzahl von nichtkeimenden Samenkörnern aus einer Packung von 1000 Körnern, wenn von diesem Saatgut durchschnittlich 1% nicht keimt.

- die zufällige Anzahl von Telefonanrufen, die in einem bestimmten Zeitabschnitt in einer Telefonzentrale einlaufen

- die zufällige Anzahl von α-Teilchen, die von einer radioaktiven Substanz in einem bestimmten Zeitintervall emittiert werden.

Satz 4.1 (von Poisson) *Betrachtet wird eine Folge binomialverteilter Zufallsgrößen* $Y_i^{(n)}$, $i = 1, ..., n$ *mit* $\mathsf{P}(Y_i^{(n)} = 1) = p_n$, $\mathsf{P}(Y_i^{(n)} = 0) = 1 - p_n$.
Dann ist $S_n = \sum\limits_{i=1}^{n} Y_i^{(n)}$ *eine binomialverteilte Zufallsgröße mit den Parametern* n *und* p_n. *Wenn* $\lim\limits_{n\to\infty} np_n = \lambda \geq 0$ *, dann gilt*

$$\lim_{n\to\infty} \mathsf{P}(S_n = k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}.$$

Eine Variable X mit den möglichen Werten $k = 0, 1, 2, ...$ und $\mathsf{P}(X = k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}$ ist Poissonverteilt mit dem Parameter $\lambda \geq 0$.

4.2 Multinomialverteilung

Ein Experiment bestehe aus der n-fachen Wiederholung eines Einzelexperiments, das genau r mögliche Ausgänge $1, ..., r$ habe. Es bezeichne $\vartheta_j = \mathsf{P}\{j\}$, $\mathcal{Y} = \{1, ..., r\}^n$.
Falls $\vartheta_1, ..., \vartheta_r$ gegeben sind, so ist die Wahrscheinlichkeit, in den n Einzelexperimenten insgesamt k_1–mal Ausgang 1, k_2–mal Ausgang 2,... und k_r–mal Ausgang r zu erhalten

$$\frac{n!}{k_1!k_2!...k_r!}\vartheta_1^{k_1}\vartheta_2^{k_2}...\vartheta_r^{k_r},$$

falls $k_i \geq 0$, $\sum k_i = n$, $\sum \vartheta_i = 1$ und sonst $= 0$.
Derartige Experimente kommen bei r möglichen unterschiedlichen Ergebnissen vor. Im Prinzip sind das alle praktisch relevanten Fälle, nur daß r und damit die Anzahl der verschiedenen ϑ_i groß ist. Hier haben wir $\vartheta = (\vartheta_1, ..., \vartheta_r)$ als unbekannten Parameter und

$$\mathcal{P} = \{\mathsf{P}_\vartheta \mid \vartheta \in [0, 1]^r\}$$

mit $\mathsf{P}_\vartheta(y) = \vartheta_1^{k_1}...\vartheta_r^{k_r}$ für k_1 mal $1, ..., k_r$ mal r in y.

4.3 Geometrische Verteilung

In einer unendlichen Folge von Bernoulli-Experiment mit der Erfolgswahrscheinlichkeit ϑ wird gezählt, in welchem Einzelversuch das erste Mal ein Erfolg auftritt. Die Wahrscheinlichkeit, den ersten Erfolg im k-ten Einzelversuch zu haben ist

$$\vartheta(1 - \vartheta)^{k-1} \qquad k = 1, 2, ...$$

Die Werte k sind hierbei die Ergebnismenge. Bezeichnen wir mit Y die Variable, die diese Werte k annimmt, so gilt

$$\mathsf{P}_\vartheta(k) = \vartheta(1 - \vartheta)^{k-1}$$

Wir haben $\sum_{k=1}^{\infty} \mathsf{P}_\vartheta(k) = 1$. Damit ist P_ϑ ein normiertes Maß.
Diese Verteilung heißt geometrische Verteilung.

4.4 Normalverteilung

a) Verteilung mit maximaler Entropie

Aus der Informationstheorie kennt man die Entropie einer Verteilung mit der Dichte $p(x)$

$$E = \int p(x) \ln \frac{1}{p(x)} \, \mathrm{d}x.$$

Die Entropie ist ein Maß, das maximal bei Gleichverteilung wird, das heißt es gibt keine besondere Ordnung, nur Unordnung. (Durchmischung von Gasen oder Kugelsorten)
Das gilt nicht mehr bei Nebenbedingungen.

Satz 4.2 *Unter allen stetigen Verteilungen (Dichte p) mit gegebenem Erwartungswert μ und gegebener Varianz σ^2, d.h.*

$$\int x p(x) \, \mathrm{d}x = \mu \ , \qquad \int (x-\mu)^2 p(x) \, \mathrm{d}x = \sigma^2 \ , \qquad (*)$$

besitzt nur die Normalverteilung maximale Entropie.

BEWEIS: Es ist für beliebige zwei Dichten bekannt

$$-\int p(x) \ln p(x) \, \mathrm{d}x \leq -\int p(x) \ln q(x) \, \mathrm{d}x$$

Jetzt wählt man $\ln q = \alpha + \beta(x-\mu) + \gamma(x-\mu)^2$ und α, β, γ derart, daß $(*)$ gilt.

$$\begin{aligned} \Rightarrow -\int p(x) \ln q(x) \, \mathrm{d}x &= -(\alpha + \gamma\sigma^2) \quad \text{unabhängig von } p \\ \Rightarrow -\int p(x) \ln p(x) \, \mathrm{d}x &\leq -(\alpha + \gamma\sigma^2). \end{aligned}$$

Hier haben wir wegen $(*)$:

$$\begin{aligned} \int x p(x) \, \mathrm{d}x = \mu &= \int x \, \mathrm{e}^{\alpha+\beta(x-\mu)+\gamma(x-\mu)^2} \, \mathrm{d}x \\ &= \mathrm{e}^{\alpha} \int x \, \mathrm{e}^{\beta(x-\mu)+\gamma(x-\mu)^2} \, \mathrm{d}x \end{aligned}$$

$$\int (x-\mu)^2 q(x) \, \mathrm{d}x = \sigma^2 \qquad \alpha = -\ln(\sqrt{2\pi}\sigma) \ , \quad \beta = 0, \ \gamma = -\frac{1}{2\sigma^2}$$

$$\Rightarrow -\int p(x) \ln p(x) \, \mathrm{d}x \leq \ln(\sqrt{2\pi}\sigma) + \frac{1}{2} \qquad \forall p.$$

Speziell für

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \, \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

wird obige Schranke angenommen.

□

b) Fehlertheorie

Es wird angenommen, daß ein deterministischer Wert nur mit einem Fehler additiv überlagert beobachtet werden kann:

$$a \; + \; \varepsilon.$$

ε entstehe selbst durch additive Überlagerung kleiner Fehler, die beschrieben werden können durch Variable $Z_1, ..., Z_r$, i.i.d., zweipunktverteilte $Z_j = \begin{cases} -\gamma \\ \gamma \end{cases}$,

$\mathsf{P}(-\gamma) = \mathsf{P}(\gamma) = \frac{1}{2} \quad \Rightarrow \quad \mathsf{E}Z_j = 0 \quad \mathsf{Var}\; Z_j = \frac{1}{2}\gamma^2 + \frac{1}{2}\gamma^2 = \gamma^2$

$\varepsilon = Z_1 + ... + Z_r \quad \Rightarrow \quad \mathsf{E}\varepsilon = 0\,, \quad \mathsf{Var}\; \varepsilon = r\gamma^2$

Die Grenzverteilung von ε ist, falls $r \to \infty$ und $\gamma \to 0$, so daß $r\gamma^2 \to \sigma^2$ (fest) durch den Zentralen Grenzwertsatz gegeben:

$$\begin{aligned} \mathfrak{L}\left((Z_1 + ... + Z_r) \cdot \frac{1}{\sqrt{r}\gamma} \right) &\to \mathsf{N}(0,1) \\ \text{d.h. } \mathfrak{L}\,(Z_1 + ... + Z_r) &\to \mathsf{N}(0, \sigma^2). \end{aligned}$$

4.5 Lebensdauerverteilungen

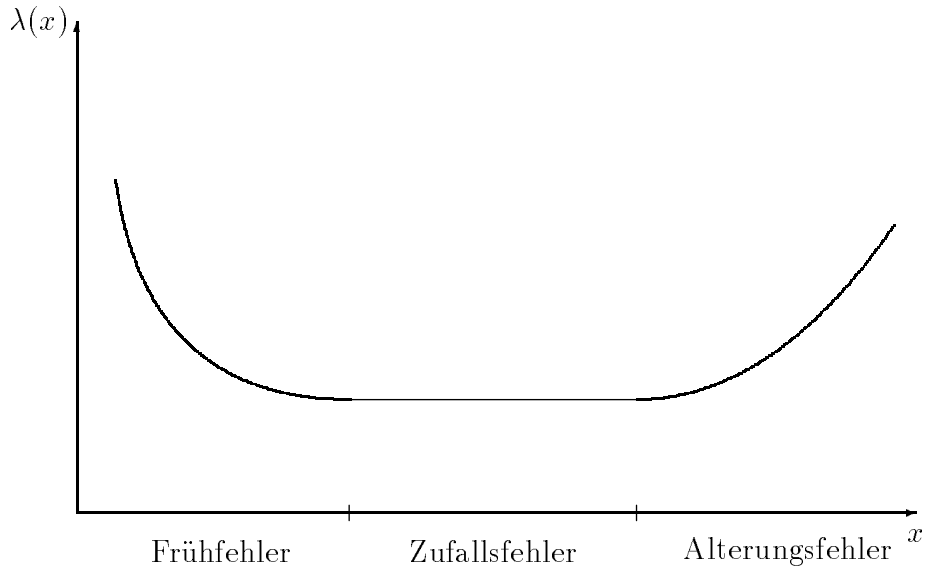
In der Zuverlässigkeitstheorie beruhen wichtige Kenngrößen auf der zufälligen Lebensdauer T eines Systems oder eines Elements. Die Verteilungsfunktion $F(x) = \mathsf{P}(T \le x)$ heißt Ausfallwahrscheinlichkeit. (Wahrscheinlichkeit dafür, daß ein Ausfall vor x stattfindet) Im allgemeinen ist die fehlerfreie Arbeitszeit T eine stetige Zufallsgröße. Sei $f(x)$ die Dichte von F. Eine wesentliche Größe ist die bedingte Wahrscheinlichkeit für den Ausfall des Systems in einem kleinen Zeitintervall $(x, x+h]$, wenn es bis zum Zeitpunkt x noch arbeitet. Als Ausfallrate bezeichnet man den Grenzwert

$$\lim_{h \to 0} \frac{1}{n} \mathsf{P}(x \le T \le x + h | T \ge x) \; .$$

Durch Umformungen erhält man

$$\begin{aligned} \lambda(x) \qquad & \text{Ausfallrate} \\ \lambda(x) \;\; = \;\; & \lim_{h \to 0} \frac{1}{h} \quad \underbrace{\mathsf{P}\,(x < T \le x + h \mid T \ge x)}_{\begin{array}{c} = \; \frac{\mathsf{P}(x < T \le x+h)}{\mathsf{P}(T > x)} \\ = \; \frac{f(x) \cdot h + o(h)}{1 - F(x)} \end{array} \quad = \quad \frac{\int_x^{x+h} f(\xi)\, \mathrm{d}\xi}{1 - F(x)}} \\ = \;\; & \frac{f(x)}{1 - F(x)} \; . \end{aligned}$$

Die Ausfallrate hat für viele Elemente die Gestalt einer „Badewannenkurve “.



Alle drei Fehlerarten besitzen Zufallscharakter.

1. Für welche Verteilung ist $\lambda(x)$=const

$$\begin{aligned} \text{d.h.} \quad f(x) &= \lambda(1-F(x)) = \lambda \int_x^\infty f(\xi)\,\mathrm{d}\xi \\ f'(x) &= -\lambda f(x) \\ \Rightarrow \quad f(x) &= a\,\mathrm{e}^{-\lambda x} \end{aligned}$$

d.h. $f(x)$ ist Dichte der Exponentialverteilung.

$$\int_0^\infty f(\xi)\,\mathrm{d}\xi = 1 = a\int_0^\infty \mathrm{e}^{-\lambda\xi}\,\mathrm{d}\xi = \frac{-1}{\lambda} a\,\mathrm{e}^{-\lambda\xi}\,|_0^\infty = \frac{a}{\lambda}\cdot 1$$

$$\Rightarrow f(x) = \lambda\,\mathrm{e}^{-\lambda x}$$

2. Weibullverteilung
hat in einem mittleren x–Bereich ähnliches Verhalten wie die Exponentialverteilung.

$$\begin{aligned} f(x) &= \lambda\alpha x^{\alpha-1}\,\mathrm{e}^{-\lambda x^\alpha} \\ \lambda(x) &= \lambda\alpha x^{\alpha-1} \end{aligned}$$

für $\alpha \le 1$: $\lambda(x)$ fallend (geeignet für Frühfehler)

für $\alpha \ge 1$: $\lambda(x)$ steigend (geeignet für Alterungsfehler)

Zusammenfassend ergibt sich, daß zu vielen Sachverhalten entsprechende Wahrscheinlichkeitsverteilungen gewählt werden können. D.h. man setzt eine Klasse $\mathcal{P}$ von Wahrscheinlichkeitsverteilungen voraus und meint damit die Beschreibung der Vorgänge:

- Meßfehler, größtmögliche Durchmischung

  Normalverteilung

- Ergebnisse sind diskret, treten mit gewissen Wahrscheinlichkeiten auf, Anzahlen sollen untersucht werden.

  Binomialverteilung, Multinomialverteilung, Poissonverteilung

- Lebensdauerverteilung

  Exponentialverteilung, Weibullverteilung

- Qualitätskontrolle

  Binomialverteilung, Geometrische Verteilung.

Grundvoraussetzung:

Daten $y_1, ...y_n$ sind Realisierungen von Zufallsgrößen $Y_1, ..., Y_n$
$(\Omega, \mathfrak{A}, \mathsf{P})$ Wahrscheinlichkeitsraum

Y heißt Zufallsgröße mit Werten in $\mathcal{Y}$, falls die Abb. $Y : (\Omega, \mathfrak{A}) \xrightarrow{Y} (\mathcal{Y}, \mathfrak{C})$ meßbar ist.
Y ist $\mathcal{Y}$-wertig. Am gebräuchlichsten sind $\mathbb{R}^1$-wertige (reellwertige) Zufallsgrößen und auch $\mathbb{R}^p$-wertige (Zufallsvektoren). $Y(\omega)$ = Realisierungen von Zufallsgrößen.
Bei den Zufallsgrößen interessieren vor allen die Verteilungen. Diese Familien von Wahrscheinlichkeitsverteilungen können durch
$\mathcal{P} = \{\mathsf{P}_\vartheta \mid \vartheta \in \Theta\}$ beschrieben werden, wobei ϑ ein Element aus einem endlichdimensionalen Raum ist, d.h. $\Theta \in \mathbb{R}^k$.
Θ heißt Parameterraum der Verteilungsfamilie.
Alle statistischen Problemstellungen oder Verfahren, die für derartige Θ untersucht werden, heißen parametrische Probleme oder Verfahren.

5 Wichtige Verteilungen, die mit der multivariaten Normalverteilung zusammenhängen

Es sei $Y = (Y_1, ..., Y_p)'$ eine p-dimensionale Zufallsgröße, wobei die ersten und zweiten Momente der Zufallsgrößen Y_i existieren sollen. Dann wollen wir folgende Bezeichnungen verwenden.

$$\mathsf{E}Y \;=\; \text{Erwartungswert von } Y \;=\; \begin{pmatrix} \mathsf{E}Y_1 \\ \vdots \\ \mathsf{E}Y_p \end{pmatrix}$$

$$\mathsf{Var}\, Y \;=\; \text{Varianzmatrix von } Y \;=\; \begin{pmatrix} \mathsf{Var}\, Y_1 & \dots & \mathsf{Cov}\,(Y_1, Y_p) \\ \vdots & \ddots & \vdots \\ \mathsf{Cov}\,(Y_p, Y_1) & \dots & \mathsf{Var}\, Y_p \end{pmatrix} .$$

$\Rightarrow \mathsf{Var}\, l'Y = l'\mathsf{Var}\, Y l$ für beliebige $l \in \mathbb{R}^p$.
$\mathsf{Var}\, Y$ heißt auch Kovarianzmatrix von Y.
Sei X eine q-dimensionale Zufallsgröße. Dann ist die Kovarianzmatrix von X und Y gegeben durch

$$\begin{aligned} \mathsf{Cov}\,(X,Y) &= \mathsf{E}\,\{(X - \mathsf{E}X)(Y - \mathsf{E}Y)'\} \\ &= \begin{pmatrix} \mathsf{Cov}\,(X_1, Y_1) & \dots & \mathsf{Cov}\,(X_1, Y_p) \\ \vdots & \ddots & \vdots \\ \mathsf{Cov}\,(X_q, Y_1) & \dots & \mathsf{Cov}\,(X_q, Y_p) \end{pmatrix} \end{aligned}$$

und es gilt für beliebige $l \in \mathbb{R}^p, h \in \mathbb{R}^q$

$$\mathsf{Cov}\,(h'X, l'Y) \;=\; h'\mathsf{Cov}\,(X,Y)l\,.$$

Falls $X_1, ..., X_k$ q-dimensionale Zufallsgrößen und $a_1, ..., a_k$ Konstanten sind, so gilt

$$\mathsf{Var}\left(\sum_{i=1}^k a_i X_i\right) \;=\; \sum_{i=1}^k a_i^2 \mathsf{Var}\, X_i + 2\sum_{i=1}^k \sum_{j=1, j\neq i}^k a_i a_j \mathsf{Cov}\,(X_i, X_j)\,,$$

$$\mathsf{Var}\left(\sum_{i=1}^k a_i X_i\right) \;=\; \sum_{i=1}^k a_i^2 \mathsf{Var}\, X_i\,, \text{ falls } X_i \text{ unkorreliert sind und}$$

$$\mathsf{Var}\left(\sum_{i=1}^k a_i X_i\right) \;=\; \left(\sum_{i=1}^k a_i^2\right) \mathsf{Var}\, X_i\,, \quad \begin{array}{l}\text{falls } X_i \text{ unkorreliert sind} \\ \text{und die gleiche Varianzmatrix haben.}\end{array}$$

Für eine $m \times q$ Matrix B gilt:

$$\mathsf{E}BX \;=\; B\mathsf{E}X \qquad \mathsf{Var}\,(BX) \;=\; B\,\mathsf{Var}\, XB'$$

Satz 5.1 *Es sei* $\mathsf{E}X = \mu\,, \mathsf{Var}\, X = \Sigma\,.$ *Dann liegt* $X - \mu$ *in* $\mathcal{R}(\Sigma)$ *mit Wahrscheinlichkeit* 1.

BEWEIS: Es gilt:

$$\begin{aligned} \mathsf{Var}\,(l'Y) &= l'\Sigma l \\ &= \mathsf{E}\,(l'(Y-\mu)(Y-\mu)'l) \\ &= \mathsf{E}\,(l'(Y-\mu))^2 \ . \end{aligned}$$

Sei $l \in \mathcal{N}(\Sigma) \Rightarrow l'\Sigma l = 0$
$\Rightarrow l'(Y-\mu) = 0$ mit Wahrscheinlichkeit 1
$\Rightarrow$ Behauptung.

□

Damit hat $X - \mathsf{E}X$ nur Werte in $\mathcal{R}(\mathsf{Var}\ X) = \mathcal{R}(\Sigma)$, was insbesondere eine Bedeutung hat, wenn Σ eine singuläre Matrix ist.

Folgerung:
X sei eine q-dimensionale Zufallsgröße, $\Sigma = \mathsf{Var}\ X$ singulär, $rg\Sigma = q_1$. Dann existiert eine Drehung $B : \mathbb{R}^q \to \mathbb{R}^q$ derart, daß $B(X-\mu) = \binom{Z}{0}$ mit Wahrscheinlichkeit 1 ist für eine q_1-dimensionale Zufallsgröße Z mit $\mathsf{Var}\ Z > 0$.

5.1 Mehrdimensionale Normalverteilung

Es sei $Y = (Y_1, ..., Y_p)'$ eine p-dimensionale Zufallsgröße.

Definition 5.1 *Y hat eine p-dimensionale Normalverteilung, falls jede lineare Funktion $l'Y$ eine eindimensionale Normalverteilung oder eine Einpunktverteilung hat.*

Damit wird die mehrdimensionale Normalverteilung durch die eindimensionale Normalverteilung erklärt.

Satz 5.2 *Y habe eine p-dimensionale Normalverteilung. Dann existieren $\mathsf{E}Y$ und $\mathsf{Var}\ Y$.*

BEWEIS: Für jedes $l \in \mathbb{R}^p$ existieren $\mathsf{E}(l'Y), \mathsf{Var}\,(l'Y)$. Wir kennen $\mathsf{E}Y_i = \mu_i$, $\mathsf{Var}\ Y_i = \sigma_{ii}$, $\mathsf{Cov}\,(Y_i, Y_j) = \mu_{ij}$.

$\Rightarrow$ $\mathsf{E}l'Y = l'\mu\ \forall l$, $\mathsf{Var}\,(l'Y) = l'\Sigma l \quad \forall l$.

$\Rightarrow$ μ ist eindeutig,
und Σ ist in der Klasse der positiv semidefiniten Matrizen eindeutig

$\Rightarrow$ μ, Σ als Erwartungswert und Varianzmatrix existieren.

□

Bezeichnung: $Y \sim \mathsf{N}_p(\mu, \Sigma)$

Satz 5.3 *Y habe eine p-dimensionale Normalverteilung und B sei beliebige $q \times p$-Matrix Dann ist $BY \sim \mathsf{N}_q(B\mu, B\Sigma B')$*

BEWEIS: $\forall t \in \mathbb{R}^q$ gilt: $t'BY \sim \mathsf{N}(t'B\mu, t'B\Sigma B't)$
$\Rightarrow$ mit $t'B = l'$ folgt die Behauptung.

□

Aus der Folgerung zu Satz 5.1 erhalten wir für normalverteilte Variable die

Folgerung:
Y habe eine p-dimensionale Normalverteilung. Dann existiert eine Drehung B in $\mathbb{R}^p$ derart, daß $B(Y - \mathsf{E}Y) = \binom{Z}{0}$ für eine p_1-dimensionale Zufallsgröße, wobei $rg\mathsf{Var}\ Y = p_1$ ist und Z eine p_1-dimensionale Normalverteilung mit nichtsingulärer Varianzmatrix $\mathsf{Var}\ Z$ hat.
BEWEIS: Es ist nur zu zeigen, daß Z normalverteilt ist. Das folgt aber aus Vorangegangenem.

□

Damit haben wir $Y = \mathsf{E}Y + B^{-1}\binom{Z}{0}$.
Setzt man $\mathsf{Var}\ Z = \Sigma_1$, so gilt für $G = \Sigma_1^{-\frac{1}{2}} Z$: $\mathsf{Var}\ G = I_{p_1}$.
Damit hat G eine p_1-dimensionale Normalverteilung mit unkorrelierten Komponenten. Da bei Normalverteilung unkorrelierte Komponenten unabhängig sind (vgl. Satz 6.16), gilt:
G hat eine p_1-dimensionale Normalverteilung mit unabhängigen Komponenten.

$$\Rightarrow\ Y = \mathsf{E}Y + B^{-1}\begin{pmatrix}\Sigma_1^{\frac{1}{2}} G\\ 0\end{pmatrix} = \mu + A\Sigma_1^{\frac{1}{2}} G \quad \text{für eine } p \times p_1 \text{ Matrix } A$$

$\Rightarrow$ Falls Y p-dimensional normalverteilt, so existiert μ und $C \in \mathfrak{M}_{p\times p_1}$ mit $rgC = p_1$ derart, daß für einen p_1-Vektor G unabhängiger, normalverteilter Größen
$G_i \sim \mathsf{N}(0,1), i = 1, ..., p_1$ gilt:

$$Y = \mu + CG$$

Umgekehrt: Falls die Darstellung $Y = \mu + CG$ gegeben ist, so ist Y p-dimensional normalverteilt.
Damit führt dieser Zugang zu einer äquivalenten Möglichkeit der Definition der p-dimensionalen Normalverteilung, denn wir haben den folgenden Satz bewiesen:

Satz 5.4 *Y ist p-dimensional normalverteilt mit Erwartungswert μ, Varianz Σ mit $rg\Sigma = p_1$ genau dann, wenn Y die Gestalt $Y = \mu + CG$ hat, wobei C eine $(p \times p_1)$-Matrix ist mit $rgC = p_1, CC' = \Sigma$ und $G = (G_1, ..., G_{p_1})$ ein p_1-dimensionaler Vektor mit standardnormalverteilten, unabhängigen Komponenten G_i besitzt.*

Dichtefunktion

Sei Y p-dimensional normalverteilt , $Y \sim \mathsf{N}_p(\mu, \Sigma)$.
Aus dem Satz von Radon-Nikodym folgt: Y besitzt keine Dichte bezüglich des Lebesgue-Maßes, wenn $rg\Sigma < p$.

Sei also $rg\Sigma = p$. $\Rightarrow Y = \mu + CG$ für $G \sim \mathsf{N}_p(0, I)$, $rgC = p$.
Wegen $\Sigma = CC'$ folgt

$$\det \Sigma = (\det C)^2 \neq 0 \text{ und } G = C^{-1}(Y - \mu).$$

Da $G \sim \mathsf{N}_p(0, I)$ gilt, hat G die Dichte

$$f_G(g) = \frac{1}{\left(\sqrt{2\pi}\right)^p} \exp\left(-\frac{1}{2}(g_1^2 + ... + g_p^2)\right), \ g \in \mathbb{R}^p.$$

So erhält man die Dichte von Y mit $g = C^{-1}(y - \mu)$ und nutzt, daß die Jacobische Determinante $\det C^{-1} = (\det \Sigma)^{-\frac{1}{2}}$ ist:

$$\begin{aligned} p(y) &= \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(y - \mu)'(CC')^{-1}(y - \mu)\right) \\ &= \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(y - \mu)'\Sigma^{-1}(y - \mu)\right). \end{aligned}$$

Satz 5.5 *Seien $Y_1, ..., Y_n$ unabhängige Zufallsgrößen, $Y_i \sim \mathsf{N}_p(\mu_i, \Sigma_i)$.*
Dann gilt: $Z = a_1 Y_1 + ... + a_n Y_n \sim \mathsf{N}_p(\sum a_i \mu_i, \sum a_i^2 \Sigma_i)$.

BEWEIS: Für beliebiges $l \in \mathbb{R}^p$ gilt $l'Z = a_1 l'Y_1 + ... + a_n l'Y_n$. Nach Definition der Normalverteilung und wegen der Unabhängigkeit der Y_i sind $l'Y_i$ unabhängig und eindimensional normalverteilt. Hieraus folgt, daß Z p-dimensional normalverteilt ist. Die Parameter ergeben sich aus:

$$\begin{aligned} \mathsf{E}Z &= \sum a_i \mathsf{E}Y_i \\ \mathsf{Var}\,(l'Z) &= l'\mathsf{Var}\,Z l \\ &= \sum a_i^2 \mathsf{Var}\,(l'Y_i) = \sum a_i^2 l'\Sigma_i l = l'\left(\sum a_i^2 \Sigma_i\right) l. \end{aligned}$$

□

Folgerung: $Y_i \sim \mathsf{N}_p(\mu, \Sigma)$, unabhängig $\Rightarrow \bar{Y} \sim \mathsf{N}_p(\mu, \frac{1}{n}\Sigma)$.

Satz 5.6 *Seien $Y_1, ..., Y_n$ unabhängige Zufallsgrößen mit $Y_i \sim \mathsf{N}_p(\mu_i, \Sigma)$ und $V_1 = \sum b_i Y_i$, $V_2 = \sum c_i Y_i$, $b = (b_1, ..., b_n)'$, $c = (c_1, ..., c_n)'$. Es gilt:*

$$V_1, V_2 \text{ unabhängig} \quad \Leftrightarrow \quad b'c = 0.$$

BEWEIS: $V_1, V_2 \sim \mathsf{N}_p$ nach vorigem Satz. Außerdem ist $\binom{V_1}{V_2} \sim \mathsf{N}_{2p}$, weil

$$\begin{aligned}\begin{pmatrix} l_{(1)} \\ l_{(2)} \end{pmatrix}' \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} &= \sum_i b_i l'_{(1)} Y_i \ + \ \sum_i c_i l'_{(2)} Y_i \\ &= \sum_i \left(b_i l'_{(1)} \ + \ c_i l'_{(2)} \right) Y_i \\ &= \sum_{i=1}^{n} t'_i Y_i\end{aligned}$$

Es gilt: $t_i Y_i \sim \mathsf{N}_1$, unabhängig.
Die Summe unabhängiger normalverteilter Zufallsgrößen ist normalverteilt.
Die Behauptung folgt, falls: $\mathsf{Cov}\ (V_1, V_2) = 0$, d. h.

$$\begin{aligned}\mathsf{Cov}\ \left(l'_{(1)} V_1, l'_{(2)} V_2 \right) &= 0 \qquad \forall l_{(1)}, l_{(2)}\,. \\ \mathsf{Cov}\ \left(l'_{(1)} V_1, l'_{(2)} V_2 \right) &= \mathsf{Cov}\ \left(\sum b_i l'_{(1)} Y_i, \sum c_j l'_{(2)} Y_j \right) \\ &\quad \mathsf{Cov}\ (Y_i, Y_j) \ = \ \Sigma \cdot \delta_{ij} \\ \Rightarrow \qquad &= \sum_{k=1}^{n} \mathsf{Cov}\ \left(b_k l'_{(1)} Y_k, c_k l'_{(2)} Y_k \right) \\ &= l'_{(1)} \sum_{k=1}^{n} b_k c_k \Sigma l_{(2)} \ = \ b' c l'_{(1)} \Sigma l_{(2)}.\end{aligned}$$

□

5.2 χ^2-Verteilung

Seien $Y_1, ..., Y_p$ unabhängige Zufallsgrößen mit $Y_i \sim \mathsf{N}_1(\mu_i, \sigma^2)$. Dann gilt
$Y \ = \ \mu \ + \ \sigma Z$ wobei $Z \sim \mathsf{N}_p(0, I)$.

Definition 5.2 *Die Zufallsgröße* $\frac{1}{\sigma^2}|Y|^2 \ = \ \frac{1}{\sigma^2} \sum_{i=1}^{n} Y_i^2$ *hat eine* χ^2*-Verteilung mit* n *Freiheitsgraden (FG). Wenn* $\mu_i = 0 \ \forall i$*, so heißt die Verteilung zentral, anderenfalls nichtzentral. Der Parameter* $\lambda^2 = \frac{1}{\sigma^2} \sum \mu_i^2$ *heißt Nichtzentralitätsparameter (NZP).*

$$\begin{aligned}\text{Bezeichnung:} \qquad \frac{1}{\sigma^2}|Y|^2 &\sim \chi^2_{n;\lambda^2} \\ \text{oder} \qquad |Y|^2 &\sim \sigma^2 \chi^2_{n;\lambda^2}\end{aligned}$$

Satz 5.7 *Sei* $Y \sim \mathsf{N}_p(\mu, I)$*. Falls* A *idempotent ist und* $rg A = r$*, so ist* $Y'AY$ χ^2*-verteilt mit* r *Freiheitsgraden.*

BEWEIS: Aus der Idempotenz von A folgt $A' = A$, $A^2 = A$ und, daß A eine orthogonale Projektion ist.

Ferner gilt $AY \sim \mathsf{N}_p(A\mu, A)$. Es existiert eine Drehung B des $\mathbb{R}^p$ derart, daß

$$BA = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & & \ddots & \\ 0 & & & & 0 \end{pmatrix} \qquad \text{mit } r \text{ Einsen}$$

und $B'B = I$. Hieraus folgt

$$BAY \sim \mathsf{N}_p\left(\underbrace{BA\mu}_{\nu}, BAAB'\right) \text{ und } BAAB' \text{ ist diagonal.}$$

Wir haben $Y'AY = Y'AB'BAY = \sum_{i=1}^{n} (\nu_i + Z_i)^2$ für $Z \sim \mathsf{N}_p(0, BAAB')$.
Da $Z_1, ..., Z_n \sim \mathsf{N}(0,1)$ unabhängig sind und $Z_{n+1}, ..., Z_p = 0,\ \nu_{n+1} = ... = \nu_p = 0$, folgt $Y'AY = \sum_{i=1}^{n}(\nu_i + Z_i)^2 \sim \chi^2_{n;\lambda^2}$ mit

$$\lambda^2 = \sum_{i=1}^{n} \nu_i^2 = |\nu|^2 = \mu' AB'BA\mu = \mu' A\mu \ .$$

□

Satz 5.8 *Sei $Y \sim \mathsf{N}_p(\mu, \Sigma)$, A symmetrisch. Falls $\Sigma(A\Sigma A - A)\Sigma = 0$, so gilt:*

$$Q := (Y-\mu)'A(Y-\mu) \ \sim \chi^2_{n;0} \qquad \textit{für} \ \ n = rg(A\Sigma).$$

BEWEIS: Wir haben $Y - \mu = CG$ für $G \sim \mathsf{N}_m(0, I)$ und $CC' = \Sigma$.
$\Rightarrow\ Q = G'C'ACG$. G hat unabhängige, normalverteilte G_i. Wir müssen nach Satz 5.7 zeigen, daß $C'AC$ idempotent ist:

$$\begin{aligned} C'ACC'AC &= C'AC \\ CC'ACC'ACC' &= CC'ACC' \quad , \qquad CC' = \Sigma \\ \Leftrightarrow \Sigma A\Sigma A\Sigma &= \Sigma A\Sigma. \end{aligned}$$

Die Verteilung von Q ist zentral, da $\mathsf{E}(Y - \mu) = 0$.

□

Satz 5.9 *$Y \sim \mathsf{N}_p(\mu, I)$. Sei $A : p \times p$ idempotent und C eine $m \times p$- Matrix. Dann sind $Y'AY$ und CY unabhängig, falls $CA = 0$. Wenn C idempotent, so sind $Y'AY$ und $Y'CY$ unabhängig, falls $CA = 0$.*

OHNE BEWEIS

Dichte der χ^2-Verteilung

χ^2_n hat eine Dichte bezüglich λ_1:

$$\psi_n(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\,\mathrm{e}^{-\frac{x}{2}}x^{\frac{n}{2}-1}\,, \quad 0 < x < \infty\,.$$

Hierbei ist $\Gamma(p) = \int\limits_0^\infty x^{p-1}\,\mathrm{e}^{-x}\,\mathrm{d}x \Rightarrow\ \Gamma(p) = (p-1)!$ für $p \geq 1\,,\ p \in \mathbb{N}$.

Die nichtzentrale ${\chi^2}_n$-Verteilung hat eine Dichte als unendliche Reihe:

$$f_{n,\lambda}(x) \quad = \quad \begin{cases} \displaystyle\sum_{m=0}^{\infty}\sum_{k=m+\frac{n}{2}}^{\infty} \frac{\left(\frac{\lambda}{2}\right)^m\left(\frac{x}{2}\right)^k}{m!k!}\,\mathrm{e}^{-\frac{\lambda+x}{2}} & \text{für } x \geq 0 \\ \\ 0 & \text{sonst}\,. \end{cases}$$

Momente der χ^2-Verteilung:

$$\text{Sei}\quad U \sim \frac{1}{\sigma^2}|Y|^2 \;=\; \frac{1}{\sigma^2}\sum_{i=1}^{n} Y_i^2 \qquad \text{für } Y_i \sim \mathsf{N}(\mu_i, \sigma^2)\,.$$

$$\Rightarrow \quad \mathsf{E}U \;=\; \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathsf{E}Y_i^2 \;=\; \frac{1}{\sigma^2}\sum_{i=1}^{n}(\mu_i^2+\sigma^2) \;=\; n + \sum_{i=1}^{n}\frac{\mu_i^2}{\sigma^2}$$

$$= \#\mathrm{FG} + \mathrm{NZP}\,.$$

$\Rightarrow$ Für $U \sim \chi^2_{n;0}$ gilt: $\mathsf{E}U = \#FG = n$

5.3 F-Verteilung

Definition 5.3 *Es seien X_1, X_2 unabhängige, χ^2-verteilte Zufallsgrößen mit m_1 bzw. m_2 Freiheitsgraden. Dann hat*

$$F \quad = \quad \frac{\frac{1}{m_1}X_1}{\frac{1}{m_2}X_2} \qquad \textit{eine F-Verteilung mit } (m_1, m_2) \textit{ Freiheitsgraden.}$$

Bezeichnung: $F \sim \mathsf{F}_{m_1,m_2,\lambda^2}$.
Falls X_1 und X_2 zentral verteilt sind, so hat F eine zentrale F-Verteilung, anderenfalls eine nichtzentrale.

Dichte einer zentral F_{m_1,m_2}-verteilten Variablen:

$$f(x) \quad = \quad \begin{cases} \dfrac{\Gamma\left(\frac{m_1+m_2}{2}\right)\left(\frac{m_1}{m_2}\right)^{\frac{m_1}{2}}}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)} \dfrac{x^{\frac{m_1}{2}-1}}{\left(1+\frac{m_1}{m_2}x\right)^{\frac{m_1+m_2}{2}}} & \text{für } x \geq 0 \\ \\ 0 & \text{sonst}\,. \end{cases}$$

5.4 Studentsche t-Verteilung

Seien Z und X unabhängige Zufallsgrößen mit $Z \sim \mathsf{N}(0,1)$ und $X \sim \chi^2_m$.

Definition 5.4 *Die Zufallsgröße* $t = \frac{Z}{\sqrt{m^{-1}X}}$ *hat eine Studentsche* t*-Verteilung mit* m *Freiheitsgraden.*

Bezeichnung: $t \sim \mathsf{t}_m$
Dichte einer t_m-Verteilung:

$$\tau(x) = \frac{1}{\sqrt{\pi m}} \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}} \qquad \text{für } x \in \mathbb{R}^1 \ .$$

$$\Rightarrow \qquad t^2 \sim \mathsf{F}_{1,m}$$

6 Charakteristische Funktionen

Bisher wurden reellwertige Zufallsgrößen ξ betrachtet. Funktionen heißen komplexwertig, wenn $\xi = \xi_1 + \mathrm{i}\,\xi_2$ für reellwertige Zufallsgrößen ξ_1, ξ_2 gilt. Es gibt dann die eineindeutige Zuordnung zwischen den komplexwertigen Größen und zufälligen Vektoren:
$\xi(\omega) = \xi_1(\omega) + \mathrm{i}\,\xi_2(\omega) \Leftrightarrow (\xi_1, \xi_2)$ zufälliger Vektor.
Damit übertragen sich die Rechenregeln für zweidimensionale Zufallsgrößen auf die komplexwertigen Zufallsgrößen, z.B.:
$\mathsf{E}\xi = \mathsf{E}\xi_1 + \mathrm{i}\,\mathsf{E}\xi_2$
Komplexwertige Zufallsgrößen ξ, η heißen unabhängig, falls $(\xi_1, \xi_2, \eta_1, \eta_2)$ unabhängig sind. Aus der Unabhängigkeit folgt wie bei den reellwertigen Zufallsgrößen $\mathsf{E}\xi\eta = \mathsf{E}\xi\,\mathsf{E}\eta$.
Spezielle komplexwertige Zufallsgrößen sind $\mathrm{e}^{\mathrm{i}tX}$ für reellwertige Zufallsgrößen X und $t \in \mathbb{R}$. Durch Erwartungswertbildung erhält man die sogenannten *charakteristischen Funktionen.* Diese charakterisieren Verteilungen eindeutig und sind für die Bestimmung von Grenzverteilungen, die Untersuchung von Summen von Zufallsgrößen und die Herleitung von Momenten wichtig.

Gnedenko Lehrbuch der Wahrscheinlichkeitsrechnung, Akademie Verlag
E. Lukacs Charakteristische Funktionen, Wiley Verlag
Rao C.R. Lineare statistische Methoden, Wiley Verlag.

Definition 6.1 *Als charakteristische Funktion (c.F.) einer reellen Zufallsgröße* X

bezeichnet man die komplexwertige Funktion

$$\varphi_X(t) \;=\; \mathsf{E}\exp(\mathrm{i}\,tX) \;=\; \mathsf{E}\,\mathrm{e}^{\mathrm{i}tX} \;=\; \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\,tz}\,\mathrm{d}F(z),$$

wobei $t \in \mathbb{R}$, $\mathrm{i} = \sqrt{-1}$, $\mathrm{e}^{\mathrm{i}\,tz} \;=\; \cos(tz) + \mathrm{i}\,\sin(tz)$.

In der Theorie der Integraltransformation ist φ_X die Fourier-Stieltjes- Transformation der Funktion F. Wenn X eine Dichte p bezüglich des Lebesgue-Maßes hat, so gilt

$$\varphi_X(t) \;=\; \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\,tz} p(z)\,\mathrm{d}z \qquad \text{Fourier-Transformation.}$$

Rechtfertigung: φ_X existiert,da

$$\int_{-\infty}^{\infty} |\,\mathrm{e}^{\mathrm{i}\,tz}|\,\mathrm{d}F(z) \;=\; 1$$

Beispiel 6.1 Normalverteilung: $X \sim \mathsf{N}(a, \sigma^2)$

$$\begin{aligned}
\mathsf{E}\,\mathrm{e}^{\mathrm{i}tX} \;&=\; \frac{1}{\sigma\sqrt{2\pi}} \int \mathrm{e}^{\mathrm{i}\,tz - \frac{(z-a)^2}{2\sigma^2}}\,\mathrm{d}z \;=\; \varphi_X(t) \\
&\qquad \text{Substitution: } v \;=\; \frac{z-a}{\sigma} - \mathrm{i}\,t\sigma \;, \quad \mathrm{d}v \;=\; \frac{1}{\sigma}\,\mathrm{d}z \\
&\qquad\quad v^2 \;=\; \frac{(z-a)^2}{\sigma^2} - t^2\sigma^2 - 2\,\mathrm{i}\,t(z-a) \\
&\qquad\quad v^2 \;=\; \frac{(z-a)^2}{\sigma^2} - 2\,\mathrm{i}\,tz - t^2\sigma^2 + 2\,\mathrm{i}\,ta \\
&\qquad -\frac{1}{2}v^2 \;=\; -\frac{(z-a)^2}{2\sigma^2} + \mathrm{i}\,tz + \frac{t^2\sigma^2}{2} - \mathrm{i}\,ta \\
\Rightarrow\; \varphi_X(t) \;&=\; \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{\mathrm{i}\,at - \frac{1}{2}\sigma^2 t^2} \underbrace{\int_{-\infty - \mathrm{i}\,t\sigma}^{\infty - \mathrm{i}\,t\sigma} \mathrm{e}^{-\frac{v^2}{2}}\,\mathrm{d}v}_{\sqrt{2\pi}} \\
&=\; \mathrm{e}^{\mathrm{i}\,at - \frac{1}{2}\sigma^2 t^2}
\end{aligned}$$

Eigenschaften

Satz 6.1 *Eine c.F. ist auf der ganzen reellen Achse gleichmäßig stetig und es ist*

$$\varphi_X(0) \;=\; 1 \quad , \quad |\varphi_X(t)| \;\le\; 1 \;.$$

BEWEIS: Gleichmäßige Stetigkeit:

$$\begin{aligned} |\varphi_X(t+h)-\varphi_X(t)| &= \Big|\int \mathrm{e}^{\mathrm{i}tz}\left(\mathrm{e}^{\mathrm{i}hz}-1\right)\,\mathrm{d}F(z)\Big| \\ &\le \int \underbrace{|\mathrm{e}^{\mathrm{i}tz}|}_{=1}|\mathrm{e}^{\mathrm{i}hz}-1|\,\mathrm{d}F(z). \end{aligned}$$

Wegen $|\mathrm{e}^{\mathrm{i}hz}-1|\le 2$ folgt bei geeigneter Wahl von a :

$$\le \underbrace{2\int_{-\infty}^{-a}\mathrm{d}F(z)}_{\le\frac{\varepsilon}{5}} + \underbrace{2\int_{a}^{\infty}\mathrm{d}F(z)}_{\le\frac{\varepsilon}{5}} + \int_{-a}^{a}|\mathrm{e}^{\mathrm{i}hz}-1|\,\mathrm{d}F(z).$$

Wir wählen dann δ so, daß $|\mathrm{e}^{\mathrm{i}\delta z}-1|\le\frac{\varepsilon}{5}$ für $z\in[-a,a]$ gilt. $\Rightarrow\quad |\mathrm{e}^{\mathrm{i}hz}-1|\le\varepsilon/5$ für $z\in[-a,a]$. $\Rightarrow$ Behauptung.

□

Satz 6.2 *Gilt $Y=aX+b$, so ist*

$$\varphi_X(t) = \mathrm{e}^{\mathrm{i}tb}\varphi_X(ta).$$

Satz 6.3 *Wenn $X_1,...,X_n$ unabhängige Zufallsgrößen sind, so gilt für die c.F. von $S_n=\sum_{i=1}^{n}X_i$*

$$\varphi_{S_n}(t) = \prod_{i=1}^{n}\varphi_{X_i}(t).$$

Satz 6.4 *Sei $\mathsf{E}|X|^k<\infty$, $k\ge 1$. Dann ist φ_X k-mal stetig differenzierbar, und es gilt*

$$\varphi_X^{(k)}(0) = \mathrm{i}^k\mathsf{E}X^k.$$

BEWEIS: k-maliges formales Differenzieren von φ_X liefert

$$\begin{aligned} \varphi_X^{(k)}(t) &= \mathrm{i}^k\int z^k\,\mathrm{e}^{\mathrm{i}tz}\,\mathrm{d}F(z) \qquad (*) \\ \Big|\int z^k\,\mathrm{e}^{\mathrm{i}tz}\,\mathrm{d}F(z)\Big| &\le \int |z|^k\,\mathrm{d}F(z) < \infty\ . \end{aligned}$$

Nach Voraussetzung existiert das Integral in $(*)$. Formales Differenzieren ist gerechtfertigt, da (vgl. Smirnow, Bd. II):

$\varphi_X(t)<\infty$ existiert,

$\int_{-a}^{b} z^k\,\mathrm{e}^{\mathrm{i}tz}\,\mathrm{d}F(z)$ gleichmäßig konvergent bezüglich t.

Daraus folgt die Behauptung.

□

Folgerung:

$$\mathsf{E}|X|^k < \infty \;\Rightarrow\; \varphi_X(t) \;=\; 1 + \sum_{j=1}^{k} \frac{t^j\,\mathrm{i}^{\,j}}{j!}\mathsf{E}X^j + o(t^k) \qquad \text{für } t \to 0\ .$$

Satz 6.5 *Für alle $t \in \mathbb{R}$ gilt $\varphi_X(-t) = \overline{\varphi_X(t)}$.*

BEWEIS:

$$\varphi_X(-t) \;=\; \mathsf{E}\,\mathrm{e}^{-\mathrm{i}tX} \;=\; \overline{\mathsf{E}\,\mathrm{e}^{\mathrm{i}tX}} \;=\; \overline{\varphi_X(t)}\ .$$

□

Folgerung:
Falls X eine symmetrische Verteilung bezüglich 0 hat, dann ist $\varphi_X(t)$ reell, denn

$$\varphi_X(t) \;=\; \varphi_{-X}(t) \;=\; \varphi_X(-t) \;=\; \overline{\varphi_X(t)}\ .$$

Satz 6.6 *Umkehrformel: Ist F die Verteilungsfunktion von X und ist φ_X deren c.F., so gilt für alle Stetigkeitspunkte x und y der Funktion F*

$$F(y) - F(x) \;=\; \lim_{a\to 0} \int_{-a}^{a} \frac{\mathrm{e}^{-\mathrm{i}tx} - \mathrm{e}^{-\mathrm{i}ty}}{2\pi\,\mathrm{i}\,t}\varphi_X(t)\,\mathrm{d}t$$

BEWEIS: Es wird der Cauchy'sche Hauptwert des Integrals berechnet.

□

Satz 6.7 *(Eindeutigkeitssatz) Die Verteilungsfunktion F ist durch ihre c.F. eindeutig bestimmt.*

BEWEIS: Wegen Satz 6.6 gilt

$$F(y) \;=\; \frac{1}{2\pi} \lim_{x\to-\infty} \lim_{a\to\infty} \int_{-a}^{a} \frac{\mathrm{e}^{-\mathrm{i}tx} - \mathrm{e}^{-\mathrm{i}ty}}{\mathrm{i}\,t}\varphi_X(t)\,\mathrm{d}t$$

für alle Stetigkeitspunkte x von F. Wir bemerken, daß es höchstens abzählbar unendlich viele Unstetigkeitspunkte von F gibt und $F(x) \to 0$ für $x \to -\infty$ gilt.

□

Satz 6.8 *Sei φ_X absolut integrierbar über $\mathbb{R}$, $\int |\varphi_X(t)|\,\mathrm{d}t < \infty$, F sei die Verteilungsfunktion von X. Dann ist F stetig differenzierbar, und es gilt*

$$F'(x) \;=\; \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}tx}\varphi_X(t)\,\mathrm{d}t\ .$$

BEWEIS: Aus

$$\frac{\mathrm{e}^{-\mathrm{i}tx} - \mathrm{e}^{-\mathrm{i}ty}}{\mathrm{i}\,t} = [1 - \mathrm{i}\,tx + o(t) - 1 + \mathrm{i}\,ty + o(t)]\frac{1}{\mathrm{i}\,t} = y - x + \frac{o(t)}{\mathrm{i}\,t}$$

folgt

$$\left|\frac{\mathrm{e}^{-\mathrm{i}tx} - \mathrm{e}^{-\mathrm{i}ty}}{\mathrm{i}\,t}\right| \leq C \qquad \text{für } |t| \leq 1.$$

Außerdem gilt für $|t| \geq 1$

$$\left|\frac{\mathrm{e}^{-\mathrm{i}tx} - \mathrm{e}^{-\mathrm{i}ty}}{\mathrm{i}\,t}\right| \leq 2.$$

Da $\varphi_X(t)$absolut integrierbar ist, d.h. $\int |\varphi_X(t)|\,\mathrm{d}t < \infty$, folgt $\int_{-\infty}^{\infty} \left|\frac{\mathrm{e}^{-\mathrm{i}tx} - \mathrm{e}^{-\mathrm{i}ty}}{\mathrm{i}\,t}\right| |\varphi_X(t)|\,\mathrm{d}t < \infty$. Deshalb kann die Umkehrformel (vg. Satz 6.6) als uneigentliches Integral geschrieben werden:

$$F(y) - F(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\mathrm{e}^{-\mathrm{i}tx} - \mathrm{e}^{-\mathrm{i}ty}}{\mathrm{i}\,t}\varphi_X(t)\,\mathrm{d}t.$$

Sei x_0 fest und sei h derart, daß $x = x_0 - h$, $y = x_0 + h$ Stetigkeitspunkte von F sind. Wir formen um:

$$F(x_0 + h) - F(x_0 - h) =$$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\cos(t(x_0 - h)) - \cos(t(x_0 + h)) - \mathrm{i}\,\sin(t(x_0 - h)) + \mathrm{i}\,\sin(t(x_0 + h))}{\mathrm{i}\,t} \cdot \varphi_X(t)\,\mathrm{d}t.$$

Mit

$$\left.\begin{array}{rcl} \cos(tx_0 - th) &=& \cos(tx_0)\cos(th) + \sin(tx_0)\sin(th) \\ \cos(tx_0 + th) &=& \cos(tx_0)\cos(th) - \sin(tx_0)\sin(th) \end{array}\right\} \Rightarrow (I)$$

$$\left.\begin{array}{rcl} \sin(tx_0 - th) &=& \sin(tx_0)\cos(th) - \cos(tx_0)\sin(th) \\ \sin(tx_0 + th) &=& \sin(tx_0)\cos(th) + \cos(tx_0)\sin(th) \end{array}\right\} \Rightarrow (II)$$

folgt

$$\begin{array}{rcll} \text{Realteil} &=& (I) &= 2\sin(tx_0)\sin(th) \\ \text{Imaginärteil} &=& (II) &= 2\cos(tx_0)\sin(th) \end{array}$$

$$\begin{array}{rcl} (I) + \mathrm{i}\,(II) &=& 2\sin(th)\sin(tx_0) + 2\,\mathrm{i}\,\cos(tx_0)\sin(th) \\ &=& 2\sin(th)\,(\sin(tx_0) + \mathrm{i}\,\cos(tx_0)) \\ &=& 2i\sin(th)\,(\cos(tx_0) - \mathrm{i}\,\sin(tx_0)) \\ \Rightarrow \dfrac{(I) + \mathrm{i}\,(II)}{\mathrm{i}\,t} &=& \dfrac{2}{t}\sin(th)\,\mathrm{e}^{-\mathrm{i}\,tx_0}. \end{array}$$

Wir erhalten schließlich

$$F(x_0+h)-F(x_0-h) \;=\; 2h\cdot\frac{1}{2\pi}\int\limits_{-\infty}^{\infty}\frac{\sin th}{th}\,\mathrm{e}^{-\,\mathrm{i}\,tx_0}\varphi_X(t)\,\mathrm{d}t \tag{1}$$

Wegen $|\frac{\sin th}{th}|\le 1$ haben wir

$$F(x_0+h)-F(x_0-h)\le 2h\frac{1}{2\pi}\int\limits_{-\infty}^{\infty}|\varphi_X(t)|\,\mathrm{d}t\;<\infty\;,$$

d. h. wir können den Grenzübergang vollziehen und erhalten für $h\to 0$:

$$F(x_0+0)-F(x_0-0)\;=\;0\;.$$

Damit ist gezeigt, daß F absolut stetig ist. Ferner gilt

$$\frac{F(x_0+h)-F(x_0-h)}{2h} \;=\; \frac{1}{2\pi}\int\limits_{-\infty}^{\infty}\frac{\sin th}{th}\,\mathrm{e}^{-\,\mathrm{i}\,tx_0}\varphi_X(t)\,\mathrm{d}t. \tag{2}$$

Aus dem Satz von Lebesgue folgt nun die Umkehrformel für die Dichte:

$$\begin{aligned}\lim_{h\to 0}\frac{1}{2\pi}\int\limits_{-\infty}^{\infty}\frac{\sin th}{th}\,\mathrm{e}^{-\,\mathrm{i}\,tx_0}\varphi_X(t)\,\mathrm{d}t \;&=\; \frac{1}{2\pi}\int\limits_{-\infty}^{\infty}\lim_{h\to 0}\frac{\sin th}{th}\,\mathrm{e}^{-\,\mathrm{i}\,tx_0}\varphi_X(t)\,\mathrm{d}t\\ &=\; \frac{1}{2\pi}\int\limits_{-\infty}^{\infty}\mathrm{e}^{-\,\mathrm{i}\,tx_0}\varphi_X(t)\,\mathrm{d}t\;=\;p(x_0).\end{aligned}$$

Es bleibt zu zeigen, daß $p=F'$ stetig ist. Wegen

$$\begin{aligned}p(x_0+h)-p(x_0) \;&=\; \frac{1}{2\pi}\int\limits_{-\infty}^{\infty}\left[\,\mathrm{e}^{-\,\mathrm{i}\,tx_0-\,\mathrm{i}\,th}-\,\mathrm{e}^{-\,\mathrm{i}\,tx_0}\right]\varphi_X(t)\,\mathrm{d}t\\ &=\; \frac{1}{2\pi}\int\limits_{-\infty}^{\infty}\mathrm{e}^{-\,\mathrm{i}\,tx_0}\left[\,\mathrm{e}^{-\,\mathrm{i}\,th}-1\right]\varphi_X(t)\,\mathrm{d}t\end{aligned}$$

folgt

$$|p(x_0+h)-p(x_0)| \;\le\; \varepsilon \;\text{ für }|h|\le\delta\;\text{ für ein geeignetes }\delta\;.$$

□

Grenzwertsätze für c.F.

Die Zuordnung zwischen c.F. und Verteilungsfunktionen ist nicht nur eineindeutig, sondern auch stetig.

Satz 6.9 *Direkter Grenzwertsatz* *Wenn die Folge von Verteilungsfunktionen $F_1, F_2, ...$ im wesentlichen gegen die Verteilungsfunktion F konvergiert, so konvergiert auch die Folge der entsprechenden c.F. $\varphi_1, \varphi_2, ...$ gegen die c.F. φ von F. Diese Konvergenz ist in jedem endlichen Intervall gleichmäßig.*

Das heißt also

$$\begin{aligned} F_n(x) &\to F(x) \text{ in jedem Stetigkeitspunkt von } F\ . \\ \Rightarrow\ \varphi_n(t) &= \int \mathrm{e}^{\mathrm{i}tx}\,\mathrm{d}F_n(x) \xrightarrow{\text{glm.}} \varphi(t) = \int \mathrm{e}^{\mathrm{i}tx}\,\mathrm{d}F(x) \end{aligned}$$

Satz 6.10 *Umkehrung des direkten Grenzwertsatzes* *Wenn die Folge der c.F. $\varphi_1, \varphi_2, ...$ gegen eine stetige Funktion φ konvergiert, so konvergiert die Folge der entsprechenden Verteilungsfunktionen $F_1, F_2, ...$ im wesentlichen gegen eine Verteilungsfunktion F und es ist $\varphi(t) = \int \mathrm{e}^{\mathrm{i}tx}\,\mathrm{d}F(x)$.*

Das heißt, es seien

$$F_j(x) = \frac{1}{2\pi} \lim_{z\to-\infty} \lim_{a\to\infty} \int_{-a}^{a} \frac{\mathrm{e}^{-\mathrm{i}tz} - \mathrm{e}^{-\mathrm{i}tx}}{\mathrm{i}\,t} \varphi_j(t)\,\mathrm{d}t, \quad j = 1, 2, ...$$

$\varphi_j \to \varphi$ (stetig) $\Rightarrow$ $\quad F_j \to F$ (im wesentlichen).

Beispiele für c.F.:

1. $X_j \sim \mathsf{N}(\mu_j, \sigma_j^2), unabhängig \Rightarrow$ Für $S = \sum_{j=1}^{n} X_j$ gilt:

$$\varphi_S(t) = \prod_{j=1}^{n} \mathrm{e}^{\mathrm{i}\mu_j t - \frac{1}{2}\sigma_j^2 t^2} = \mathrm{e}^{\mathrm{i}(\sum \mu_j)t - \frac{1}{2}(\sum \sigma_j^2)t^2}.$$

   Wegen der eineindeutigen Zuordnung von c.F. und Verteilungen folgt hieraus: $S \sim \mathsf{N}(\sum \mu_j, \sum \sigma_j^2)$.

2. X sei Poisson-verteilt: $\mathsf{P}(X = k) = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!}$, $k = 0, 1, 2, ...$ für $\lambda > 0$. Dann gilt

$$\begin{aligned} \varphi_X(t) &= \mathsf{E}\,\mathrm{e}^{\mathrm{i}tX} = \sum_{k=0}^{\infty} \mathrm{e}^{\mathrm{i}kt} \mathsf{P}(X = k) = \sum_{k=0}^{\infty} \mathrm{e}^{\mathrm{i}kt} \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!} \\ &= \mathrm{e}^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda\,\mathrm{e}^{\mathrm{i}t})^k}{k!} = \mathrm{e}^{-\lambda}\,\mathrm{e}^{\lambda\,\mathrm{e}^{\mathrm{i}t}} = \mathrm{e}^{\lambda(\mathrm{e}^{\mathrm{i}t} - 1)} \end{aligned}$$

Aus Satz 6.4 berechnen wir Erwartungswert und Varianz einer Poisson-verteilten Variablen:

$$\begin{aligned}
\varphi_X^{(l)}(0) &= \mathrm{i}^l \mathsf{E}(X^l), \text{d.h.} \\
\mathsf{E}X &= -\mathrm{i}\,\varphi_X^{(1)}(0), \\
\mathsf{Var}\ X &= \mathsf{E}X^2 - (\mathsf{E}X)^2 = -\varphi_X^{(2)}(0) - \left(\mathrm{i}^2 \left(\varphi_X^{(1)}(0)\right)^2\right) \\
&= -\varphi_X^{(2)}(0) + \left(\varphi_X^{(1)}(0)\right)^2 \\
\varphi_X^{(1)}(t) &= \mathrm{e}^{\lambda(\mathrm{e}^{\mathrm{i}t}-1)} \cdot \lambda \mathrm{e}^{\mathrm{i}t}\mathrm{i} = \mathrm{i}\lambda \mathrm{e}^{\mathrm{i}t} \mathrm{e}^{\lambda(\mathrm{e}^{\mathrm{i}t}-1)} \\
\varphi_X^{(2)}(t) &= \mathrm{i}\lambda \mathrm{e}^{\mathrm{i}t} \mathrm{e}^{\lambda(\mathrm{e}^{\mathrm{i}t}-1)} \cdot \mathrm{i} + \mathrm{i}\lambda \mathrm{e}^{\mathrm{i}t} \mathrm{e}^{\lambda(\mathrm{e}^{\mathrm{i}t}-1)} \cdot \lambda \mathrm{e}^{\mathrm{i}t} \cdot \mathrm{i} \\
&= -\lambda \mathrm{e}^{\mathrm{i}t} \mathrm{e}^{\lambda(\mathrm{e}^{\mathrm{i}t}-1)} [1 + \lambda \mathrm{e}^{\mathrm{i}t}]
\end{aligned}$$

$\Rightarrow\ \varphi_X^{(1)}(0) = \mathrm{i}\lambda\ ,\ \varphi_X^{(2)}(0) = -\lambda^2 + (-\lambda) = -\lambda(1+\lambda)$
$\Rightarrow\ \mathsf{E}X = \lambda\ ,\ \ \mathsf{Var}\ X = \lambda(1+\lambda) - \lambda^2 = \lambda$

Anwendung des Satzes über die Dichte:
Wenn c.F. $\varphi_X(t)$ integrierbar über $\mathbb{R}^1$, so existiert Dichte p und es gilt:

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}tx} \varphi_X(t)\,\mathrm{d}t\ .$$

Beispiel 6.2 Normalverteilung:

$$\begin{aligned}
& X \sim \mathsf{N}(0,1) \\
\Rightarrow\ & \varphi_X(t) = \mathrm{e}^{-\frac{1}{2}t^2} \\
\Rightarrow\ & \varphi_X \text{ ist integrierbar über } \mathbb{R}^1\ .
\end{aligned}$$

Hier haben wir

$$\begin{aligned}
\int_{-\infty}^{\infty} \sin(tx)\,\mathrm{e}^{-\frac{1}{2}t^2}\,\mathrm{d}t &= 0 \\
\int_{-\infty}^{\infty} \cos(tx)\,\mathrm{e}^{-\frac{1}{2}t^2}\,\mathrm{d}t &= \sqrt{2\pi}\,\mathrm{e}^{-\frac{1}{2}x^2} \\
\Rightarrow\ \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}x^2} \varphi_X(t)\,\mathrm{d}t &= \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}x^2}.
\end{aligned}$$

Charakteristische Funktionen eines zufälligen Vektors

Sei X eine p-dimensionale Zufallsgröße $\Rightarrow \quad t'X = \sum_{j=1}^{p} t_j X_j$ ist eine eindimensionale Zufallsgröße.

Definition 6.2 *Als c.F. von X definiert man*

$$\varphi_X(t) = \varphi_X(t_1, ..., t_p) = \mathsf{E}\, \mathrm{e}^{\mathrm{i}\, t'X} = \int \dots \int \mathrm{e}^{\mathrm{i} \sum_{j=1}^{p} t_j z_j} \, \mathrm{d}F(z_1, ..., z_p)$$
$$= \int \dots \int \mathrm{e}^{\mathrm{i}\, t'z} \, \mathrm{d}F(z).$$

Die Eigenschaften sind analog zu den Eigenschaften der c.F. eindimensionaler Zufallsgrößen:

$$\varphi_X(0) = 1\,, \qquad |\varphi_X(t)| \leq 1\,, \qquad \varphi_X(-t) = \overline{\varphi_X(t)}.$$

Aus der charakteristischen Funktion von X erhält man die c.F. jeder einzelnen Komponente X_s bzw. einer beliebigen Menge von Komponenten,
zum Beispiel

$$\varphi_{X_1}(t_1) = \varphi_X(t_1, 0, ..., 0).$$

Ähnlich wie im eindimensionalen Fall ist die c.F. einer p-dimensionalen Zufallsgröße im ganzen Raum $\mathbb{R}^p$ gleichmäßig stetig, außerdem gilt für $X = (X_1, ..., X_p)'$

$$\mathsf{E} X_1^{k_1} ... X_p^{k_p} = i^{\sum_{j=1}^{p} k_j} \cdot \left. \frac{\partial^{k_1 + ... + k_p} \varphi_X(t_1, ..., t_p)}{\partial t_1^{k_1} ... \partial t_p^{k_p}} \right|_{t_1 = ... = t_p = 0}.$$

Es folgt auch direkt

Satz 6.11 *Hat X die c.F. φ_X , so ist die c.F. von $(b_1 X_1 + a_1, ..., b_p X_p + a_p)$ gleich*

$$\mathrm{e}^{\mathrm{i} \sum_{j=1}^{p} a_j t_j} \varphi_X(b_1 t_1, ..., b_p t_p)$$

Satz 6.12 *Ist φ_X die charakteristische Funktion von X, so ist die c.F. der Summe $\sum_{j=1}^{p} X_j$ gleich $\varphi_X(t, ..., t)$ für $t \in \mathbb{R}^1$.*

(folgt unmittelbar aus der Definition)

Beispiel 6.3 Charakteristische Funktion einer Normalverteilung $\mathsf{N}(\mu, W)$.
Für $X \sim \mathsf{N}(\mu, W)$ gilt: $t'X$ ist eindimensional normalverteilt mit $\mathsf{N}(t'\mu, t'Wt)$.
$\Rightarrow$ Die c.F. von $t'X$ ist

$$\varphi_{t'X}(\tau) = \mathsf{E}\, \mathrm{e}^{\mathrm{i}\, \tau t'X} = \mathrm{e}^{\mathrm{i}\, \tau t'\mu - \frac{1}{2}\tau^2 t'Wt}$$

Für $\tau = 1$ ergibt sich

$$\varphi_X(t) \;=\; \mathrm{e}^{\mathrm{i}\,t'\mu - \frac{1}{2}t'\Sigma t}$$

Wichtig ist der folgende

Satz 6.13 *Die Verteilungsfunktion ist durch ihre charakteristische Funktion eindeutig bestimmt.*

BEWEIS: Wie im eindimensionalen über die Umkehrformel.

Satz 6.14 *(Umkehrformel) Ist $\varphi(t_1, ..., t_p)$ die charakteristische Funktion und $F(x_1, ..., x_p)$ die Verteilungsfunktion von $(X_1, ..., X_p) = X$, so gilt $\forall\, a, b \in \mathbb{R}^p$*

$$\mathsf{P}(X \le b) - \mathsf{P}(X \le a) = \lim_{T\to\infty} \frac{1}{(2\pi)^p} \int_{-T}^{T} \cdots \int_{-T}^{T} \prod_{j=1}^{p} \frac{\mathrm{e}^{-\mathrm{i}\,t_j a_j} - \mathrm{e}^{-\mathrm{i}\,t_j b_j}}{\mathrm{i}\,t_j} \varphi(t_1, ..., t_p)\,\mathrm{d}t_1 \ldots \mathrm{d}t_p.$$

Es gelten auch für mehrdimensionale c.F. der **direkte GWS** und die **Umkehrung des direkten GWS**.

Satz 6.15 *Die Komponenten von X sind unabhängig genau dann, wenn die charakteristische Funktion von X das Produkt der charakteristischen Funktionen der Komponenten ist, d.h.*

$$\varphi_X(t_1, ..., t_p) \;=\; \prod_{j=1}^{p} \varphi_{X_j}(t_j).$$

BEWEIS:

$(\Rightarrow)$ $\mathsf{E}\,\mathrm{e}^{\mathrm{i}\,t'X} \;=\; \mathsf{E} \prod_{j=1}^{p} \mathrm{e}^{\mathrm{i}\,t_j X_j} \;=\; \prod_{j=1}^{p} \mathsf{E}\,\mathrm{e}^{\mathrm{i}\,t_j X_j}.$

$(\Leftarrow)$ Sei $\varphi_X(t_1, ..., t_p) \;=\; \prod_{j=1}^{p} \varphi_{X_j}(t_j)$

Aus der Umkehrformel folgt

$$\begin{aligned} \mathsf{P}(X \le b) \;&=\; \lim_{a\to-\infty} \lim_{T\to\infty} \frac{1}{(2\pi)^p} \prod_{j=1}^{p} \int_{-T}^{T} \frac{\mathrm{e}^{-\mathrm{i}\,t_j a_j} - \mathrm{e}^{-\mathrm{i}\,t_j b_j}}{\mathrm{i}\,t_j} \varphi_{X_j}(t_j)\,\mathrm{d}t_j \\ &=\; \lim_{a\to-\infty} \prod_{j=1}^{p} [F_j(b_j) - F_j(a_j)] \;=\; \prod_{j=1}^{p} F_j(b_j). \end{aligned}$$

□

Anwendung für die Normalverteilung:

Satz 6.16 *Die Komponenten eines normalverteilten Vektors sind unabhängig genau dann, wenn sie unkorreliert sind.*

BEWEIS: o. B. d. A. $\mathsf{E}X_i = 0$. Für $X \sim \mathsf{N}(0, W)$ ist $W = diag(\sigma_{11}, ..., \sigma_{nn})$ und die c.F. von X ist das Produkt der charakteristischen Funktionen der X_j.

□

Zentraler Grenzwertsatz für unabhängig identisch verteilte Zufallsgrößen

Satz 6.17 *Es seien $X_1, X_2, ...$ unabhängige und identisch verteilte Zufallsgrößen mit $\mathsf{E}X_i = \mu$, $\mathsf{Var}\ X_i = \sigma^2 < \infty$. Dann gilt für die standardisierte Summe*

$$V_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}.$$

für $S_n = \sum_{i=1}^{n} X_i$:

$$\mathsf{P}(V_N \le x) \ \to \ \Phi(x) \qquad \textit{gleichmäßig in } x,$$

wobei Φ die Verteilungsfunktion der $\mathsf{N}(0,1)$-Verteilung ist.

BEWEIS: Sei φ_0 die c.F. einer $\mathsf{N}(0,1)$-verteilten Variablen, d.h. $\varphi_0(t) = \mathrm{e}^{-\frac{t^2}{2}}$. Es genügt zu zeigen: $\varphi_{V_n}(t) \to \varphi_0(t)$.
Wir haben

$$\varphi_{V_n}(t) \ = \ \left[\varphi_{X_1-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

$\varphi_{X_1-\mu}$ besitzt eine Taylorreihenentwicklung:

$$\varphi_{X_1-\mu}(t) \ = \ 1 + \sum_{j=1}^{2} \frac{t^j}{j!}\varphi^{(j)}_{X_1-\mu}(0) \ + o(t^2) \ = \ 1 - \frac{t^2\sigma^2}{2} + o(t^2),$$

wobei $\varphi^{(j)}_{X_1-\mu}(0) = \mathrm{i}^j \mathsf{E}(X_1 - \mu)^j$. Damit erhält man

$$\varphi_{V_n}(t) \ = \ \left(1 - \frac{t^2}{2n} + o(\frac{t^2}{n})\right)^n.$$

Ferner gilt

$$\ln\left(1 - \frac{t^2}{2n} + o(\frac{t^2}{n})\right)^n \ = \ n\ln\left(1 - \frac{t^2}{2n} + o(\frac{t^2}{n})\right) \to -\frac{t^2}{2},$$

d.h. $\varphi_{V_n}(t) \to \varphi_0(t)$.

□

7 Entscheidungsfunktionen

7.1 Entscheidungsprobleme

Untersucht werden in der mathematischen Statistik statistische Entscheidungen, d.h. Aussagen über die Verteilung der Zufallsgrößen Y aufgrund einer Beobachtung $y = Y(w)$. Es sei Δ die Menge der Entscheidungen, mit der σ-Algebra $\widetilde{\mathfrak{D}}$ sei $(\Delta, \widetilde{\mathfrak{D}})$ ein meßbarer Raum. Dieser Raum heißt Entscheidungsraum.

Definition 7.1 *Unter einer nichtrandomisierten Entscheidungsfunktion (EF) versteht man eine meßbare Abbildung* $(Y, \mathfrak{C}) \to (\Delta, \widetilde{\mathfrak{D}})$.

Beispiele für Entscheidungsprobleme

7.1.1 Testen von Hypothesen

Allgemeine Problemstellung:
Das statistische Modell sei $Y \sim \mathsf{P}_\vartheta$, $\vartheta \in \Theta$. Man formuliert Annahmen ber den Parameter ϑ:

$$\begin{array}{lll} \text{Hypothese} & Y \sim \mathsf{P}_\vartheta & \text{für} \quad \vartheta \in \Theta_H \subset \Theta \\ \text{Alternative} & Y \sim \mathsf{P}_\vartheta & \text{für} \quad \vartheta \in \Theta_K = \Theta \backslash \Theta_H \ . \end{array}$$

Die Hypothese ist eine Aussage über eine angenommene Verteilung von Y , ausgedrückt durch die Lage des Parameters ϑ.
Kurz schreibt man:

$$H : \ \vartheta \in \Theta_H \quad \text{gegen} \quad K : \ \vartheta \in \Theta_K = \Theta \backslash \Theta_H .$$

Der Stichprobenraum ist $(\mathcal{Y}, \mathfrak{C})$, der Entscheidungsraum ist $(\Delta, \tilde{\mathfrak{D}})$ für $\Delta = \{d_H, d_K\}$: d_H Entscheidung für H, d_K Entscheidung für K. Man wählt $\tilde{\mathfrak{D}} = \mathfrak{P}(\Delta)$ die Potenzmenge von Δ. Wird die Entscheidung zwischen H und K durch eine Entscheidungsfunktion T gegeben, so gilt $T(y) = d_H$ oder $T(y) = d_K$.

Definition 7.2 *Eine meßbare Funktion* $T : (\mathcal{Y}, \mathfrak{C}) \to (\Delta, \tilde{\mathfrak{D}})$ *heißt Testfunktion.*

Eine Testfunktion ist beschrieben durch

$$T(y) \ = \ \begin{cases} d_K & \text{für} \quad y \in S \\ d_H & \quad\ \ sonst \end{cases}$$

S heißt Ablehnungsgebiet oder kritischer Bereich. Man hat S bzw. T so zu wählen, daß mglichst wenig Fehlentscheidungen entstehen. Dabei gibt es zwei mgliche Fehlentscheidungen:
Fehler 1. Art: Entscheidung für K, obwohl H richtig ist.
Fehler 2. Art: Entscheidung für H, obwohl K richtig ist.

Wesentlich sind die Wahrscheinlichkeiten für diese Fehler.
Wk für Fehler 1.Art = Irrtumswahrscheinlichkeit 1. Art (Signifikanzniveau)

$$\mathsf{P}_\vartheta(T(y) = d_K),\ \vartheta \in \Theta_H.$$

Wk für Fehler 2. Art = Irrtumswahrscheinlichkeit 2. Art

$$1 - \mathsf{P}_\vartheta(T(y) = d_K),\ \vartheta \in \Theta_K.$$

Damit sind die Irrtumswahrscheinlichkeiten beschrieben durch die Gütefunktion

$$\beta_T(\vartheta) := \mathsf{P}_\vartheta(T(y) = d_K)\,.$$

Definition 7.3 *Zu vorgegebenem α, $0 < \alpha < 1$, heißt eine Testfunktion T α–Testfunktion oder kurz α–Test, falls $\beta_T(\vartheta) \leq \alpha$ für alle $\vartheta \in \Theta_H$ gilt.*

$\Rightarrow$ Wenn T ein α-Test ist, so ist die Entscheidung für K, obwohl H richtig ist, nur in $100\%\cdot\alpha$ der Fälle zu erwarten. Damit ist die Entscheidung für K statistisch abgesichert. Für α nimmt man kleine Werte : α=0.05, 0.01 . Anstrebenswert sind α-Tests mit kleiner Irrtumswahrscheinlichkeit 2. Art. Ob es möglich ist, solche Tests zu finden, soll später untersucht werden.

Es gibt noch eine weitere Beschreibung für Testfunktionen. Beschreibt zu einer Funktion φ der Wert $\varphi(y)$ die Wahrscheinlichkeit, sich für d_K zu entscheiden, und ist $\varphi : (\mathcal{Y}, \mathfrak{C}) \xrightarrow{\varphi} ([0,1], \mathfrak{B} \cap [0,1])$ meßbar, so heißt eine solche Funktion φ Test. Es gibt eine einfache Beziehung zwischen T und φ bei nichtrandomisierten Tests:

$$T(y) = d_K \Leftrightarrow \varphi(y) = 1$$
$$T(y) = d_H \Leftrightarrow \varphi(y) = 0$$

Berechnung der Gütefunktion:

$$\beta_T(\vartheta) = \mathsf{P}_\vartheta\,(T(y) = d_K) = \mathsf{P}_\vartheta\,(\varphi(y) = 1) = \mathsf{E}_\vartheta\varphi(Y).$$

Es gilt stets $0 \leq \mathsf{E}_\vartheta\varphi(Y) \leq 1$.
Ein weiterer wichtiger Begriff für Tests ist die Unverfälschtheit.

Definition 7.4 *Ein α-Test heißt unverfälscht, falls*

$$\mathsf{E}_\vartheta\varphi(Y) \geq \alpha \quad \forall\ \vartheta \in \Theta_K \text{ ist.}$$

Das heißt, man entscheidet sich öfter für K, wenn K richtig ist als wenn H richtig ist.

Beispiel 7.1 (Gaußtest) $Y_1, \ldots, Y_n \sim \mathsf{N}(\vartheta, \sigma_0^2)$ i.i.d., σ_0^2 bekannt.
Sei $H : \vartheta \leq \vartheta_0$, $K : \vartheta > \vartheta_0$, ϑ_0 gegeben. Dieses Testproblem H gegen K heißt einseitiges Testproblem. Beobachtet wird $y = (y_1, \ldots, y_n)$, eine Stichprobe vom Umfang n. Bei Aussagen über den Erwartungswert ϑ zu $\mathsf{N}(\vartheta, \sigma_0^2)$ benutzt man aus heuristischen Gründen die Statistik $\bar{y}$, es gilt:

$$\bar{Y} \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i \sim \mathsf{N}\left(\vartheta, \frac{1}{n}\sigma_0^2\right)$$

$$\Rightarrow T(Y) \quad = \quad \frac{\sqrt{n}}{\sigma_0}(\bar{Y} - \vartheta_0) \sim \mathsf{N}\left(\sqrt{n}\frac{\vartheta - \vartheta_0}{\sigma_0}, 1\right)$$

Gesucht ist ein α-Test der folgenden Gestalt (heuristisch nahegelegt):

$$\varphi_c(y) \quad = \quad \begin{cases} 1\,, & \text{falls} \quad T(y) > c \\ 0\,, & \text{falls} \quad T(y) < c\,. \end{cases}$$

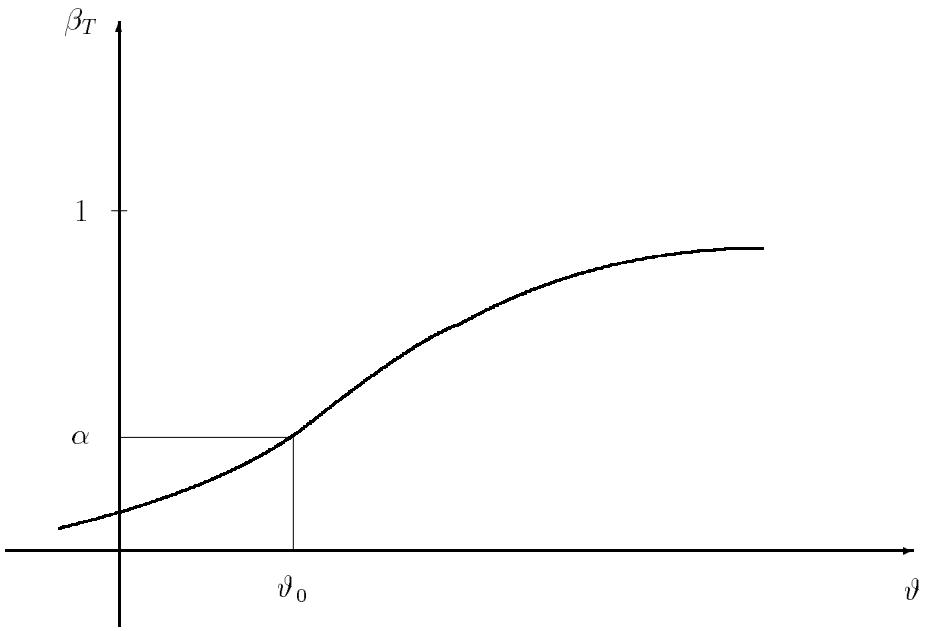
Dann gilt für $\mathsf{P}_\vartheta = \mathsf{P}^Y$:

$$\begin{aligned} \mathsf{P}_\vartheta(T(y) > c) \quad &= \quad \frac{1}{\sqrt{2\pi}} \int_c^\infty e^{\frac{-(x - \sqrt{n}\frac{\vartheta - \vartheta_o}{\sigma_o})^2}{2}}\, \mathrm{d}x = 1 - \phi\left(c - \sqrt{n}\frac{\vartheta - \vartheta_o}{\sigma_o}\right) \quad \text{mit} \\ \phi(z) \quad &= \quad \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{x^2}{2}}\, \mathrm{d}x \\ \Rightarrow \beta_T(\vartheta) \quad &= \quad \mathsf{P}_\vartheta(T(y) > c) \leq 1 - \phi(c) \quad \text{bei} \quad \vartheta \leq \vartheta_o \\ &\Rightarrow \quad \text{Falls} \quad 1 - \alpha = \phi(c_\alpha) \quad \text{so ist} \quad \varphi_{c_\alpha} \quad \text{ein} \quad \alpha - Test. \end{aligned}$$

c_α heißt α–Fraktil von $\mathsf{N}(0,1)$. Ist φ_{c_α} unverfälscht? Dann muß gelten $\mathsf{P}_\vartheta(T(y) > c_\alpha) \geq \alpha$ für $\vartheta \geq \vartheta_0$. Wir haben $\mathsf{P}_\vartheta(T(y) > c_\alpha) \quad = \quad 1 - \phi\left(c_\alpha - \sqrt{n}\frac{\vartheta - \vartheta_o}{\sigma_o}\right)$ und die Unverfälschtheit bedeutet dann

$$1 - \alpha \geq \phi\left(c_\alpha - \underbrace{\sqrt{n}\frac{\vartheta - \vartheta_o}{\sigma_o}}_{\geq 0}\right)$$

Das ist erfüllt.



Beispiel 7.2 (Zweiseitiger Gaußtest) Y sei eine zufällige Stichprobe vom Umfang n mit $Y_i \sim \mathsf{N}(\vartheta, \sigma_o^2)$. Zweiseitiges Testproblem:
$H : \vartheta = \vartheta_o$ gegen $K : \vartheta \neq \vartheta_o$.
Wesentlich ist hier also der Abstand von ϑ zu ϑ_o, d.h. es werden Aussagen über $\mid \vartheta - \vartheta_o \mid$ verlangt.

$\Rightarrow$ betrachte Statistik $\mid T(y) \mid$ für $T(y)$ aus Beispiel 7.1 Wir entscheiden uns für K, falls $\mid T(y) \mid > c$. Sei $c_{\alpha/2}$ das $\frac{\alpha}{2}$-Fraktil von $\mathsf{N}(0,1)$ und

$$\varphi_c(y) = \begin{cases} 1 & \text{für} \quad \mid T(y) \mid > c \\ 0 & \text{sonst .} \end{cases}$$

Es gilt

$$\begin{aligned} \mathsf{P}_\vartheta(\mid T(y) \mid > c_{\alpha/2}) = \mathsf{E}_\vartheta \varphi_{c/\alpha}(Y) &= \mathsf{P}_\vartheta\big(T(y) < -c_{\alpha/2} \quad \text{oder} \quad T(y) > c_{\alpha/2}\big) \\ &= \phi\left(-c_{\alpha/2} - \sqrt{n}\frac{\vartheta - \vartheta_o}{\sigma_o}\right) + 1 - \phi\left(c_{\alpha/2} - \sqrt{n}\frac{\vartheta - \vartheta_o}{\sigma_o}\right) \\ \Rightarrow \mathsf{P}_{\vartheta_o}(\mid T(y) \mid > c_{\alpha/2}) &= \phi(-c_{\alpha/2}) + 1 - \phi(c_{\alpha/2}) = \alpha. \end{aligned}$$

Damit ist $\varphi_{c_{\alpha/2}}$ ein α–Test. Außerdem gilt

$$\mathsf{P}_\vartheta(\mid T(y) \mid > c_{\alpha/2}) > \alpha \quad \text{für} \quad \vartheta \neq \vartheta_o.$$

$\Rightarrow \varphi_{c_{\alpha/2}}$ ist unverfälscht.

Die speziell formulierten Hypothesen und Alternativen haben einen entscheidenden Einfluß auf den zu wählenden Test.
Die Statistiken $T(y)$ bzw $\mid T(y) \mid$ heißen Prüfgröße, c kritischer Wert.

7.1.2 Mehrentscheidungsverfahren

Sei $Y \sim \mathsf{Q} \in \mathcal{P} = \{\mathsf{P}_\vartheta : \vartheta \in \Theta\}$. Soll eine Entscheidung zwischen k sich einander ausschließenden Aussagen getroffen werden, so sprechen wir von einem Mehrentscheidungsverfahren. Dem entspricht eine Zerlegung von Θ in k disjunkte Teilmengen $\Theta_1, ..., \Theta_k$.

Hypothesen	$H_1 : \vartheta \in \Theta_1$	
	.	
	.	EF d hat die Werte $d_1, ..., d_k$ (d_i Entscheidung für H_i).
	.	
	$H_k : \vartheta \in \Theta_k$	

Beispiel 7.3 Diskriminanzanalyse

Grenzfälle:	$k = 2$	klassisches Testproblem
	$k > 2$	Mengenschätzproblem
	$k \to \infty$,	Θ_i beliebig klein, so näherungsweise Aussagen über die einzelnen Parameterwerte.

7.1.3 Punktschätzungen

Die betrachteten Testprobleme können auch umformuliert werden. Sei z.B. $\gamma(\vartheta) = I_{\Theta_K}(\vartheta)$. Dann gilt
$H : \gamma(\vartheta) = 0$
$K : \gamma(\vartheta) = 1$.

Man hat eine Funktion $\gamma : \Theta \to \mathbb{R}^1$ und es ist der Wert $\gamma(\vartheta)$ zu bestimmen. Verallgemeinert man diese Situation, so sei eine beliebige Funktion $\gamma : \Theta \to \Gamma$ gegeben. Zum Beispiel könnte $\mathcal{P} = \{\mathsf{N}(\mu, W) : \mu \in \mathbb{R}^n, W \in \mathcal{M}_n^{\geq}\}$ sein. Dann sind etwa $\gamma(\vartheta) = \mu, \gamma(\vartheta) = W, \gamma(\vartheta) = (\mu, W)$ mögliche Funktionen γ. $\gamma(\vartheta)$ ist zu bestimmen. Das könnte man als ein Testproblem mit unendlich vielen Hypothesen formulieren. Einfacher ist es in diesem Fall, die Entscheidungen in den Raum $\gamma(\Theta)$ zu legen.

Definition 7.5 *Eine meßbare Abbildung $\hat{\gamma} : (\mathcal{Y}, \mathfrak{C}) \to (\Gamma, \mathfrak{G})$ heißt Schätzfunktion oder genauer Punktschätzfunktion.*

$\hat{\gamma}$ ist eine Schätzfunktion bzw. ein Schätzer, $\hat{\gamma}(y)$ ist die Schätzung oder der Schätzwert, $\hat{\gamma}(Y)$ heißt Schätzvariable.
Eigenschaften von Schätzern:

Definition 7.6 *$\hat{\gamma}$ heißt erwartungstreu (unverzerrt) für $\gamma(\vartheta)$, falls*

$$\mathsf{E}_\vartheta \hat{\gamma}(Y) = \gamma(\vartheta), \qquad \forall \vartheta \in \Theta$$

gilt.

$$\mathsf{E}_\vartheta \hat{\gamma}(Y) - \gamma(\vartheta) = b_\vartheta(\hat{\gamma}) \quad \textit{heißt Verzerrung von } \hat{\gamma}\,.$$

Falls $\hat{\gamma}(Y)$ eine stetige Verteilung besitzt, so wird ein einzelner Punkt mit der Wahrscheinlichkeit 0 angenommen. $\Rightarrow$ Mit Wahrscheinlichkeit 1 kommt ein falscher Schätzwert heraus. Die Bewertung eines Schätzers erfolgt deshalb nicht realisierungsweise sondern z.B. mit Hilfe der Verzerrung oder der Varianz. Dabei ist die Varianz (für $l = 1$) oder die Kovarianzmatrix (für $l > 1$) in folgender Weise definiert:

$$\begin{aligned} l = 1: \quad \mathsf{Var}_\vartheta \hat{\gamma} &= \mathsf{E}_\vartheta (\hat{\gamma}(Y) - \mathsf{E}_\vartheta \hat{\gamma}(Y))^2, \\ l > 1: \quad \mathsf{Var}_\vartheta \hat{\gamma} &= \mathsf{E}_\vartheta (\hat{\gamma}(Y) - \mathsf{E}_\vartheta \hat{\gamma}(Y))(\hat{\gamma}(Y) - \mathsf{E}_\vartheta \hat{\gamma}(Y))'. \end{aligned}$$

Schön wäre: $\hat{\gamma}$ ist erwartungstreu mit sehr kleiner Varianz. Das ist im allgemeinen nicht möglich. Stattdessen bewertet man eine Schätzung durch die mittleren quadratische Abweichung

$$\mathsf{MSE}_\vartheta \hat{\gamma} = \mathsf{E}_\vartheta (\hat{\gamma}(Y) - \gamma(\vartheta))(\hat{\gamma}(Y) - \gamma(\vartheta))'.$$

Hierfür gilt

$$\mathsf{MSE}_\vartheta \hat{\gamma} = \mathsf{Var}_\vartheta \hat{\gamma} + b_\vartheta(\hat{\gamma}) b_\vartheta(\hat{\gamma})'\,.$$

Die Bedeutung der erwartungstreuen Schätzfunktionen ist unbestritten, im folgenden Beispiel zeigt sich, daß es auch sinnlose erwartungstreue Schätzfunktionen gibt.

Beispiel 7.4 In der Qualitätskontrolle wird zur Bestimmung des Anteils defekter Produkte in einer Charge die Wahrscheinlichkeit für das erste Auftreten eines fehlerhaften Stückes in einer zu prüfenden Menge beobachtet. Die Variable mit dieser Wahrscheinlichkeit hat eine geometrische Verteilung.

$$\mathsf{P}(Y = r) = (1 - \vartheta)^{r-1}\vartheta \quad , \quad r = 1, 2, \ldots.$$

Offenbar gilt $\sum_{r=1}^{\infty} \mathsf{P}(Y = r) = 1$. $\gamma(\vartheta) = \vartheta$ soll geschätzt werden. $\hat{\gamma}(Y)$ sei eine erwartungstreue Schätzfunktion. Das bedeutet

$$\begin{aligned} \mathsf{E}\hat{\gamma}(Y) &= \sum_{r=1}^{\infty} \hat{\gamma}(r)(1 - \vartheta)^{r-1}\vartheta = \vartheta \quad \forall \vartheta, \quad \text{d.h.} \\ &\sum_{r=1}^{\infty} \hat{\gamma}(r)(1 - \vartheta)^{r-1} = 1. \end{aligned}$$

Durch Koeffizientenvergleich beider Reihen ergibt sich notwendig

$$\hat{\gamma}(1) = 1, \hat{\gamma}(r) = 0 \quad \text{für} \quad r = 2, \ldots.$$

$\Rightarrow$ Es existiert genau eine erwartungstreue Schätzfunktion, $\hat{\gamma}$ ist diese einzige.
$\hat{\gamma}(y)$ soll eigentlich eine Aussage über ϑ machen. Das ist hier nicht der Fall, so daß $\hat{\gamma}$ eine ungeeignete Schätzfunktion ist.

7.2 Entscheidungsprinzipien

Man benötigt Verfahren und Prinzipien zur Konstruktion von optimalen Entscheidungsfunktionen. Es sei

$$\mathcal{P} = \{\mathsf{P}_\vartheta : \vartheta \in \Theta\} \quad , \quad Y \sim \mathsf{Q} \in \mathcal{P}.$$

P_ϑ sind Maße über $(\mathcal{Y}, \mathfrak{C})$. Gesucht sind Aussagen und Informationen über Q. Sei $\mathsf{Q} = \mathsf{P}_{\vartheta_o}$ und sei ϑ_o unbekannt. Wir betrachten zwei Möglichkeiten:

a) Über ϑ_o ist nichts bekannt. Alle Werte von Θ sind gleichermaßen möglich.

b) ϑ_o ist Realisierung einer zufälligen Variablen mit einer Verteilung $\tilde{\mathsf{Q}}$. $\tilde{\mathsf{Q}}$ heißt a-priori-Verteilung. Kenntnis über $\tilde{\mathsf{Q}}$ kann aus früheren Erfahrungen herrühren. Dann entsteht die Aufgabe, mit Y eine Aussage über die Realisierung der Variablen mit der Verteilung $\tilde{\mathsf{Q}}$ herzuleiten.

7.2.1 Maximum-Likelihood-Prinzip

$\mathcal{P}$ sei durch das σ-endliche Maß μ dominiert, d.h. jedes P_ϑ ist μ-stetig. Nach dem Satz von Radon-Nikodym existiert eine μ-Dichte für beliebiges $\mathsf{P} \in \mathcal{P}$:

$$\mathsf{P}(A) \quad = \quad \int\limits_A p(x)\mu(\,\mathrm{d}x)\,.$$

Definition 7.7 *L heißt Likelihoodfunktion von $\mathcal{P}$, falls*

$$L(p; y) = p(y) \qquad \forall \mathsf{P} \in \mathcal{P}.$$

Hier ist y eine Realisierung von Y, y wird beobachtet. $\Rightarrow$ Die plausibelste Verteilung aus $\mathcal{P}$ zu dieser Beobachtung ist diejenige, für die $L(.; y)$ maximal ist, d.h. gilt

$$L(\hat{p}; y) = \max\left\{L(p; y); p = \frac{\mathrm{d}\mathsf{P}}{\mathrm{d}\mu}, \mathsf{P} \in \mathcal{P}\right\}\,, \qquad (*)$$

so ist $\hat{\mathsf{P}}$ die plausibelste Verteilung $\left(\frac{\mathrm{d}\hat{\mathsf{P}}}{\mathrm{d}\mu} = \hat{p}\right)$. Hier geht es um die Maximierung von L, deshalb sind monotone Transformationen erlaubt. Die Funktion

$$l(p; y) = \ln \quad L(p; y)$$

heißt log-Likelihood Funktion von $\mathcal{P}$.

Definition 7.8 *$\hat{p}$ aus (*) heißt Maximum-Likelihood-Schätzung (MLS). Gilt $\hat{p} = p_{\hat{\vartheta}}$, so ist $\hat{\vartheta}$ MLS für ϑ.*

Sei $Y = (Y_1, ..., Y_n)'$ und seien Y_i unabhängig mit der Dichte p_i. Dann gilt $p(y) = p_1(y_1)...p_n(y_n)$ und

$$\begin{aligned} L(p; y) &= \prod_{i=1}^{n} p_i(y_i). \\ \Rightarrow l(p; y) &= \sum_{i=1}^{n} \ln p_i(y_i) \end{aligned}$$

Falls Y_i außerdem identisch verteilt sind, so gilt

$$l(p; y) = \sum_{i=1}^{n} \ln \tilde{p}(y_i)$$

für $p_i = \tilde{p}$.

Beispiel 7.5 Seien $Y_1, ..., Y_n$ unabhängig, doppelt exponentiell verteilt, d.h.

$$\begin{aligned} p_i(y_i) &= \frac{a}{2} e^{-a|y_i - b|}, \qquad a > 0. \\ \Rightarrow p(y) &= \left(\frac{a}{2}\right)^n e^{-a \sum_{i=1}^{n} |y_i - b|}, \qquad \vartheta = (a, b) \quad \in \quad \mathbb{R}^+ \times \mathbb{R}^1 \\ \Rightarrow l(p; y) &= n \ln \left(\frac{a}{2}\right) - a \sum_{i=1}^{n} \mid y_i - b \mid \end{aligned}$$

Sei a fest und bekannt. Gesucht wird eine MLS $\hat{b}$ für b. Dann ist $\vartheta = b$ und $\hat{b}$ ist bestimmt durch

$$\sum_{i=1}^{n} \mid y_i - \hat{b} \mid = \min_b \sum_{i=1}^{n} \mid y_i - b \mid$$

Lösung: $\hat{b}(y) = med\{y_1, ..., y_n\}$, der Median von $y_1, ..., y_n$.

BEWEIS: Sei $f(b) = \sum_{i=1}^{n} \mid y_i - b \mid$. Dann gilt $f(b) = \sum_{i=1}^{n} \mid y_{[i]} - b \mid$ für die geordnete Statistik $y_{[\,]} = (y_{[1]}, ..., y_{[n]})'$. Für $b \in (y_{[j]}, y_{[j+1]})$ unter $y_{[j]} < y_{[j+1]}$ gilt dann $f(b) = \sum_{i=1}^{j} (b - y_{[i]}) + \sum_{i=j+1}^{n} (y_{[i]} - b)$ und das ist linear in b. Wegen der Stetigkeit von f und der Differenzierbarkeit in den offenen Intervallen folgt $f'(b) = j - (n - j) = 2j - n$. Daher ist

$$\hat{b}(y) = \begin{cases} y_{[\frac{n+1}{2}]} & \text{für} \quad n \quad \text{ungerade} \\ \frac{1}{2}(y_{[n]} + y_{[\frac{n+2}{2}]}) & \text{für} \quad n \quad \text{gerade} \end{cases}$$

die Lösung.

□

Beispiel 7.6 Wir betrachten die geometrische Verteilung

$$\mathsf{P}(Z = r) = (1-\vartheta)^{r-1}\vartheta, \quad \vartheta \in (0,1), r = 1, 2, ...n$$

Die MLS $\hat{\vartheta}$ ergibt sich aus

$$\max_{\vartheta}(1-\vartheta)^{r-1}\vartheta = (1-\hat{\vartheta})^{r-1}\hat{\vartheta}$$

für $\hat{\vartheta}(r) = \frac{1}{r}$. Verglichen mit der einzigen erwartungstreuen Schätzung aus Beispiel 7.4 ist diese MLS viel sinnvoller.

Beispiel 7.7 Es gelte $Y \sim \mathsf{N}_n(X\beta, \sigma^2\Sigma), \Sigma \in \mathcal{M}_n^>$ bekannt, ($\mathcal{M}_n^>$ bezeichnet die Menge aller positiv definiten $n \times n$ Matrizen), $X : n \times k$ bekannt, $\beta \in \mathbb{R}^k, \sigma^2$ seien unbekannt. Die Dichte von Y ist

$$p_{\beta,\sigma^2}(y) = \frac{1}{(2\pi\sigma^2)^{n/2} \mid \Sigma \mid^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'\Sigma^{-1}(y - X\beta)\right)$$

Für welche β, σ^2 ist $p_{\beta,\sigma^2}(y)$ maximal ? Sei zunächst σ^2 beliebig, fest. Dann führt die Maximierung von $p_{\beta,\sigma^2}(y)$ zur Minimierung von $(y - X\beta)'\Sigma^{-1}(y - X\beta)$ bzgl. β. Wir berechnen

$$\frac{\mathrm{d}}{\mathrm{d}\beta}(y - X\beta)'\Sigma^{-1}(y - X\beta) \;=\; 2X'\Sigma^{-1}X\beta - 2X'\Sigma^{-1}y$$

$\Rightarrow$ Die Lösung $\hat{\beta}$ der Minimierungsaufgabe erfüllt $X'\Sigma^{-1}X\hat{\beta} = X'\Sigma^{-1}y$ und ist unabhängig von σ^2.

Zu diesem $\hat{\beta}$ ist ein σ^2 zu finden, so daß $p_{\hat{\beta},\hat{\sigma^2}} = max$. Sei $s^2 = (y - X\hat{\beta})\Sigma^{-1}(y - X\hat{\beta})$.

$$\begin{aligned}
\Rightarrow\; l(p_{\hat{\beta},\sigma^2}; y) &= -\frac{1}{2}\ln\left((2\pi)^n \mid \Sigma \mid\right) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}s^2 \\
\frac{\mathrm{d}}{\mathrm{d}\sigma^2} l(p_{\hat{\beta},\sigma^2}; y) &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{s^2}{2}\frac{1}{\sigma^4} = \frac{1}{2\sigma^2}\left(-n + \frac{s^2}{\sigma^2}\right) \\
\Rightarrow p_{\hat{\beta},\sigma^2} &= \max \quad \text{für} \quad \hat{\sigma^2} = \frac{1}{n}s^2
\end{aligned}$$

$\Rightarrow$ MLS $\hat{\beta}$ $\hat{\sigma}^2$ sind gegeben durch:

$$\begin{aligned}
\hat{\beta}(y) &= \left(X'\Sigma^{-1}X\right)^{-1} X\Sigma^{-1}y, \\
\hat{\sigma}^2(y) &= \frac{1}{n}(y - X\hat{\beta}(y))'\Sigma^{-1}(y - X\hat{\beta}(y))
\end{aligned}$$

Als Erwartungswerte für $\hat{\beta}(Y)$, $\hat{\sigma}^2(Y)$ erhält man

$$\mathsf{E}_{\beta,\sigma^2}\hat{\beta}(Y) = \beta, \quad \mathsf{E}_{\beta,\sigma^2}\hat{\sigma}^2(Y) = \frac{n-k}{n}\sigma^2.$$

Damit ist $\hat{\beta}$ erwartungstreu. $\hat{\sigma}^2$ unterschätzt σ^2 im Mittel. In Abhängigkeit von k muß die MLS $\hat{\sigma}^2$ verändert werden, um ein brauchbarer Schätzer zu sein.

Das wird an folgender Situation deutlich. Sei $n = 40, k = 20$. Dann sind 99,5% aller Werte $\hat{\sigma}^2(y)$ kleiner als σ^2 und folglich sind fast alle beobachtbaren Schätzwerte $\hat{\sigma}^2(y)$ wesentlich kleiner als der zu schätzende Parameter. Das zeigt an, daß die MLS nicht immer sinnvolle Schätzfunktionen sein müssen.

In diesen drei Beispielen zeigte sich, daß die MLS explizit angegeben werden können. In allen Fällen sind die Lösungen durch meßbare Funktionen gegeben und folglich kann man dann von der statistischen Entscheidungsfunktion MLS sprechen. Es bleibt offen, ob im allgemeinen die MLS eine statistische Entscheidungsfunktion ist. Das kann in wichtigen Fällen nahezu bejaht werden. Grundlage bildet das folgende Lemma.

Lemma 7.1 *Sei $\Theta \subseteq \mathbb{R}^k, g : \mathcal{Y} \times \Theta \to \mathbb{R}^l$. Sei g bei festem ϑ meßbar, bei festem y stetig in Θ. Falls Θ kompakt ist und eine Abbildung $\eta : \mathcal{Y} \to \Theta$ derart existiert, daß $h(y) := g(y, \eta(y))$ meßbar ist, so existiert eine meßbare Abbildung $\tilde{\vartheta} : (\mathcal{Y}, \mathfrak{C}) \to (\Theta, \Theta \cap \mathfrak{B}^k)$ mit*

$$h(y) = g(y, \tilde{\vartheta}(y)).$$

Als Folgerung erhalten wir eine Aussage zur MLS.
Folgerung: Falls $p_\vartheta(y)$ bei festem y stetig in Θ ist, so kann bei kompaktem $\Theta \subset \mathbb{R}^k$ die MLS so gewählt werden, daß sie meßbar und somit eine statistische Entscheidungsfunktion ist.

BEWEIS: Da $p_\vartheta(y)$ bei festem ϑ meßbar ist, ist auch $h(y) := \sup_{\vartheta \in \Theta} p_\vartheta(y)$ meßbar. Folglich existiert ein meßbares $\tilde{\vartheta}$ mit $h(y) = p_{\tilde{\vartheta}}(y)$ und $\tilde{\vartheta}$ ist damit eine MLS und eine statistische Entscheidungsfunktion.

□

Bemerkung: Damit ist noch nicht gesagt, daß jede MLS eine meßbare Funktion ist. Aber man kann unter den nicht eindeutigen maximalen Werten einen derartigen finden, daß die MLS statistische Entscheidungsfunktionen sind.

Desweiteren bleibt zu bemerken, daß die MLS unabhängig von der speziellen Wahl des dominierenden Maßes ist. Das beweist man folgendermaßen. Sei $p_\vartheta(y) = \frac{\mathrm{d}\mathsf{P}_\vartheta}{\mathrm{d}\mu} \quad [\mathfrak{C}, \mu]$. Für ein zweites dominierendes Maß ν sei

$$q_\vartheta(y) = \frac{\mathrm{d}\mathsf{P}_\vartheta}{\mathrm{d}\nu} \quad [\mathfrak{C}, \nu]$$

Wir wählen ρ als ein zu $\mathcal{P}$ äquivalentes Maß. Dann gilt $r_\vartheta(y) = \frac{\mathrm{d}\mathsf{P}_\vartheta}{\mathrm{d}\rho} \quad [\mathfrak{C}, \rho]$ und $\frac{\mathrm{d}\rho}{\mathrm{d}\mu} = k(y) \quad [\mathfrak{C}, \mu], \frac{\mathrm{d}\rho}{\mathrm{d}\nu} = l(y) \quad [\mathfrak{C}, \nu]$. Daraus folgt

$$p_\vartheta(y) = r_\vartheta(y) \cdot k(y) \quad [\mathfrak{C}, \mu] \quad \text{d.h.} \quad [\mathfrak{C}, \rho]$$
$$q_\vartheta(y) = r_\vartheta(y) \cdot l(y) \quad [\mathfrak{C}, \nu] \quad \text{d.h.} \quad [\mathfrak{C}, \rho].$$

Hieraus folgt, daß die MLS unabhängig von der speziellen Wahl des dominierenden Maßes ist.

Beispiel 7.8 Es sei $\mathcal{P} = \{\mathsf{P}_\vartheta : \vartheta \in \Theta\}$. Es soll entschieden werden, welche der Hypothesen

$$H_i : \vartheta \in \Theta_i, i = 1, ..., k$$

für Y zutrifft. Bezeichnen wir die Dichten mit p_ϑ, so haben wir die Likelihoodfunktion

$$L(p_\vartheta; y) = p_\vartheta(y)$$

Es existiere ein $\hat{\vartheta}$ mit $\sup_{\vartheta \in \Theta} L(p_\vartheta; y) = L(p_{\hat{\vartheta}}; y)$. Die Entscheidungsfunktion

$$d(y) = \begin{cases} d_1 & \text{für} \quad \hat{\vartheta}(y) \in \Theta_1 \\ . & . \\ . & . \\ . & . \\ d_k & \hat{\vartheta}(y) \in \Theta_k \end{cases}$$

liefert eine Lösung für das gestellte Problem. Diese Lösung heißt MLS für $\Theta_1, ..., \Theta_k$. Sie ist Maximum-Likelihood-Entscheidungsfunktion für das Mehrentscheidungsproblem.

7.2.2 Kleinste-Quadrate-Methode

Zur Bestimmung der ML-Lösungen benötigt man die funktionelle Gestalt der Dichten. Sind die Dichten unbekannt, so kann man häufig die Kleinste-Quadrate-Methode (KQ-Methode) anwenden.

Es sei $f : \Theta \to \mathbb{R}^n, \Theta \subseteq \mathbb{R}^k$ eine Funktion. Zu einem $Z \in \mathcal{M}_n^{\geq}$ ist ein $\hat{\vartheta}$ mit

$$\min_{\vartheta \in \Theta} \mid y - f(\vartheta) \mid_Z^2 = \mid y - f(\hat{\vartheta}(y)) \mid_Z^2$$

zu bestimmen $\left(\mid a \mid_Z^2 = a'Za = \sum_{i,j=1}^{n} a_i a_j z_{ij} \right)$. $\hat{\vartheta}(y)$ heißt Kleinste-Quadrate-Schätzer (KQS) für ϑ.

Die KQ-Methode wird häufig dann angewendet, wenn eine Aussage über das erste Moment von Y gemacht werden soll.

Beispiel 7.9 Es sei Y eine zufällige Variable mit

$$\mathsf{E}Y = X\beta, \quad \mathsf{Var}\, Y = \sigma^2\Sigma$$

Sei $Z = I, f(\vartheta) = X\beta$. Dann haben wir $\hat{\beta}$ zu bestimmen mit

$$\begin{aligned} \mid y - X\hat{\beta} \mid^2 &= \min_{\beta} \mid y - X\beta \mid^2 \quad \text{und folglich gilt} \\ \hat{\beta}(y) &= (X'X)^{-1}X'y. \end{aligned}$$

Bei $Z = \Sigma^{-1}, f(\vartheta) = X\beta$ gilt

$$\begin{aligned} \hat{\beta} &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \\ &= MLS \quad \text{für } \hat{\beta} \quad \text{von Beispiel 7.7.} \end{aligned}$$

Bei $Z = \Sigma^{-1}$ ist also die MLS von Beispiel 7.7 auch KQS. Genau wie bei der MLS hat man zu klären, ob die so bestimmte KQS eine statistische Entscheidungsfunktion ist. Dabei gilt das folgende Resultat.

Satz 7.2 *Sei $\Theta \subseteq \mathbb{R}^k$ kompakt und sei f stetig. Dann kann die KQS so gewählt werden, daß sie meßbar und somit eine statistische Entscheidungsfunktion ist.*

BEWEIS: Wenn man berücksichtigt, daß die KQS für Normalverteilungen eine MLS ist, so folgt die Behauptung aus der entsprechenden Aussage für MLS.

Die Berechnung der KQS erfolgt über die Lösung der Normalgleichungen (NGL) $\frac{\mathrm{d}}{\mathrm{d}\beta} \mid y - X\beta \mid_Z^2 = 0$. Nehmen wir an, $f(\vartheta)$ sei differenzierbar und die KQS $\hat{\vartheta}(y)$ sei innerer Punkt von Θ. $\Rightarrow \hat{\vartheta}(y)$ ist Lösung von

$$\frac{\mathrm{d}}{\mathrm{d}\vartheta} \mid y - f(\vartheta) \mid_Z^2 = 0 \qquad (*).$$

Sei $\frac{\partial}{\partial \vartheta_j} f_i(\vartheta) = g_{ij}(\vartheta)$. Dann gilt (*) gdw.

$$\sum_{i,t=1}^{n} z_{it} g_{ij}(\vartheta) f_t(\vartheta) = \sum_{i,t=1}^{n} z_{it} g_{ij}(\vartheta) y_{it} \qquad j = 1, ..., k.$$

Falls $f(\vartheta)$ nichtlinear ist , so sind diese NGL nur iterativ lösbar. Falls $f(\vartheta)$ linear, so sind die NGL explizit lösbar.

□

7.2.3 Bayessches Prinzip

Aus den Beobachtungen und der a-priori Verteilung soll eine Entscheidungsfunktion gefunden werden.

$$\mathcal{P} = \{\mathsf{P}_\vartheta : \vartheta \in \Theta\}, \qquad \tilde{\mathsf{Q}} \quad a-priori Verteilung$$

p_ϑ bzw. $\tilde{q}$ seien Dichten von $\mathsf{P}_\vartheta, \tilde{\mathsf{Q}}$ bzgl. μ bzw. $\tilde{\mu}$.
Dann wählen wir die Bezeichnung $p_\vartheta(y) = p(y \mid \vartheta)$ und betrachten P_ϑ als bedingte Verteilung von Y bei gegebenem ϑ. Die gemeinsame Dichte von Y und ϑ ist darstellbar als

$$p(y, \vartheta) = \tilde{q}(\vartheta) p(y \mid \vartheta)$$

Satz 7.3 (Satz von Bayes:)

$$p(\vartheta \mid y) = \tilde{q}(\vartheta) p(y \mid \vartheta) \cdot \frac{1}{p(y)} \qquad mit$$

$$p(y) = \int_\Theta \tilde{q}(\vartheta) p(y \mid \vartheta) \tilde{\mu}(d\vartheta) \qquad für\ \tilde{q} = \frac{\mathrm{d}\tilde{\mathsf{Q}}}{\mathrm{d}\tilde{\mu}}$$

Interpretation: $p(\vartheta \mid y)$ ist die Dichte derjenigen Verteilung, die unsere Kenntnis über die verschiedenen Parmater ϑ nach der Beobachtung y ausdrückt. $p(\vartheta \mid y)$ heißt a-posteriori Dichte von ϑ. $\tilde{\mathsf{Q}}$ wird also korrigiert zu einer a-posteriori Verteilung.

Definition 7.9 *$\tilde{\vartheta}(y)$ heißt Bayessche ML–Schätzung für ϑ, falls*

$$p(\tilde{\vartheta} \mid y) = \max_{\vartheta} \quad p(\vartheta \mid y)$$

gilt.

Beispiel 7.10 Es seien $Y_1, ..., Y_n$ unabhängig, identisch binomialverteilt, d.h.

$$\begin{aligned} Y_i &= \begin{cases} 0 & \text{mit Wk} \quad 1-\vartheta \\ 1 & \text{mit Wk} \quad \vartheta. \end{cases} \qquad 0 \le \vartheta \le 1 \\ \text{Sei} \quad Z &= \sum_{i=1}^{n} Y_i. \\ \Rightarrow \ \mathsf{P}(Z = r \mid \vartheta) &= \binom{n}{r} \vartheta^r (1-\vartheta)^{n-r} \end{aligned}$$

Wenn $\tilde{\mathsf{Q}}$ die Gleichverteilung über $[0,1]$ ist, so gilt $\tilde{q} = 1$ für $0 \le \vartheta \le 1$. Die a-posteriori Dichte ist proportional zu $\tilde{q}(\vartheta) p(r \mid \vartheta)$. Folglich gilt

$$p(\vartheta \mid r) = \frac{\vartheta^r (1-\vartheta)^{n-r}}{\mathsf{B}(r+1, n-r+1)} \quad \text{für } \mathsf{B}(r+1, n-r+1) = \int_0^1 \vartheta^r (1-\vartheta)^{n-r} \, \mathrm{d}\vartheta \quad \text{und}$$

$p(\vartheta \mid r)$ ist maximal für $\hat{\vartheta} = \frac{r}{n}$. Diese Bayessche Schätzung ist auch MLS.

Bisher haben wir in diesem Kapitel Schätzungen und Tests kennengelernt und Eigenschaften wie Erwartungstreue bzw. Unverfälschtheit diskutiert. Jetzt soll das asymptotische Verhalten von Schätzungen und die Optimalität von Entscheidungsfunktionen untersucht werden.

7.3 Konsistenz von Schätzern

Das ist eine Aussage über das Verhalten der Schätzfunktionen bei großen Stichprobenumfängen. Sei

$$Y \sim \mathsf{P}_\vartheta, \quad \vartheta \in \Theta$$

Zu schätzen ist $\gamma(\vartheta)$. Y habe Werte im $\mathbb{R}^n$.
Es sei zum Beispiel $\mathsf{E}_\vartheta Y = \mu \mathbf{1}$ für alle ϑ. Dann gilt nach dem starken Gesetz der großen Zahlen

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow{f.s.} \mu,$$

d.h.

$$\mathsf{P}\left(\omega:\ \sup_{m\le n}\left|\frac{1}{n}\sum_{i=1}^{n}Y_i(\omega)-\mu\right|>\varepsilon\right)\ \underset{m\to\infty}{\longrightarrow}\ 0\ .$$

Das heißt, $\bar{Y}_n$ konvergiert fast sicher gegen μ.

Definition 7.10 *1. $\hat{\gamma}_n$ heißt schwach konsistent für γ, falls*

$$\mathsf{P}\left(|\hat{\gamma}_n-\gamma|>\varepsilon\right)\ \underset{n\to\infty}{\longrightarrow}\ 0.$$

2. $\hat{\gamma}_n$ heißt stark konsistent für γ, falls $\hat{\gamma}_n \overset{f.s.}{\longrightarrow} \gamma$, d.h.

$$\mathsf{P}\left(\sup_{n\ge N}|\hat{\gamma}_n-\gamma|>\varepsilon\right)\ \underset{N\to\infty}{\longrightarrow}\ 0.$$

3. $\hat{\gamma}_n$ heißt konsistent im quadratischen Mittel, falls

$$\mathsf{MSE}_\vartheta\hat{\gamma}_n\ \underset{n\to\infty}{\longrightarrow}\ 0.$$

Schwache Konsistenz ist eine mindeste Eigenschaft der Schätzungen. Die Maximum-Likelihood-Schätzungen haben diese Eigenschaften.

Vor.: $Y_1, Y_2, \ldots$, unabhängig, Dichte $f(z,\vartheta)$

Likelihoodgleichung: $L_n(\vartheta)=\ln\underbrace{\prod_{i=1}^{n}f(y_i,\vartheta)}_{f_{n,\vartheta}(y)}$.

Satz 7.4 *Es existiere $\mathsf{E}_{\vartheta_0}\ln\frac{f(Y_1,\vartheta)}{f(Y_1,\vartheta_0)}$. Dann gilt*

$$\lim_{n\to\infty}\mathsf{P}_{\vartheta_0}\left(y:\ f_{n,\vartheta_0}(y)>f_{n,\vartheta}(y)\right)=1.$$

Beweis: Wir haben

$$f_{n,\vartheta_0}(y)>f_{n,\vartheta}(y)\Longleftrightarrow\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(y_i,\vartheta)}{f(y_i,\vartheta_0)}<0.$$

$$\Longrightarrow\qquad \mathsf{P}_{\vartheta_0}\left(y:\ f_{n,\vartheta_0}(y)>f_{n,\vartheta}(y)\right)\ =\ \mathsf{P}_{\vartheta_0}\left(\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(y_i,\vartheta)}{f(y_i,\vartheta_0)}<0\right).$$

Die Zufallsgrößen $Z_i=\ln\frac{f(Y_i,\vartheta)}{f(Y_i,\vartheta_0)}$ sind i.i.d. und $\mathsf{E}_{\vartheta_0}Z_i$ existiert. Daher folgt für

$$\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(Y_i,\vartheta)}{f(Y_i,\vartheta_0)}=\frac{1}{n}\sum_{i=1}^{n}Z_i=\bar{Z}_n,$$

daß $$\bar{Z}_n\ \overset{f.s}{\longrightarrow}\ \underbrace{\mathsf{E}_\vartheta Z_1}_{=\mu}$$

und $$\mathsf{P}\left(\sup_{n\geq N}|\bar{Z}_n-\mu|>\varepsilon\right)\underset{N\to\infty}{\longrightarrow}0\ .$$

Nach der Jensen-Ungleichung haben wir

$$\mathsf{E}_{\vartheta_0}\ln\frac{f(Y_i,\vartheta)}{f(Y_i,\vartheta_0)}<\ln\mathsf{E}_{\vartheta_0}\frac{f(Y_i,\vartheta)}{f(Y_i,\vartheta_0)},\ \text{denn ln ist konkav.}$$

$$\Longrightarrow\quad \mu=\mathsf{E}_{\vartheta_0}Z_1<0.$$

$$\Longrightarrow\quad \mathsf{P}\left(\sup_{n\geq N}|\bar{Z}_n-\mu|>\left|\frac{\mu}{2}\right|\right)\longrightarrow 0.$$

$$\Longrightarrow\quad \mathsf{P}\left(\bar{Z}_n<\frac{\mu}{2}\right)\longrightarrow 1.$$

$$\Longrightarrow\quad \mathsf{P}_{\vartheta_0}(y:\ f_{n,\vartheta_0}(y)>f_{n,\vartheta}(y))\longrightarrow 1.$$

□

Satz 7.5 *Sei $\Theta\subset\mathbb{R}^1,\vartheta_0$ innerer Punkt. Sei $\ln\frac{f(z,\vartheta)}{f(z,\vartheta_0)}$ integrierbar und es existiere $\frac{\partial}{\partial\vartheta}f(y,\vartheta)$ und sei stetig in ϑ. Sei $\mathcal{A}_n(y):=\{\vartheta:\ \frac{\partial}{\partial\vartheta}\ln f_{n,\vartheta}(y)=0\}\neq\emptyset$. Dann existiert ein Schätzer $\hat{\vartheta}_n(y)\in\mathcal{A}_n(y)$ mit $\hat{\vartheta}_n\to\vartheta$ P_{ϑ_0}–f.s..*

Beweis: Sei ϑ_0 innerer Punkt von Θ, $(\vartheta_0-a,\vartheta_0+a)\subset\Theta$. Sei weiterhin

$$C_n=\{y:\ f_{n,\vartheta_0}(y)>f_{n,\vartheta_0-a}(y)\text{ und }f_{n,\vartheta_0}(y)>f_{n,\vartheta_0+a}(y)\}.$$

Nach Satz 7.4 gilt $\mathsf{P}_{\vartheta_0}(C_n)\to 1$. Für $y\in C_n$ existiert $\hat{\vartheta}_n(y)\in(\vartheta_0-a,\vartheta_0+a)$ mit $\frac{\partial}{\partial\vartheta}f(y,\vartheta)\big|_{\hat{\vartheta}_n(y)}=0$. $\hat{\vartheta}_n$ kann meßbar gewählt werden und wir haben folglich

$$\mathsf{P}_{\vartheta_0}\left(\left|\hat{\vartheta}_n-\vartheta_0\right|<a\right)\longrightarrow 1$$

für beliebiges $a>0$. Damit ist $\hat{\vartheta}_n$ schwach konsistent.

□

Nun folgt ein Beispiel für Konsistenz im quadratischen Mittel.
Betrachten wir das lineare Modell $Y=X_n\beta+\varepsilon$ mit $Y\in\mathbb{R}^n$, $\mathsf{Var}\,\varepsilon=\sigma^2I_n$, X_n habe vollen Rang k. Dann ist

$$\hat{\beta}_n(y)=(X_n'X_n)^{-1}X_n'y$$

Kleinste–Quadrate–Schätzung für β.

Satz 7.6 *Es gelte $\frac{1}{n}X_n'X_n\underset{n\to\infty}{\longrightarrow}A\in\mathfrak{M}_k^>$. Dann konvergiert $\hat{\beta}_n$ im quadratischen Mittel gegen β, d.h. $\mathsf{MSE}_\vartheta(\hat{\beta}_n)\to 0$.*

Beweis: Zunächst ist $\hat{\beta}_n$ erwartungstreu.

$$\begin{aligned}\Longrightarrow\quad \mathsf{MSE}_\vartheta\hat{\beta}_n &= \mathsf{E}_\vartheta\left(\hat{\beta}_n(Y)-\beta\right)\left(\hat{\beta}_n(Y)-\beta\right)'\\ &= \mathsf{E}_\vartheta\hat{\beta}_n(\varepsilon)\hat{\beta}_n(\varepsilon)'=\sigma^2(X_n'X_n)^{-1}\\ &= \frac{\sigma^2}{n}\cdot\left(\frac{1}{n}X_n'X_n\right)^{-1}\underset{n\to\infty}{\longrightarrow}0.\end{aligned}$$

$\Longrightarrow$ $\hat{\beta}_n$ konvergiert im quadratischen Mittel gegen β.

□

7.4 Optimalität von Entscheidungsfunktionen

Im allgemeinen existieren bei einem Problem mehrere Enscheidungsfunktionen, die als Lsung in Frage kommen. Notwendig sind die Bewertung der Entscheidungsfunktionen und die Auswahl einer optimalen Entscheidungsfunktion.
Δ: Menge aller Entscheidungen
$\mathcal{E}$: Menge von Entscheidungsfunktionen
$Y \sim \mathsf{Q} \quad \in \quad \{\mathsf{P}_\vartheta : \vartheta \in \Theta\} = \mathcal{P}$.

Definition 7.11 (Verlust) *Eine auf $\Theta \times \Delta$ definierte nichtnegative reelle Funktion L heißt Verlustfunktion.*

$L(\vartheta, \delta)$ drückt den Verlust aus, den man beim Treffen der Entscheidung δ und bei Vorliegen der Verteilung P_ϑ erleidet.

Die Wahl von L hängt vom speziellen Problem ab. Im allgemeinen werden wir $L(\vartheta, \delta) = 0$ annehmen, falls δ eine für ϑ richtige Entscheidung ist. $L(\vartheta, \delta)$ soll die Bewertung der Abweichung der richtigen von der tatsächlich getroffenen Entscheidung beschreiben.

Wir betrachten Spezialfälle:
Schätzen von Parametern $\gamma(\vartheta)$:
$L(\vartheta, \delta) = (\delta - \gamma(\vartheta))^2$ bzw.
$L = (\delta - \gamma(\vartheta))(\delta - \gamma(\vartheta))' \Rightarrow$ quadratisches Anwachsen des Verlustes mit der Grße des Fehlers.

Ein lineares Anwachsen liegt bei dem Ansatz $L(\vartheta, \delta) = \mid \delta - \gamma(\vartheta) \mid$ vor.

Testen von Hypothesen: (Neyman-Pearson Verlustfunktion)

$$L(\vartheta, d_H) = \begin{cases} 0 & \text{für} \quad \vartheta \in \Theta_H \\ L_0 & \text{für} \quad \vartheta \in \Theta_K \end{cases} \qquad L(\vartheta, d_K) = \begin{cases} L_1 & \text{für} \quad \vartheta \in \Theta_H \\ 0 & \text{für} \quad \vartheta \in \Theta_K \end{cases}$$

$\Rightarrow$ keine unterschiedliche Bewertung der $\vartheta \in \Theta_H$ bzw. $\in \Theta_K$.

Definition 7.12 (Risiko) *Unter einer Risikofunktion R verstehen wir eine auf $\Theta \times \mathcal{E}$ definierte reelle Funktion, deren Werte den bei Benutzen der Entscheidungsfunktion d und Vorliegen der Verteilung P_ϑ erwarteten Verlust angeben.*

Für nichtrandomisierte Entscheidungsfunktion d gilt

$$R(\vartheta, d) = \mathsf{E}_\vartheta L(\vartheta, d(Y)) = \int L(\vartheta, d(y)) \mathsf{P}_\vartheta(\,\mathrm{d}y).$$

Speziell bei der Schätzung mit quadratischem Verlust ($l = 1$) erhalten wir

$$R(\vartheta, d) = \mathsf{E}_\vartheta(d(Y) - \gamma(\vartheta))^2 = \mathsf{Var}_\vartheta d(Y) + b_\vartheta(d)^2$$

$\Rightarrow R$ ist durch die Streuung und die Verzerrung bestimmt.

Betrachten wir das Hypothesentesten

$$\text{H: } \vartheta \in \Theta_H \quad \text{gegen K: } \vartheta \in \Theta_K$$

Beispiel 7.11 Seien $Y_1, .., Y_n$ unabhängige Zufallsgrßen und $Y_i \sim B(1, \vartheta)$. Gesucht ist ein unverfälschter α–Test für

$$\text{H: } \vartheta \leq \vartheta_0 \quad \text{gegen K: } \vartheta > \vartheta_0.$$

Wir knnen später rechtfertigen, daß ein bester Test nur über $T(y) = \sum_{i=1}^{n} y_i$ von y abhängt. Sei

$$\varphi_c(y) = \begin{cases} 1 & \text{für } T(y) > c \\ 0 & \text{sonst} \end{cases} \qquad \text{für } \; c = 0, 1, 2, ...$$

Die Gütefunktion ist

$$\beta(\vartheta) = \mathsf{E}_\vartheta \varphi_c(Y) = \mathsf{P}_\vartheta(T(y) > c) = \sum_{t=c+1}^{n} \binom{n}{t} \vartheta^t (1-\vartheta)^{n-t}$$

Mit der unvollständigen β–Funktion erhält man

$$\beta(\vartheta) = \binom{n}{c} (n-c) \int_0^\vartheta t^c (1-t)^{n-c-1} \, \mathrm{d}t$$

und damit $\beta'(\vartheta) > 0$ und $0 < \vartheta < 1$. Damit ist β streng monoton wachsend. Ein α–Test muß

$$\beta(\vartheta) \leq \alpha \text{ für } \vartheta \leq \vartheta_0$$

erfüllen. Das Problem besteht darin, daß unendlich viele ϑ_0, aber nur endlich viele φ_c mglich sind. Nur für endlich viele ϑ_0 kann ein c bestimmt werden, so daß φ_c ein unverfälschter α–Test ist.

Damit man aber für jedes ϑ_0 eine Lsung findet, führt man Tests in der Form

$$\tilde{\varphi}_{c,\gamma}(y) = \begin{cases} 1 & \text{für } \; T(y) > c \\ \gamma & \text{für } \; T(y) = c \quad , 0 \leq \gamma \leq 1 \\ 0 & \text{für } \; T(y) < c \quad , c = 0, 1, 2, ... \end{cases}$$

ein. Es gilt

$$\tilde{\beta}(\vartheta) = \mathsf{E}_\vartheta \tilde{\varphi}_{c,\gamma}(Y) = \sum_{t=c+1}^{n} \binom{n}{t} \vartheta^t (1-\vartheta)^{n-t} + \gamma \binom{n}{c} \vartheta^c (1-\vartheta)^{n-c}$$

und auch $\tilde{\beta}'(\vartheta) > 0$. Man bestimmt c, γ derart, daß $\tilde{\beta}(\vartheta_0) = \alpha$. Die Funktion

$$F(t) = \mathsf{P}_{\vartheta_0}(T(y) \leq t)$$

ist stets rechtsseitig stetig. Daher existiert $c = \min\{t : F(t) \geq \alpha\}$. c ist also der kleinste Wert mit $F(t) \geq \alpha$, d.h.

$$\begin{aligned} \mathsf{P}_{\vartheta_0}(T(y) > c) &\leq \alpha \\ \mathsf{P}_{\vartheta_0}(T(y) \geq c) &> \alpha \end{aligned}$$

Es soll $\tilde{\beta}(\vartheta_0) = \alpha$ sein, d.h.

$$\mathsf{P}_{\vartheta_0}(T(y) > c) + \gamma \mathsf{P}_{\vartheta_0}(T(y) = c) = \alpha$$

Falls $\mathsf{P}_{\vartheta_0}(T(y) > c) \neq \alpha$, d.h. $< \alpha$, aber $\mathsf{P}_{\vartheta_0}(T(y) > c) + \mathsf{P}_{\vartheta_0}(T(y) = c) > \alpha$, so wählt man

$$\gamma = \frac{\alpha - \mathsf{P}_{\vartheta_0}(T(y) > c)}{\mathsf{P}_{\vartheta_0}(T(y) = c)}.$$

Andernfalls ist $\gamma = 0$. Also ist $\tilde{\varphi}_{c,\gamma}$ ein unverfälschter α–Test.

Wir brauchen folglich Tests, die sich mit Wahrscheinlichkeit γ für d_k entscheiden (randomisierte Tests).

Beispiel 7.12 (Numerisches Beispiel) Sei $n = 7,\ \vartheta_0 = 0.35,\ \alpha = 0.05$.
Dann gilt

$$\begin{aligned} \mathsf{P}_{\vartheta_0}(T(y) > 5) &= 0.0090 \\ \mathsf{P}_{\vartheta_0}(T(y) = 5) &= 0.0466 \end{aligned}$$

Folglich nutzt der Test φ_5 bei ϑ_0 nicht die gesamte Irrtumswahrscheinlichkeit 1.Art $\alpha = 0.05$ aus, es ist $\beta(\vartheta_0) = 0.0090$. Für φ_4 gilt bereits $\mathsf{P}_{\vartheta_0}(T(y) > 4) = 0.0556 > \alpha$. Der randomisierte Test

$$\tilde{\varphi}_{5,\gamma}(y) = \begin{cases} 1 & \text{für } T(y) > 5 \\ \gamma & \text{für } T(y) = 5 \\ 0 & \text{für } T(y) < 5 \end{cases}$$

mit $\gamma = \frac{0.05 - 0.009}{0.0466} = 0.88$ ist ein unverfälschter α–Test.

Definition 7.13 *Unter einer randomisierten Entscheidungsfunktion d versteht man eine reellwertige Funktion, die auf dem Produktraum $\mathfrak{D} \times \mathfrak{Y}$ definiert ist und bei festem y eine Wahrscheinlichkeitsverteilung W_y über $(\Delta, \mathfrak{D})$ ist und bei festem $D \in \mathfrak{D}$ eine $\mathfrak{C}$–meßbare Funktion von y ist.*
Speziell: $\Delta = \{d_H, d_K\}, \mathfrak{D} = \mathfrak{P}(\Delta)$. Eine Wahrscheinlichkeitsverteilung W_y ist durch die Angabe einer Wahrscheinlichkeit $\varphi(y)$ bestimmt, mit der d_k getroffen wird.

Die Risikofunktion für Tests sieht dann folgendermaßen aus:

Wir benutzen die Neyman-Pearsonsche Verlustfunktion $L(\vartheta, \delta)$.

a) nichtrandomisierte Tests

$$\varphi(y) = \begin{cases} 1 & \text{für } y \in S \\ 0 & \text{sonst} \end{cases}$$

Folglich gilt

$$\begin{aligned} R(\vartheta, \varphi) &= \int L(\vartheta, d_K)\varphi(y)\mathsf{P}_\vartheta(\,\mathrm{d}y) + \int L(\vartheta, d_H)[1-\varphi(y)]\mathsf{P}_\vartheta(\,\mathrm{d}y) \\ &= L(\vartheta, d_K)\mathsf{P}_\vartheta(S) + L(\vartheta, d_H)\mathsf{P}_\vartheta(\overline{S}) \\ &= \begin{cases} L_1\mathsf{P}_\vartheta(S) & \text{für } \vartheta \in \Theta_H \\ L_0\mathsf{P}_\vartheta(\overline{S}) & \text{für } \vartheta \in \Theta_K\ . \end{cases} \end{aligned}$$

Es war $\mathsf{P}_\vartheta(S) = \mathsf{E}_\vartheta\varphi(Y)$.

b) randomisierte Tests

Auch hier gilt

$$\begin{aligned} R(\vartheta, \varphi) &= \int \left[L(\vartheta, d_K)\varphi(y) + L(\vartheta, d_H)(1-\varphi(y))\right] \mathsf{P}_\vartheta(\,\mathrm{d}y) \\ &= \begin{cases} L_1\mathsf{E}_\vartheta\varphi(Y) & \vartheta \in \Theta_H \\ L_0\mathsf{E}_\vartheta(1-\varphi(Y)), & \vartheta \in \Theta_K \end{cases} \end{aligned}$$

$\Rightarrow R(\vartheta, \varphi)$ wird durch die Gütefunktion $\mathsf{E}_\vartheta\varphi(Y)$ bestimmt.

Sei d eine randomisierte Entscheidungsfunktion. $d(y)$ habe die Verteilung W_y. Dann gilt

$$R(\vartheta, d) = \int\int L(\vartheta, \delta)\mathsf{W}_y(\,\mathrm{d}\delta)\mathsf{P}_\vartheta(\,\mathrm{d}y)$$

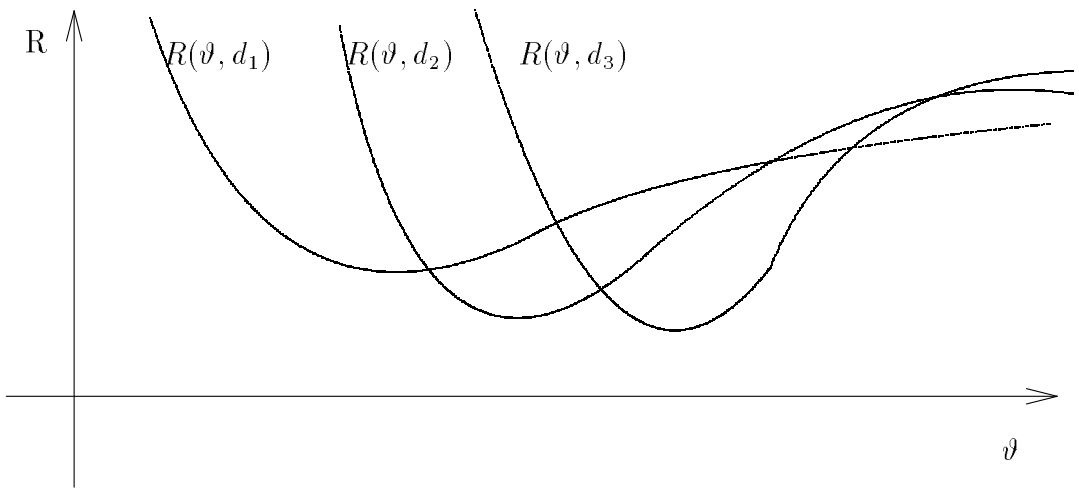
Wenn man annimmt, daß die Maße $\mathsf{P}_\vartheta, \mathsf{W}_y$ Lebesgue–Dichten haben, d.h.

$$\begin{aligned} \mathsf{P}_\vartheta(C) &= \int\limits_C p_\vartheta(y)\,\mathrm{d}y \\ \mathsf{W}_y(D) &= \int\limits_D w_y(\delta)\,\mathrm{d}\delta \end{aligned}$$

so ergibt sich

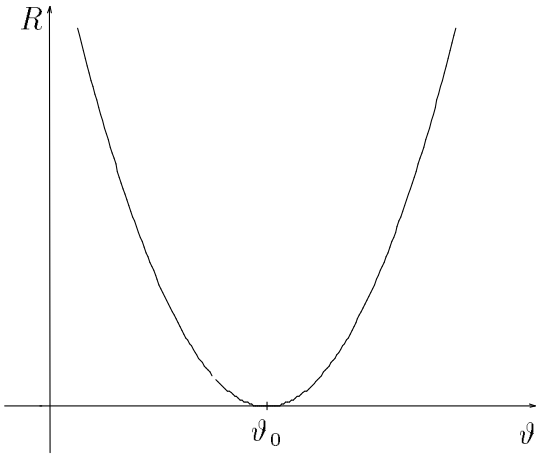
$$R(\vartheta, d) = \int\int L(\vartheta, \delta)w_y(\delta)p_\vartheta(y)\,\mathrm{d}\delta\,\mathrm{d}y$$

Jeder Entscheidungsfunktion d wird durch das Risiko eine Funktion von ϑ zugeordnet. Diese Risikofunktion ist Grundlage für die Bestimmung optimaler Entscheidungsfunktionen. Vergleich von Entscheidungsfunktionen'en:
d_1 heißt besser als d_2, falls $R(\vartheta, d_1) \leq R(\vartheta, d_2) \quad \forall \quad \vartheta \in \Theta$ und $R(\vartheta, d_1) < R(\vartheta, d_2)$ für ein ϑ. Im allgemeinen sind d_1, d_2 nicht vergleichbar in diesem Sinne. Meist hat man



Ist $\mathcal{E}$ sehr umfassend, so existieren "extreme" Entscheidungsfunktion.

Schätzproblem: Wähle $d(y) \equiv \vartheta_o \in \Theta \Rightarrow R(\vartheta, d) = (\vartheta_o - \vartheta)^2$.



$\Rightarrow d$ ist für $\vartheta = \vartheta_o$ eine beste Entscheidungsfunktion.

Testproblem: H: $\vartheta \in \Theta_H$ gegen K: $\vartheta \in \Theta_K$. Wähle $d(y) \equiv d_H \Rightarrow$ Test ist für $\vartheta \in \Theta_H$ ideal: Irrtumswahrscheinlichkeit 1. Art $= 0$, aber die Irrtumswahrscheinlichkeit 2. Art ist groß.

Diese Entscheidungsfunktionen sind unbrauchbar, denn sie sind unabhängig von den Beobachtungen $\Rightarrow$ man muß $\mathcal{E}$ einschränken, damit derartige Entscheidungsfunktionen ausgeschlossen werden.

Definition 7.14 *$d^* \in \tilde{\mathcal{E}}$ heißt beste Entscheidungsfunktion, falls $R(\vartheta, d^*) \leq R(0, d) \quad \forall \vartheta \in \Theta, d \in \tilde{\mathcal{E}}$ gilt.*

Beispiel 7.13 Es seien Y_1, Y_2 zufällige unkorrelierte Variable mit

$$\mathsf{E}Y_1 = \mathsf{E}Y_2 = \mu,\ \mathsf{Var}\ Y_1 = \mathsf{Var}\ Y_2 = \sigma^2,$$

d.h. $\mathcal{P} = \{\mathsf{P}_\vartheta : \vartheta \in \Theta\}$ mit $Y = \binom{Y_1}{Y_2} \sim \mathsf{P}_\vartheta$ und $\vartheta = (\mu, \sigma, \kappa) \in \mathbb{R}^1 \times \mathbb{R}^+ \times K$. Es gilt $\mathsf{E}_\vartheta Y = \mu\mathbf{1}, \mathsf{Var}_\vartheta Y = \sigma^2 I$. Es ist $\mathcal{E}$ die Menge aller für μ erwartungstreuen, in Y_1, Y_2 linearen Schätzer, d.h.

$$\mathcal{E} = \{d : d(y) = a_1 y_1 + a_2 y_2 + b :\ \mathsf{E}_\vartheta(a_1 Y_1 + a_2 Y_2 + b) = \mu\ \forall\vartheta\}\,.$$

Gesucht ist in $\mathcal{E}$ eine beste Entscheidungsfunktion. Es ergibt sich $\mathcal{E} = \{d : d(y) = a_1 y_1 + (1 - a_1) y_2\}$ und für $d(Y) = a_1 Y_2 + (1 - a_1) Y_2$ hat man

$$\mathsf{E}_\vartheta \left(\frac{d(Y) - \mu}{\sigma^2} \right)^2 = a_1^2 + (1 - a_1)^2$$

zu minimieren. Die Lsung ist $a_1 = \frac{1}{2}$. Damit ist d^* mit $d^*(Y) = \frac{1}{2}(Y_1 + Y_2)$ beste erwartungstreue lineare SF für μ.

Wenn ein derartiges d^* nicht existiert, so sucht man wenigstens eine solche Entscheidungsfunktion in $\tilde{\mathcal{E}}$, die nicht durch eine andere für alle ϑ dominiert wird.

Definition 7.15 *$\tilde{d} \in \tilde{\mathcal{E}}$ heißt zulässig in $\tilde{\mathcal{E}}$, falls kein $d \in \tilde{\mathcal{E}}$ derart existiert, daß $R(\vartheta, d) \leq R(\vartheta, \tilde{d})\ \ \forall\vartheta, (< \text{für ein } \vartheta)$ gilt. Alle anderen Entscheidungsfunktionen in $\mathcal{E}$ heißen nichtzulässig.*

Wir betrachten das obige Beispiel. Man stellt fest, daß $d^*(Y)$ die einzige zulässige Schätzfunktion in $\mathcal{E}$ ist.

Beispiel 7.14 Sei $Y \sim \mathsf{N}_2(\mu\mathbf{1}, \sigma^2 I)$. Dann ist die MLS für σ^2 gegeben durch $\hat{\sigma}^2(y) = \frac{(y_1 - y_2)^2}{4}$ und damit $E\hat{\sigma}^2(y) = \frac{1}{2}\sigma^2$. Eine erwartungstreue Schätzfunktion für σ^2 ist gegeben durch $\tilde{\sigma}^2 = 2\hat{\sigma}^2$. Nimmt man als Verlust $L(\sigma^2, d) = (\sigma^2 - d)^2$, so erhält man

$$R(\sigma^2, \hat{\sigma}^2) = \frac{3}{4}\sigma^4,$$
$$R(\sigma^2, \tilde{\sigma}^2) = 2\sigma^4,$$

so daß $\tilde{\sigma}^2$ in der Klasse aller quadratischen Schätzfunktionen für σ^2 nicht zulässig ist.

Bei der Suche von optimalen Entscheidungsfunktionen betrachtet man nur zulässige Entscheidungsfunktionen. Es existieren mehrere Mglichkeiten für die Bestimmung optimaler Entscheidungsfunktionen. Die meist benutzten Kriterien beruhen auf dem Minimax- und dem Bayes-Risiko. Interpretiert man die gegebenen Verteilungen als die Strategien der Natur und die Entscheidungsfunktion als die des Statistikers, so kann man das Ziel verfolgen, den grßten erwarteten Verlust, den die Naturstrategien unter $d \in \tilde{\mathcal{E}}$ verursachen knnen, zu minimieren.

Definition 7.16 *$d^* \in \tilde{\mathcal{E}}$ heißt Minimaxlsung in $\tilde{\mathcal{E}}$, falls*

$$\sup_{\vartheta\in\Theta} R(\vartheta, d^*) = \min_{d\in\tilde{\mathcal{E}}} \sup_{\vartheta\in\Theta} R(\vartheta, d)$$

gilt.

Das wird sowohl bei Schätzungen als auch bei Tests benutzt.

Minimax-Entscheidungsfunktionen benutzt man dann, wenn keine Vorkenntnisse über die Verteilung, d.h. über ϑ vorliegen. Vielfach glaubt man, Vorkenntnisse zu haben. Hat man Hinweise über die mgliche Lage von ϑ, so kann man diese durch eine a-priori Verteilung $\tilde{\mathsf{Q}}$ beschreiben. Das Funktional

$$\tilde{R}_{\tilde{Q}}(d) = \int R(\vartheta, d)\tilde{\mathsf{Q}}(\,\mathrm{d}\vartheta)$$

ist ein gemitteltes Risiko. Diejenigen ϑ werden stärker bewichtet, an deren Richtigkeit wir besonders glauben. Ein d^*, das $\tilde{R}$ minimiert, heißt Bayessche Lsung (bzgl. $\tilde{\mathsf{Q}}$).

Es gibt einen Zusammenhang zwischen Minimax- und Bayesschen Lsungen. Ist $\tilde{\mathsf{Q}}$ nicht gut bekannt, so wird man mit der ungünstigsten Strategie der Natur rechnen müssen. Dabei ist Q^* ungünstigste a-priori Verteilung, falls

$$\sup_{\tilde{Q}} \inf_{d\in\tilde{\mathcal{E}}} \tilde{R}_{\tilde{Q}}(d) = \inf_{d\in\tilde{\mathcal{E}}} \tilde{R}_{Q^*}(d)$$

Satz 7.7 *Unter der Bedingung der Vertauschbarkeit*

$$\inf_d \sup_{\hat{\mathsf{Q}}} \hat{R}_{\hat{\mathsf{Q}}}(d) = \sup_{\tilde{\mathsf{Q}}} \inf_d \tilde{R}_{\tilde{\mathsf{Q}}}(d)$$

ist eine Minimax-Lsung eine Bayessche Lsung bzgl. Q^*. Das heißt,es muß also für eine Minimax Entscheidungsfunktion d^* gelten:

$$\tilde{R}_{\mathsf{Q}^*}(d^*) \le \tilde{R}_{\mathsf{Q}^*}(d) \quad \forall d :$$

BEWEIS: d^* sei Minimax–Entscheidungsfunktion. Dann gilt:

$$\begin{aligned}
\sup_\vartheta R(\vartheta, d^*) &= \inf_d \sup_{\tilde{\mathsf{Q}}} \tilde{R}_{\tilde{\mathsf{Q}}}(d), \quad \text{denn} \\
\sup_\vartheta R(\vartheta, d^*) &= \inf_d \underbrace{\sup_\vartheta R(\vartheta, d)}_{=\sup_{\tilde{\mathsf{Q}}} \tilde{R}_{\tilde{\mathsf{Q}}}(d)} = \inf_d \sup_{\tilde{\mathsf{Q}}} \tilde{R}_{\tilde{\mathsf{Q}}}(d). \\
\Rightarrow \tilde{R}_{\mathsf{Q}^*}(d^*) &\le \sup_\vartheta R(\vartheta, d^*) = \inf_d \sup_{\tilde{\mathsf{Q}}} \tilde{R}_{\tilde{\mathsf{Q}}}(d) \\
&= \sup_{\tilde{\mathsf{Q}}} \inf_d \tilde{R}_{\tilde{\mathsf{Q}}}(d) = \inf_d \tilde{R}_{\mathsf{Q}^*}(d).
\end{aligned}$$

□

Es gelten verschiedene Aussagen über den Zusammenhang zwischen Minimax– und Bayesschen Lsungen.

Satz 7.8 *Sei* Q_0 *eine a–priori Verteilung,* d_0 *eine Bayessche Entscheidungsfunktion bzgl.* Q_0 *und sei* $R(\vartheta, d_0) = const.$ *Dann ist* d_0 *eine Minimax–Entscheidungsfunktion.*

BEWEIS: Angenommen, d_0 sei keine Minimax–Entscheidungsfunktion. Dann existiert ein d' mit

$$\sup_\vartheta R(\vartheta, d') < \sup_\vartheta R(\vartheta, d_0) = R(\vartheta, d_0) \quad \forall \vartheta. \Rightarrow \tilde{R}_{\mathsf{Q}_0}(d') < \tilde{R}_{\mathsf{Q}_0}(d_0),$$

was ein Widerspruch ist.

□

Beispiel 7.15 Sei $\mathsf{P} = \{\mathsf{P}_1, \mathsf{P}_2\}$, $\Theta = \{1, 2\}$. P_i habe die Dichte p_i. Sei $\Delta = \{d_1, d_2\}$. Eine a–priori Verteilung über Θ ist durch ζ beschrieben: ζ ist die Wahrscheinlichkeit für $\vartheta = 1$, $1 - \zeta$ ist die Wahrscheinlichkeit für $\vartheta = 2$. Die Verlustfunktion sei

$$L(1, d_2) = L_1, \; L(1, d_1) = 0, \; L(2, d_1) = L_0, \; L(2, d_2) = 0.$$

d sei eine Entscheidungsfunktion. Es ergibt sich

$$\begin{aligned}
R(1, d) &= \int_{\overline{S}} L(1, d_1) p_1(y)\, \mathrm{d}y + \int_S L(1, d_2) p_1(y)\, \mathrm{d}y \\
&= L_1 \int_S p_1(y)\, \mathrm{d}y \\
R(2, d) &= \int_{\overline{S}} L(2, d_1) p_2(y)\, \mathrm{d}y + \int_S L(2, d_2) p_2(y)\, \mathrm{d}y \\
&= L_0 \int_{\overline{S}} p_2(y)\, \mathrm{d}y\,.
\end{aligned}$$

Für das Bayessche Risiko erhalten wir

$$\begin{aligned}
\tilde{R}_\zeta &= \zeta R(1, d) + (1 - \zeta) R(2, d) \\
&= \zeta L_1 \int_S p_1(y)\, \mathrm{d}y + (1 - \zeta) L_0 (1 - \int_S p_2(y)\, \mathrm{d}y) \\
&= \zeta L_1 \int_S p_1(y)\, \mathrm{d}y + (1 - \zeta) L_0 - (1 - \zeta) L_0 \int_S p_2(y)\, \mathrm{d}y \\
&= (1 - \zeta) L_0 + \int_S [\zeta L_1 p_1(y) - (1 - \zeta) L_0 p_2(y)]\, \mathrm{d}y\,.
\end{aligned}$$

Folglich wird $\min_d \tilde{R}_\zeta(d)$ erreicht für

$$S = \left\{ y : \frac{\zeta L_1}{(1 - \zeta) L_0} < \frac{p_2(y)}{p_1(y)} \right\}.$$

$$\Rightarrow \varphi_c(y) \;=\; \begin{cases} d_2 & > \\ \text{bel. für } \frac{p_2(y)}{p_2(y)} & = c = \frac{\zeta L_1}{(1-\zeta)L_2} \\ d_1 & < \end{cases}$$

ist eine Bayessche Lsung. Nach dem vorherigen Satz ist φ_c eine Minimax–Lsung, falls

$$R(1,\varphi_c) = R(2,\varphi_c)$$

gilt. Es sei

$$S_\zeta \;=\; \left\{ y : \frac{p_2(y)}{p_1(y)} > \frac{\zeta}{1-\zeta}\frac{L_1}{L_0} \right\}$$
$$G_\zeta \;=\; \left\{ y : \frac{p_2(y)}{p_1(y)} = \frac{\zeta}{1-\zeta}\frac{L_1}{L_0} \right\}$$

$\frac{\zeta}{1-\zeta}$ ist monoton wachsend in ζ, also ist $\mathsf{P}_1(S_\zeta)$ monoton fallend in ζ. Sei

$$a(\zeta) \;=\; \int\limits_{S_\zeta} p_1(y)\,\mathrm{d}y\,, \qquad a_=(\zeta) = \int\limits_{G_\zeta} p_1(y)\,\mathrm{d}y,$$
$$b(\zeta) \;=\; \int\limits_{\overline{S}\setminus G_\zeta} p_2(y)\,\mathrm{d}y\,, \quad b_=(\zeta) = \int\limits_{G_\zeta} p_2(y)\,\mathrm{d}y\,.$$

Für einen Bayesschen Test

$$\varphi_{c,\gamma}(y) \;=\; \begin{cases} d_2 & > \\ \gamma \quad \text{für } \frac{p_2(y)}{p_1(y)} & = c = \frac{\zeta L_1}{(1-\zeta)L_0} \\ d_1 & < \end{cases}$$

ist

$$R(1,\varphi_{c,\gamma}) = L_1 a(\zeta) + \gamma L_1 a_=(\zeta), R(2,\varphi_{c,\gamma}) = L_0 b(\zeta) + (1-\gamma)L_0 b_=(\zeta)$$

und man findet stets Werte c^*, γ^*, so daß

$$R(1,\varphi_{c^*,\gamma^*}) = R(2,\varphi_{c^*,\gamma^*})$$

gilt. Man zeigt auch, daß es zu beliebigen α einen Test $\varphi_{c,\gamma}$ mit

$$\mathsf{E}_1\varphi_{c,\gamma}(Y) = \alpha$$

gibt. Damit existiert zu jedem α eine a–priori Verteilung, so daß der entsprechende Bayessche Test die Wahrscheinlichkeit für den Fehler 1.Art gleich α hat.

Beste Tests für spezielle Testprobleme

1. Einfache Hypothese gegen einfache Alternative

$$P = \{\mathsf{P}_{\vartheta_0}, \mathsf{P}_{\vartheta_1}\}$$
$$H : \vartheta = \vartheta_0$$
$$K : \vartheta = \vartheta_1$$

$\Rightarrow$ es existiert ein σ–endliches Maß μ derart, daß $p_i(y)$ Dichten von P_{ϑ_i}. (z.B. $\mu = \mathsf{P}_{\vartheta_0} + \mathsf{P}_{\vartheta_1}$).

Für die heuristischen Betrachtungen nehmen wir diskrete Verteilungen P_{ϑ_0}, P_{ϑ_1}. Wir knnen uns auf diejenigen $y \in \mathcal{Y}$ beschränken, die unter P_{ϑ_0} oder P_{ϑ_1} positive Wahrscheinlichkeiten haben. Die Punkte aus $\mathcal{N} := \{y \in Y : p_0(y) = p_1(y) = 0\}$ tragen keine Wahrscheinlichkeiten. Wir suchen kritische Bereiche $\Psi \subset \mathcal{Y} \setminus \mathcal{N}$. Die Eigenschaft: φ ist α–Test, heißt für einen nichtrandomisierten Test φ

$$\int \varphi(y)\mathsf{P}_{\vartheta_0}(\,\mathrm{d}y) = \mathsf{P}_{\vartheta_0}(\Psi) = \sum_{y\in\Psi} p_0(y) \le \alpha. \qquad (*)$$

Unter allen kritischen Bereichen Ψ mit $(*)$ ist derjenige mit der maximalen Güte unter P_{ϑ_1} zu bestimmen:

$$\int \varphi(y)\mathsf{P}_{\vartheta_1}(\,\mathrm{d}y) = \mathsf{P}_{\vartheta_1}(\Psi) = \sum_{y\in\Psi} p_1(\mathcal{Y}) = \max_{\Psi} \qquad (**)$$

Wie sucht man ein derartiges Ψ ?
Gehört y zu Ψ, so wird in $(*)$ $p_0(y)$ und in $(**)$ $p_1(y)$ einen Betrag leisten. Am günstigsten sind diejenigen y, die unter $p_1(y)$ relativ zu p_0 einen großen Wert haben, d.h. diejenigen y mit $\frac{p_1(y)}{p_o(y)}$ groß.
D.h., man wählt eine mglichst kleine Zahl $k_0 \ge 0$ mit

$$\sum_{y\in\mathcal{Y}\setminus\mathcal{N}} p_0(y) 1_{[k_0,\infty)} \left(\frac{p_1(y)}{p_0(y)}\right) \le \alpha$$

und setzt $S = \left\{y :\ y \in \mathcal{Y} \setminus \mathcal{N},\ \frac{p_1(y)}{p_0(y)} \ge k_\alpha\right\}$
(Wir setzen $\frac{p_1(y)}{p_0(y)} = \infty$ für $p_0(y) = 0,\ p_1(y) \neq 0$).
Diese Heuristik führt zu folgenden Satz, wobei jetzt beliebige Maße P_{ϑ_0}, P_{ϑ_1} zugelassen werden.

Satz 7.9 (Neyman–Pearson–Test) *Gegeben seien eine Konstante $k \ge 0$ und der Test*

$$\varphi_k(y) = \begin{cases} 1 & \textit{falls } p_1(y) \ge k p_0(y) \\ 0 & \textit{sonst} \end{cases} .$$

Für $\alpha := \mathsf{E}_{\vartheta_0}\varphi_k(Y)$ ist φ_k ein bester α–Test für H gegen K.

Beweis: Sei ψ ein anderer α–Test, d.h. $\int \psi \mathsf{P}_{\vartheta_0}(\,\mathrm{d}y) \le \alpha$. Für die meßbaren Mengen

$$M^+ = \{y :\ \varphi_k(y) > \psi(y)\} \quad \text{und} \quad M^- = \{y :\ \varphi_k(y) < \psi(y)\}$$

gilt

$$y \in M^+ \Rightarrow \varphi_k(y) > 0 \Rightarrow p_1(y) - k p_0(y) \ge 0$$
$$y \in M^- \Rightarrow \varphi_k(y) < 1 \Rightarrow p_1(y) - k p_0(y) < 0$$

$$\Rightarrow \int (\varphi_k - \psi)(p_1 - kp_o)\,\mathrm{d}\mu = \int\limits_{M^+} \underbrace{(\varphi_k - \psi)}_{\geq 0}\underbrace{(p_1 - kp_0)}_{\geq 0}\,\mathrm{d}\mu + \int\limits_{M^-} \underbrace{\underbrace{(\varphi_k - \psi)}_{<0}\underbrace{(p_1 - kp_0)}_{<0}}_{\geq 0}\,\mathrm{d}\mu$$

$$\Rightarrow \int (\varphi_k - \psi)p_1\,\mathrm{d}\mu \geq k\int (\varphi_k - \psi)p_0\,\mathrm{d}\mu \;=\; k\left(\alpha - \underbrace{\int \psi\,\mathrm{d}\mathsf{P}_{\vartheta_0}}_{\leq \alpha}\right) \geq 0$$

$$\Rightarrow \int \varphi_k\,\mathrm{d}\mathsf{P}_{\vartheta_1} \geq \int \psi\,\mathrm{d}\mathsf{P}_{\vartheta_1}, \quad n.V.$$

$\Rightarrow \psi$ hat eine kleinste Güte unter K.

□

Wir hatten $\alpha = \int \varphi_k(y)\,\mathrm{d}\mathsf{P}_{\vartheta_0}$.
Bezeichnen wir $T(y) = \frac{p_1(y)}{p_0(y)}1_{(0,\infty)}(p_0(y))$, so ist $\varphi_k = 1$ für $T(y) \geq k$.

$$\Rightarrow \alpha = \mathsf{P}_{\vartheta_0}(T(y) \geq k) = 1 - \mathsf{P}_{\vartheta_0}(T(y) < k)$$

Ist also die Verteilung von $T(Y)$ unstetig, so nimmt $\mathsf{P}_{\vartheta_0}(T(y) \geq k)$ nicht alle reellen $\alpha \in (0,1)$ an. Damit man auch dann wieder einen optimalen α–Test für beliebige $\alpha \in (0,1)$ bestimmen kann, betrachtet man randomisierte Neyman–Pearson–Tests:

$$\phi_k(y) = \begin{cases} 1 & > \\ \gamma \quad \text{für} \quad p_1(y) & = \quad kp_0(y) \\ 0 & < \end{cases} \qquad (1)$$

mit $0 \leq k < \infty,\ 0 \leq \gamma \leq 1$.

Satz 7.10 (Neyman–Pearson–Lemma) *a) Jeder Test φ_k der Form (1) ist ein bester α–Test für $\alpha = \int \varphi_k\,\mathrm{d}\mathsf{P}_{\vartheta_0}$.*

b) Für jedes $\alpha \in (0,1)$ gibt es einen α–Test (1) mit einem $\gamma \in [0,1]$, d.h. es gibt ein $k = k(\alpha)$ und ein $\gamma = \gamma(\alpha)$ mit $\mathsf{P}_{\vartheta_0}(p_1 > kp_0) + \gamma\mathsf{P}_{\vartheta_0}(p_1 = kp_0) = \alpha$. Ein solcher Test ist ein bester α–Test.

BEWEIS: *a)* ist genauso zu beweisen wie der vorherige Satz.
b) Mit dem T von oben gilt

$$\mathsf{E}_{\vartheta_0}\varphi_k(Y) = \int \varphi_k(y)\,\mathrm{d}\mathsf{P}_{\vartheta_0}(y) = \mathsf{P}_{\vartheta_0}(T(y) > k) + \gamma\mathsf{P}_{\vartheta_0}(T(y) = k)$$

Gilt nun $\mathsf{P}_{\vartheta_0}(T(y) > k) = \alpha$, so sind wir mit $\gamma = 0$ fertig.
Gilt aber $\mathsf{P}_{\vartheta_0}(T(y) > k) < \alpha$ und $\mathsf{P}_{\vartheta_0}(T(y) \geq k) \geq \alpha$, so setzt men γ so, daß $\mathsf{P}_{\vartheta_0}(T(y) > 0) + \gamma\mathsf{P}_{\vartheta_0}(T(y) = k) = \alpha$. Nach *a)* ist φ_k optimal.

□

Bemerkungen:
1. Der optimale Test ist im wesentlichen eindeutig.
Sei φ_0 ein optimaler Neyman-Pearson-Test zum Niveau $\alpha \in (0,1)$. Sei φ^* ebenfalls ein bester α–Test für H gegen K.
$\Rightarrow$ auf $\{y:\ p_1(y) \neq kp_0(y)\}$ gilt $\varphi_0 = \varphi^*\ \ \mu$– f.s.

(folgt wie im Satz 7.9 N–P–Test).
Auf $R_0 = \{y : p_1(y) = kp_0(y)\}$ ist der Test nicht eindeutig.
2. Ein bester α–Test für H gegen K schpft das Niveau α aus ($\int \psi(y)\mathsf{P}_{\vartheta_0}(\,\mathrm{d}y) = \alpha$), oder es gilt $\int \psi \mathsf{P}_{\vartheta_1}(\,\mathrm{d}y) = 1$. Denn angenommen: $\mathsf{E}_\vartheta \psi < \alpha$, $\mathsf{E}_{\vartheta_1}\psi < 1$ und E_{ϑ_1} ist maximal.
$\Rightarrow$ wählen $\eta \in (0,1)$ mit $\mathsf{E}_{\vartheta_0}\psi + \eta \int (1-\psi)\mathsf{P}_{\vartheta_0}(\,\mathrm{d}y) \leq \alpha$
$\Rightarrow$ mit $\psi^* = \psi + \eta(1-\psi)$ gilt

$$\mathsf{E}_{\vartheta_1}\psi^* = \mathsf{E}_{\vartheta_1}\psi + \eta(1 - \mathsf{E}_{\vartheta_1}\psi) > \mathsf{E}_{\vartheta_1}\psi \quad \text{Wdspr.}$$

3. Ein bester α–Test ist unverfälscht ($\mathsf{E}_{\vartheta_1}\psi \geq \alpha$),
denn $\tilde{\psi}(y) \equiv \alpha$ ist α–Test und wegen der Optimalität ist $\mathsf{E}_{\vartheta_1}\psi \geq \mathsf{E}_{\vartheta_1}\tilde{\psi} = \alpha$.

Tests für zusammengesetzte Hypothesen

Definition 7.17 *Eine Hypothese heißt einfach, falls sie nur aus einem Element besteht. Eine Hypothese heißt zusammengesetzt, falls sie aus mehreren Elementen besteht.*

Beispiel 7.16 (Optimalität des Gauß–Tests) Sei

$$Y_1, ..., Y_n \sim N(\vartheta, \sigma_0^2), \quad H:\ \vartheta \leq \vartheta_0,\ K:\ \vartheta > \vartheta_0.$$

Es sei $\tilde{\Psi}_0 = \{\psi:\ \alpha\text{–Test für } \tilde{H}_0 : \vartheta = \vartheta_0 \text{ gegen } K\}$.
Sei nun $\vartheta_1 > \vartheta_0$ beliebig, fest. Das N–P–Lemma gibt einen optimalen α–Test $\psi^* \in \tilde{\Psi}_0$ für $\tilde{H}_0$ gegen die einfache Alternative $\tilde{K}:\ \vartheta = \vartheta_1$, nämlich

$$\psi^* = \begin{cases} 1 & > \\ \gamma \quad \text{falls} \quad \frac{p_1(y)}{p_0(y)} & = k \\ 0 & < \end{cases} \qquad (1)$$

für die Dichten

$$p_j(y) = \frac{1}{(\sqrt{2\pi}\sigma_0)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (y_i - \vartheta_j)^2\right\} \quad j = 0, 1$$

$$\text{und} \quad \mathsf{P}_{\vartheta_0}\left(\frac{p_1(y)}{p_0(y)} > k\right) + \gamma \mathsf{P}_{\vartheta_0}\left(\frac{p_1(y)}{p_0(y)} = k\right) = \alpha. \qquad (2)$$

Es gilt jetzt die Gleichungskette

$$\left\{y:\ \frac{p_1(y)}{p_0(y)} > k\right\} \ = \ \left\{y:\ \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n \left[(y_i - \vartheta_1)^2 - (y_i - \vartheta_0)^2\right]\right\} \ > \ k\right\}$$

$$
\begin{aligned}
&= \left\{ y : \ (\vartheta_1 - \vartheta_0) \sum y_i - \frac{n}{2}(\vartheta_1^2 - \vartheta_0^2) \ > \ \sigma_0^2 \ln k \right\} \\
&= \left\{ y : \ \overline{y} - \frac{1}{2}(\vartheta_1 + \vartheta_0) \ > \ \frac{\sigma_0^2 \ln k}{n(\vartheta_1 - \vartheta_0)} \right\} \\
&= \left\{ y : \ \overline{y} - \vartheta_0 - \frac{1}{2}(\vartheta_1 - \vartheta_0) \ > \ \frac{\sigma_0^2 \ln k}{n(\vartheta_1 - \vartheta_0)} \right\} \\
&= \left\{ y : \ \frac{\sqrt{n}(\overline{y} - \vartheta_0)}{\sigma_0} \ > \ \frac{\sigma_0 \ln k}{\sqrt{n}(\vartheta_1 - \vartheta_0)} + \frac{\sqrt{n}}{2\sigma_0}(\vartheta_1 - \vartheta_0) \right\}
\end{aligned}
$$

(Ebenso für $=, <$.)

Dieser Test ist definiert durch k und γ und zwar durch (2).
Es gilt $T_0(Y) = \frac{\sqrt{n}(\overline{y} - \vartheta_0)}{\sigma_0} \sim N(0,1)$ ist unabhängig von ϑ_1.
Wir haben beim Gauß–Test: für $\tilde{H}$ gegen $\tilde{K}$:

$$
\begin{aligned}
\left\{ y : \ \frac{p_1(y)}{p_0(y)} > k \right\} &= \{ y : \ T_0(y) > l \} \\
\text{für} \ \ T_0(y) &= \frac{\sqrt{n}(\overline{y} - \vartheta_0)}{\sigma_0} \\
\text{und} \qquad l &= \frac{\sigma_0 \ln k}{\sqrt{n}(\vartheta_1 - \vartheta_0)} + \frac{\sqrt{n}}{2\sigma_0}(\vartheta_1 - \vartheta_0)
\end{aligned}
$$

Das gleiche gilt für $=$ und $<$.
Dann folgt

$$
\left\{ y : \ \frac{p_1(y)}{p_0(y)} > k \right\} = \{ y : \ T_0 = (y) > l \}
$$

(analog für $= k, < k$). $\Rightarrow$ (1) ist identisch mit dem Test

$$
\psi^*(y) = \begin{cases} 1 & > \\ \gamma \quad \text{falls} \quad T_0(y) & = c \\ 0 & < \ , \end{cases}
$$

falls c, γ Lsungen von

$$
\begin{aligned}
\mathsf{P}_{\vartheta_0}(T_0(y) > c) + \gamma \mathsf{P}_{\vartheta_0}(T_0(y) = c) &= 1 - \Phi(c) + \gamma \cdot 0 \\
&= 1 - \Phi(c)
\end{aligned}
$$

sind. Bestimmt wird c_α mit $1 - \Phi(c_\alpha) = \alpha$. Setzt man $\gamma = 1$, so ist

$$
\psi^*(y) = \begin{cases} 1 & \geq \\ \quad \text{falls} \quad T_0(y) & \quad c \\ 0 & < \end{cases} \tag{3}
$$

als bester α–Test für $\tilde{H} : \ \vartheta = \vartheta_0$ gegen $\tilde{K} : \ \vartheta = \vartheta_1$. Da im Test ψ^* in (3) kein ϑ_1 vorkommt, ist er für alle $\vartheta_1 > \vartheta_0$ bester Test, d.h. ein bester α–Test für $\tilde{H}$ gegen K. Wir hatten früher beim Gauß–Test gesehen, daß er ein α–Test für H gegen K ist.

Grundidee des Optimalitätsbeweises: Beschreibung des Ablehnungsbereiches $\frac{p_1}{p_0} > k$ bzw. $< k$ für ϑ_0 gegen ϑ_1 in der Gestalt $T_0 > c$ bzw. $< c$ mit einer Statistik T_0, die unabhängig von ϑ_1 ist. Das war möglich, weil $\frac{p_1(y)}{p_0(y)} = h(T_0, \vartheta_0, \vartheta_1)$ für eine streng monoton wachsende Funktion h galt. Viele Verteilungsklassen haben diese wichtige Eigenschaft.

8 Bedingter Erwartungswert und suffiziente Statistiken

Wir benötigen die bedingte Wahrscheinlichkeit. Zur Erinnerung gilt:

$$\mathsf{P}_\vartheta(C|\ A) = \frac{\mathsf{P}_\vartheta(CA)}{\mathsf{P}_\vartheta(A)} \quad \text{für } \mathsf{P}_\vartheta(A) > 0\ .$$

Sei $A = \{y : T(y) = t\} = T^{-1}(t)$ und $\mathsf{P}_\vartheta(A) > 0$.

$$\begin{aligned}
\Longrightarrow \quad \underbrace{\mathsf{P}_\vartheta(C|\ T(y) = t)}_{h_C(t)} &= \frac{\mathsf{P}_\vartheta(CT^{-1}(t))}{\mathsf{P}_\vartheta(T^{-1}(t))} \\
\Longrightarrow \quad \mathsf{P}_\vartheta(CT^{-1}(t)) &= h_C(t)\mathsf{P}_\vartheta(T^{-1}(t)) \\
&= h_C(t)\mathsf{P}_\vartheta{}^T(t) \\
\Longrightarrow \quad \mathsf{P}_\vartheta(CT^{-1}(D)) &= \int\limits_D h_C(t)\mathsf{P}_\vartheta^T(\,\mathrm{d}t) \\
&= \int\limits_{T^{-1}(D)} h_C(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) \quad \forall\, D \in \mathfrak{D} \qquad (4)
\end{aligned}$$

Wegen $\mathsf{P}_\vartheta(CT^{-1}(D)) = \int\limits_{T^{-1}(D)} I_C(y)\mathsf{P}_\vartheta(\,\mathrm{d}y)$ folgt

$$\int\limits_{T^{-1}(D)} I_C(y)\mathsf{P}_\vartheta(\,\mathrm{d}y) = \int\limits_{T^{-1}(D)} h_C(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) \quad \forall\, D \in \mathfrak{D}. \qquad (5)$$

Es gilt: Es existiert für beliebige $C \in \mathfrak{C}$ immer eine $\mathfrak{D}$–meßbare Funktion h_C mit (5). Wenn man $\mathfrak{F} = \{T^{-1}(D),\ D \in \mathfrak{D}\}$ bezeichnet, so gilt $\mathfrak{F} \subseteq \mathfrak{C}$ und h_C ist eindeutig bestimmt bis auf P_ϑ–Nullmengen von $\mathfrak{F}$. Das rechtfertigt die Definition:

Definition 8.1 *$h_C(t)$ heißt bedingte Wahrscheinlichkeit von C bei gegebenem t.*

Dieser Begriff läßt sich erweitern. Jede nichtnegative $\mathfrak{C}$–meßbare Funktion G ist Grenzwert primitiver Funktionen:

$$G(y) = \lim \sum_{i=1}^{n} \gamma_i I_{C_i}(y) \quad ,\ C_i \in \mathfrak{C},\ \gamma_i \geq 0\ .$$

Damit ist begründet

$$\begin{aligned}
\int\limits_{T^{-1}(D)} G(y)\mathsf{P}_\vartheta(\,\mathrm{d}y) &= \lim_{n\to\infty}\sum_{i=1}^{n}\gamma_i \int\limits_{T^{-1}(D)} I_{C_i}(y)\mathsf{P}_\vartheta(\,\mathrm{d}y) \\
&= \lim_{n\to\infty}\sum_{i=1}^{n}\gamma_i \int\limits_{T^{-1}(D)} h_{C_i}(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) \\
&= \int\limits_{T^{-1}(D)} \lim_{n\to\infty}\sum_{i=1}^{n}\gamma_i h_{C_i}(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) \\
&= \int\limits_{T^{-1}(D)} l(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y)
\end{aligned}$$

für die $\mathfrak{D}$–meßbare Funktion l, d.h.

$$\int\limits_{T^{-1}(D)} G(y)\mathsf{P}_\vartheta(\,\mathrm{d}y) = \int\limits_{T^{-1}(D)} l(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y)\ . \qquad \forall\, D\in\mathfrak{D} \tag{6}$$

Da sich jede $\mathfrak{C}$–meßbare Funktion als Differenz zweier nichtnegativer $\mathfrak{C}$–meßbarer Funktionen darstellen läßt, überträgt sich die letzte Beziehung für beliebige $\mathfrak{C}$–meßbare Funktionen G. Aus (6) folgt insbesondere für $D = t$:

$$\int\limits_{T^{-1}(t)} G(y)\mathsf{P}_\vartheta(\,\mathrm{d}y) = \int\limits_{T^{-1}(t)} l(t)\mathsf{P}_\vartheta(\,\mathrm{d}y) = l(t)\int\limits_{T^{-1}(t)} \mathsf{P}_\vartheta(\,\mathrm{d}y)$$

$$\Rightarrow \qquad l(t) = \int\limits_{T^{-1}(t)} G(y)\frac{1}{\int\limits_{T^{-1}(t)}\mathsf{P}_\vartheta(\,\mathrm{d}z)}\mathsf{P}_\vartheta(\,\mathrm{d}y) = l_\vartheta(t)$$

$\Rightarrow$ $l_\vartheta(t)$ ist Erwartungswert bezüglich einer Verteilung, deren Träger $T^{-1}(t)$ ist.

$\Rightarrow$ $l_\vartheta(t)$ ist der Erwartungswert von $G(Y)$ unter der Bedingung, daß Y Werte in $T^{-1}(t)$ hat.

$$l_\vartheta(t) := \mathsf{E}_\vartheta(G(Y)|t)$$

Definition 8.2 *Sei G eine beliebige $\mathfrak{C}$–meßbare Funktion. Jede $\mathfrak{D}$–meßbare Funktion $l_\vartheta = l$ mit (6) heißt bedingter Erwartungswert von $G(Y)$ bei gegebenem t.*

Satz 8.1 *Ist $l_\vartheta(t) = \mathsf{E}_\vartheta(G(Y)|t)$, so gilt*

$$\mathsf{E}_\vartheta G(Y) = \mathsf{E}_\vartheta l(T(Y))\ .$$

Beweis: Für $D = \mathcal{T}$ gilt

$$\int\limits_{\mathcal{Y}} G(y)\mathsf{P}_\vartheta(\,\mathrm{d}y) = \int\limits_{\mathcal{Y}} l(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y)$$

$$\Rightarrow \quad \mathsf{E}_\vartheta G(Y) \;=\; \mathsf{E}_\vartheta l(T(Y)) \;.$$

□

Zur Berechnung des bedingten Erwartungswertes ist es manchmal hilfreich, die Atome einer σ–Algebra zu bestimmen. Wir wissen, die Bestimmungsgleichung des bedingten Erwartungswertes ist

$$\int\limits_{T^{-1}(D)} G(y)\mathsf{P}_\vartheta(\,\mathrm{d}y) \;=\; \int\limits_{T^{-1}(D)} l(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) \qquad\qquad \forall\, D \in \mathfrak{D}.$$

und l ist eine $\mathfrak{F}$–meßbare Funktion für $\mathfrak{F} = T^{-1}(\mathfrak{D})$.

Definition 8.3 *A ist Atom von $\mathfrak{F}$, wenn $A \in \mathfrak{F}$ und aus $B \in \mathfrak{F},\ B \neq \emptyset,\ B \subseteq A$ folgt $A = B$.*

Dann gilt: Eine $\mathfrak{F}$–meßbare Funktion ist auf den Atomen von $\mathfrak{F}$ konstant. Besteht $\mathfrak{F}$ aus abzählbar vielen Atomen, so ist eine auf den Atomen konstante Funktion $\mathfrak{F}$–meßbar.

Beispiel 8.1 Sei $T = const \;\Rightarrow\; \mathfrak{F} = \{\emptyset, \mathcal{Y}\}$. Der bedingte Erwartungswert von $G(Y)$ bei gegebenem t muß also auf $\mathcal{Y}$ konstant sein.

$$l(T(y)) = \mathsf{E}_\vartheta G(Y).$$

Definition 8.4 *T heißt suffizient, falls $l_\vartheta(t)$ beziehungsweise $h_C(t)$ unabhängig von ϑ gewählt werden kann.*

Motivation: Bei festem t steckt in y mit $T(y) = t$ keine Information über ϑ.

Beispiel 8.2 $Y_1, \ldots, Y_n$ unabhängig, $\qquad \sim \mathsf{B}(1, \vartheta)$
Sei $\mathsf{P}_\vartheta = \mathsf{P}^Y \quad$ für $Y = (Y_1, \ldots, Y_n)$,
$\mathsf{P}_\vartheta(y) = \prod_{i=1}^n \mathsf{P}_\vartheta{}^{Y_i}(y_i) = \vartheta^{\Sigma y_i}(1-\vartheta)^{n-\Sigma y_i}$ bei $y \in \{0,1\}^n$.

Es gilt für $T(y) = \sum y_i$

$$\mathsf{E}_\vartheta\big(I_C(Y) \mid T(Y) = t\big) = \begin{cases} \dfrac{\sum\limits_{y:T(y)=t} I_C(y)\mathsf{P}_\vartheta(y)}{\mathsf{P}^T_\vartheta(t)} & \text{für} \quad \mathsf{P}^T_\vartheta(t) > 0 \\[2ex] \text{beliebig} & \text{sonst} \end{cases}$$

$$\Rightarrow \mathsf{E}_\vartheta\big(I_C(Y) \mid T(Y) = t\big) = \begin{cases} \dfrac{\sum\limits_{y:T(y)=t} I_C(y)\vartheta^t(1-\vartheta)^{n-t}}{\sum\limits_{y:T(y)=t} \vartheta^t(1-\vartheta)^{n-t}} & \text{für Nenner } > 0 \\[2ex] \text{beliebig} \end{cases}$$

Für $\mathsf{P}^T_\vartheta(t) > 0 \quad$ gilt:

$$\frac{\sum\limits_{y:T(y)=t} I_C(y)\vartheta^t(1-\vartheta)^{n-t}}{\sum\limits_{y:T(y)=t} \vartheta^t(1-\vartheta)^{n-t}} = \frac{1}{\binom{n}{t}} \sum_{y:T(y)=t} I_C(y)$$

Wir bemerken, daß $\mathsf{E}_\vartheta\big(I_C(Y) \mid T(Y)=t\big)$ unabhängig von ϑ ist.

Die grundlegende Beziehung (6) besagt, daß auf der rechten Seite eine $\mathfrak{F}$–meßbare, und auf der linken Seite eine $\mathfrak{C}$–meßbare Funktion unter dem Integral steht. Wenn also G selbst schon $\mathfrak{F}$–meßbar ist, so ist $G(y) = l(T(y))$. Das ist eine Konsequenz aus folgendem Satz:

Satz 8.2 *Für jede Funktion $L(T(\cdot))$ gilt*

$$\int\limits_{T^{-1}(D)} G(y)L(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) \;=\; \int\limits_{T^{-1}(D)} l(T(y))L(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) \qquad \forall\, D.$$

$$\textit{d.h.} \qquad \mathsf{E}(G(Y)L(T(Y))|t) \;=\; l(t)L(t) \;=\; L(t)\mathsf{E}(G(Y)|t)\ .$$

Beweis folgt aus (6)

Der nun folgende Faktorisierungssatz von Neyman gibt ein einfaches Kriterium zur Prüfung, ob eine Statistik suffizient ist.

Satz 8.3 (Faktorisierungssatz) *Sei $\mathcal{P}$ dominiert durch ein Maß μ und seien p_ϑ die Dichten von P_ϑ bezüglich μ. Falls für eine Statistik T*

$$p_\vartheta(y) \;=\; g_\vartheta(T(y))h(y) \qquad [\mathfrak{C},\mu] \tag{7}$$

für eine $\mathfrak{D}$–meßbare Funktion g_ϑ, so ist T suffizient.

Beweis: Wir wissen $\frac{\mathrm{d}\mathsf{P}_\vartheta}{\mathrm{d}\mu} = g(T(y))h(y)$
Es gibt dann ein zu $\mathcal{P}$ äquivalentes Maß $\tilde{\mu}$ der Gestalt

$$\tilde{\mu}(C) = \sum_{i=0}^{\infty} \rho_i \mathsf{P}_{\vartheta i}(C) \quad ,\ \rho_i > 0,\ \ \sum \rho_i = 1,\ \mathsf{P}_{\vartheta i} \in \mathcal{P}\ .$$

Folglich gilt wegen $\frac{\mathrm{d}\mathsf{P}_\vartheta}{\mathrm{d}\mu} = \frac{\mathrm{d}\mathsf{P}_\vartheta}{\mathrm{d}\tilde{\mu}}\frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\mu}$

$$g_\vartheta(T(y))h(y) \;=\; \frac{\mathrm{d}\mathsf{P}_\vartheta}{\mathrm{d}\tilde{\mu}} \cdot \sum_{i=1}^{\infty} \rho_i g_{\vartheta i}(T(y))h(y) \tag{8}$$

Hieraus folgt auch $\tilde{\mu}(\{y:\ h(y)=0\}) = 0$, woraus wegen (8)

$$g_\vartheta(T(y)) = \frac{\mathrm{d}\mathsf{P}_\vartheta}{\mathrm{d}\tilde{\mu}} \sum_{i=1}^{\infty} \rho_i g_{\vartheta i}(T(y)) \quad [\mathfrak{C},\tilde{\mu}] \tag{9}$$

gilt. Da wegen der quivalenz von $\tilde{\mu}$ zu $\mathcal{P}$ und wegen (7)

$$\tilde{\mu}\left(y:\ \sum_{i=1}^{\infty} \rho_i g_{\vartheta i}(T(y)) = 0\right) \;=\; 0$$

gilt, erhalten wir aus (9)

$$\frac{\mathrm{d}\mathsf{P}_\vartheta}{\mathrm{d}\tilde{\mu}} = l_\vartheta(T(y)).$$

Offenbar ist $l_\vartheta(T(y))$ abhängig von y nur über $T(y)$. Sei $G(Y)$ eine beliebige Zufallsgröße, wobei $\mathsf{E}_\vartheta G(Y)$ existiert. Dann existiert auch der bedingte Erwartungswert $k(t)$ von $G(Y)$ bezüglich $\tilde{\mu}$ bei gegebenem t:

$$\int\limits_{T^{-1}(D)} G(y)\tilde{\mu}(\,\mathrm{d}y) = \int\limits_{T^{-1}(D)} k(T(y))\tilde{\mu}(\,\mathrm{d}y) \quad \forall\, D \in \mathfrak{D}\ .$$

Wir zeigen, daß $k(T(y))$ auch bedingter Erwartungswert von $G(Y)$ bezüglich beliebigem P_ϑ ist.

$$\begin{aligned}
\int\limits_{T^{-1}(D)} k(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) &= \int\limits_{T^{-1}(D)} k(T(y))l_\vartheta(T(y))\tilde{\mu}(\,\mathrm{d}y) \\
&= \int\limits_{T^{-1}(D)} G(y)l_\vartheta(T(y))\tilde{\mu}(\,\mathrm{d}y) \\
&= \int\limits_{T^{-1}(D)} G(y)\mathsf{P}_\vartheta(\,\mathrm{d}y) \quad \forall\, D \in \mathfrak{D}
\end{aligned}$$

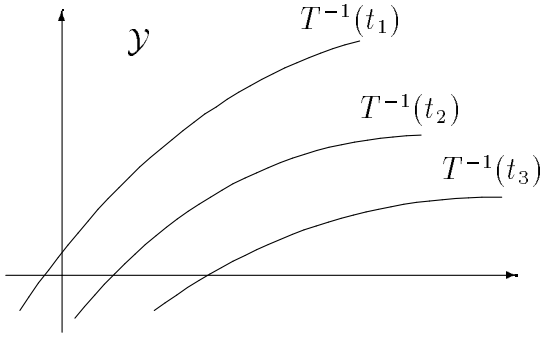
Hier wurde im wesentlichen Satz 8.2 benutzt.

□

Beispiel 8.3 Sei $Y \sim N(X\beta, \sigma^2\Sigma)$. Dann ist $\left(P_{\mathcal{R}(X)}\hat{\beta}, s^2\right)$ suffizient.

Merke: Der bedingte Erwartungswert ist eine Verallgemeinerung des Satzes von Fubini. Denn: Wir haben

$$\begin{aligned}
\mathsf{E}G(Y) = \int G(y)\mathsf{P}_\vartheta(\,\mathrm{d}y) &= \int l(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) = \int l(t)\mathsf{P}_\vartheta{}^T(\,\mathrm{d}t) \\
&= \int \underbrace{\left(\frac{\int\limits_{T^{-1}(t)} G(y)\mathsf{P}_\vartheta(\,\mathrm{d}y)}{\int\limits_{T^{-1}(t)} \mathsf{P}_\vartheta(\,\mathrm{d}y)}\right)}_{\text{Integration entlang } T^{-1}(t)} \mathsf{P}_\vartheta{}^T(\,\mathrm{d}t)
\end{aligned}$$



Im Satz von Fubini gilt:

$$\int\int f(y_1,y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 \;=\; \int \underbrace{\left[\int f(y_1,y_2)\,\mathrm{d}y_1\right]}_{\text{Integration entlang } y_2 = \text{const}} \mathrm{d}y_2$$

Beispiel 8.4 Sei $\mathcal{P} = \{\mathsf{P}_\vartheta : \vartheta \in \Theta\}$ und diese Familie sei dominiert durch ein σ-endliches Maß μ. Dann ex. die Dichten $p_\vartheta(y)$. Sei $y \in \mathbb{R}^1$. Wenn $T(y) = \mid y \mid = t$, haben wir gezeigt:

$$h_\vartheta(t) = \begin{cases} \frac{\mathbf{I}_C(t)p_\vartheta(t)+I_C(-t)p_\vartheta(-t)}{p_\vartheta(t)+p_\vartheta(-t)} & \text{für} \quad p_\vartheta(t)+p_\vartheta(-t) < 0 \\ \frac{1}{2}\left(I_C(t)+I_C(-t)\right) & \text{sonst.} \end{cases}$$

Dieses $h_\vartheta(t)$ ist von ϑ abhängig. Falls P_ϑ symmetrisch, d.h. $p_\vartheta(t) = p_\vartheta(-t)$, so ist $h_\vartheta(t) = \frac{1}{2}\left(I_C(t)+I_C(-t)\right)$ eine Festlegung des bedingten Erwartungswertes. Wegen der Unabhängigkeit von ϑ folgt: $T(y) = \mid y \mid$ ist für symmetrische Verteilungen suffizient.
Schreibweise bei Unabhängigkeit von ϑ: $h_\vartheta(t) = h_\cdot(t)$

Satz 8.4 *Ist e eine $1,1$-Abb. von $\mathcal{T} = T(\mathcal{Y})$ auf $\tilde{\mathcal{T}} = \tilde{T}(\mathcal{Y})$ mit $\tilde{T}(y) = e\left(T\left(y\right)\right)$ und gilt $\tilde{D} = e(D)$ und ist $T(y)$ suffizient, so ist auch $\tilde{T}(y)$ suffizient.*

BEWEIS: Es ist $h_C(t) = \mathsf{P}_\vartheta\left(y \in C \mid T(y) = t\right)$ unabhängig von ϑ.
$\Rightarrow k_C(Y) = h_C(T(y))$ ist unabhängig von ϑ
Daher ist $h_C(T(y))$ Lösung von

$$\int\limits_{T^{-1}(D)} k(y)\mathsf{P}^Y(dy) \;=\; \int\limits_{T^{-1}(D)} g(y)\mathsf{P}^Y(dy)\,,$$

wegen $T^{-1}(D) = T^{-1}e^{-1}e(D) = \tilde{T}^{-1}(\tilde{D})$ ist $k(y)$ auch Lösung von

$$\int\limits_{\tilde{T}^{-1}(\tilde{D})} k(y)\mathsf{P}^Y(dy) \;=\; \int\limits_{\tilde{T}^{-1}(\tilde{D})} g(y)\mathsf{P}^Y(dy)\,.$$

Wegen $k(y) = h_C\left(T(y)\right) = h_C\left(e^{-1}\tilde{T}(y)\right) = \tilde{h}_C\left(\tilde{T}(y)\right)$
ist $\tilde{h}_C(\tilde{t})$ bedingter Erwartungswert von $I_C(Y)$ bei gegebenem $\tilde{T}(y) = \tilde{t}$ und unabhängig von ϑ.

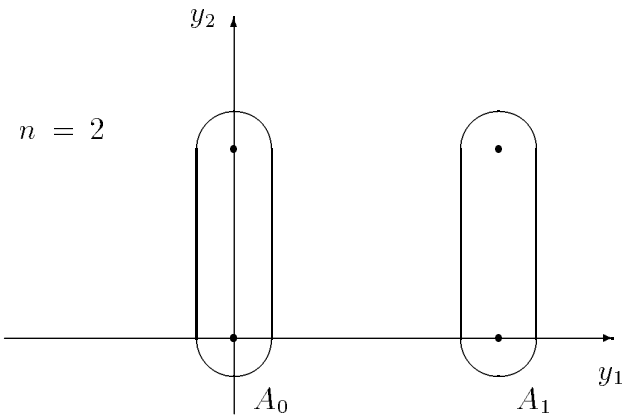
□

Betrachten wir noch einmal den bedingten Erwartungswert. Gegeben ist Y mit der möglichen Verteilungsfamilie $\mathcal{P}$. Bezüglich eines Maßes μ seien die Dichten $p_\vartheta(y)$. Dann hatten wir die bedingte Wahrscheinlichkeit einer Menge C bei gegebenen $t \in \mathcal{T}$. Das war $h_{C,\vartheta}(t)$ und mußte die Gleichung

$$\int\limits_{T^{-1}(D)} I_C(y)\mathsf{P}_\vartheta(\,\mathrm{d}y) \;=\; \int\limits_{T^{-1}(D)} h_{C,\vartheta}(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y)$$

für eine gegebene Statistik $T\colon (\mathcal{Y},\mathfrak{C}) \to (\mathcal{T},\mathfrak{D})$ erfüllen.
$\Rightarrow$ es ist natürlich, daß $h_{C,\vartheta}$ über $\mathcal{T}$ definiert ist. Damit muß sie bezüglich $\mathfrak{D}$ meßbar sein: $h_{C,\vartheta} : (\mathcal{T},\mathfrak{D}) \to (\mathbb{R}^1,\mathfrak{B}^1)$ (genauer: $\mathbb{R}^1 \cap [0,1]$).
Es gilt $h_{c,\vartheta}^{-1}([a,b]) \in \mathfrak{D}$. Offenbar ist $h_{C,\vartheta}$ über $\mathcal{Y}$ definiert, ist aber nicht $\mathfrak{C}$–meßbar. $h_{C,\vartheta}(T)$ ist $\mathfrak{F}$–meßbar für $\mathfrak{F} = T^{-1}(\mathfrak{D})$.

Beispiel 8.5 $Y_1,...,Y_n$ unabhängig, $Y_i \sim \mathsf{B}(1,\vartheta)$, $\mathsf{P}_\vartheta = \mathsf{P}^Y$ für $Y = (Y_1,...,Y_n)$
Es gilt $\mathsf{P}_\vartheta(y) = \prod\limits_{i=1}^{n} \mathsf{P}_\vartheta{}^{Y_i}(y_i) = \vartheta^{\sum y_i}(1-\vartheta)^{n-\sum y_i}$ für $y \in \{0,1\}^n$. Hier ist also $\mathcal{Y} = \{0,1\}^n$, $\mathfrak{C} = \mathfrak{P}(\mathcal{Y})$. Für $T(y) = y_1$ gilt dann: $\mathfrak{F}$ hat 4 Elemente:
$\mathfrak{F} = \{\emptyset, \mathcal{Y}, A_0, A_1\}$ für $A_0 = \{y : y_1 = 0\} = T^{-1}(0)$, $A_1 = \{y : y_1 = 1\} = T^{-1}(1)$



Eine Funktion $h_{C,\vartheta}^{(1)}(t)$ kann dann keine unterschiedlichen Werte auf den beiden Punkten von A_0 bzw. von A_1 haben. Es muß ein $h_{C,\vartheta}^{(1)}(t)$ bestimmt werden mit

$$\int\limits_{F} I_C(y)\mathsf{P}_\vartheta(\,\mathrm{d}y) \;=\; \int\limits_{F} h_{C,\vartheta}^{(1)}(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) \qquad \forall F \in \mathcal{F}\,.$$

$F = \emptyset$:

klar

$F = \mathcal{Y}$:

$$\begin{aligned} \mathsf{P}_\vartheta(C) &= \int\limits_{A_0} h^{(1)}_{C,\vartheta}(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) + \int\limits_{A_1} h^{(1)}_{C,\vartheta}(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) \\ &= h^{(1)}_{C,\vartheta}(0)\mathsf{P}_\vartheta(A_0) + h^{(1)}_{C,\vartheta}(1)\mathsf{P}_\vartheta(A_1) \end{aligned}$$

$F = A_0$:

$$\mathsf{P}_\vartheta(CA_0) = \int\limits_{A_0} h^{(1)}_{C,\vartheta}(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) = h_{C,\vartheta}(0)\mathsf{P}_\vartheta(A_0)$$

$F = A_1$:

$$\mathsf{P}_\vartheta(CA_1) = \int\limits_{A_1} h^{(1)}_{C,\vartheta}(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y) = h_{C,\vartheta}(1)\mathsf{P}_\vartheta(A_1)\,.$$

Es gilt

$$\mathsf{P}_\vartheta(A_0) = \mathsf{P}_\vartheta{}^{Y_1}(0) = 1-\vartheta \quad , \; \mathsf{P}_\vartheta(A_1) = \mathsf{P}_\vartheta{}^{Y_1}(1) = \vartheta\,.$$

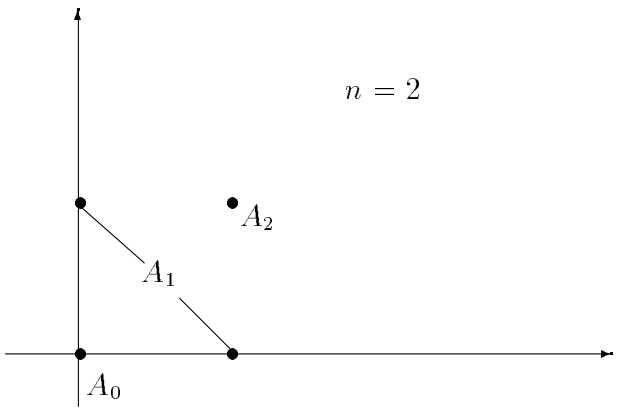
$$\Rightarrow \quad h^{(1)}_{C,\vartheta}(0) = \frac{\mathsf{P}_\vartheta(CA_0)}{1-\vartheta}\,, \qquad h^{(1)}_{C,\vartheta}(1) = \frac{\mathsf{P}_\vartheta(CA_1 0)}{\vartheta}$$

$$= \frac{\mathsf{P}_\vartheta(CA_0)}{\mathsf{P}_\vartheta(A_0)} \qquad = \frac{\mathsf{P}_\vartheta(CA_1)}{\mathsf{P}_\vartheta(A_1)}$$

$$\text{Damit ist} \quad \begin{aligned} h^{(1)}_{C,\vartheta}(0) &= \mathsf{P}_\vartheta(C|A_0)\,, \\ h^{(1)}_{C,\vartheta}(1) &= \mathsf{P}_\vartheta(C|A_1)\,. \end{aligned}$$

Die Funktion $h^{(1)}_{C,\vartheta}$ gibt also die bedingten Wahrscheinlichkeiten für die unterschiedlichen Bedingungen an. Es kann aber auch als Funktion $T(y) = \sum y_i$ genommen werden.
$\mathcal{Y} = \{0,1\}^n\,,\ \mathfrak{C} = \mathfrak{P}(\mathcal{Y})\,,\ \mathcal{T} = \{0,1,...,n\}\,,\ \mathfrak{D} = \mathfrak{P}(\mathcal{T})\,,$
$T^{-1}(\mathfrak{D}) = \sigma$–Algebra von $T^{-1}(i)\,,\ i = 0,...,n.$
Auf alle Fälle gehören die y mit $T(y) = const$ zu einer Menge aus $T^{-1}(\mathfrak{D})$.

Bei $n = 2$ haben wir für $\mathfrak{F}$ die Potenzmenge von A_0, A_1, A_2 mit $A_0 = \{(0,0)\}$, $A_1 = \{(0,1),(1,0)\}$, $A_2 = \{(1,1)\}$. (disjunkte Mengen)



Es gilt $\mathsf{P}_\vartheta(A_0) = (1-\vartheta)^2$, $\mathsf{P}_\vartheta(A_1) = 2\vartheta(1-\vartheta)$, $\mathsf{P}_\vartheta(A_2) = \vartheta^2$.
Wir haben $h_{C,\vartheta}(t)$ zu bestimmen aus

$$\begin{aligned} \int\limits_{A_i} I_C(y)\mathsf{P}_\vartheta(\,\mathrm{d}y) &= \int\limits_{A_i} h_{C,\vartheta}(T(y))\mathsf{P}_\vartheta(\,\mathrm{d}y)\,, \\ \text{d.h.} \qquad \mathsf{P}_\vartheta(CA_i) &= h_{C,\vartheta}(i)\mathsf{P}_\vartheta(A_i) \\ \Rightarrow \quad h_{C,\vartheta}(i) &= \frac{\mathsf{P}_\vartheta(CA_i)}{\mathsf{P}_\vartheta(A_i)}\,. \end{aligned}$$

Auch hier geben $h_{C,\vartheta}(t)$ die bedingten Wahrscheinlichkeiten für die Werte $t = 0\,,1\,,2$ an. Aber $h^{(1)}_{C,\vartheta}$ und $h_{C,\vartheta}$ sind beide nicht ineinander überführbar. Es sind unterschiedliche Definitionsbereiche für die beiden bedingten Wahrscheinlichkeitsfunktionen.
Berechnung der bedingten Wahrscheinlichkeiten bei $n = 2$.

$$\text{Wir betrachten } h^{(1)}_{C,\vartheta}(0) = \frac{\mathsf{P}_\vartheta(CA_0)}{1-\vartheta}$$

$$\begin{aligned} \text{Sei } C_1 &= \{(0,0)\} && \Rightarrow h^{(1)}_{C_1,\vartheta}(0) = \tfrac{\mathsf{P}_\vartheta((0,0))}{1-\vartheta} = \tfrac{(1-\vartheta)^2}{1-\vartheta} = 1-\vartheta\,, \\ C_2 &= \{(0,0),(0,1)\} && \Rightarrow h^{(1)}_{C_2,\vartheta}(0) = 1 \\ C_3 &= \{(0,0),(1,0)\} && \Rightarrow h^{(1)}_{C_3,\vartheta}(0) = \tfrac{\mathsf{P}_\vartheta((0,0))}{1-\vartheta} = 1-\vartheta\,, \\ C_4 &= \{(1,1)\} && \Rightarrow h^{(1)}_{C_4,\vartheta}(0) = 0\,. \end{aligned}$$

jetzt betrachten wir $h_{C,\vartheta}(i) \;=\; \dfrac{\mathsf{P}_\vartheta(CA_i)}{\mathsf{P}_\vartheta(A_i)}$.

Für obige Menge n gilt:
$$h_{C_1,\vartheta}(0) = 1\,,\ h_{C_2,\vartheta}(0) = \frac{(1-\vartheta)^2}{(1-\vartheta)^2} = 1$$
$$h_{C_3,\vartheta}(0) = 1\,,\ h_{C_4,\vartheta}(0) = 0\ .$$
$$h_{C_1,\vartheta}(1) = 0\,,\ h_{C_2,\vartheta}(1) = \frac{\vartheta(1-\vartheta)}{2\vartheta(1-\vartheta)} = \frac{1}{2}$$
$$h_{C_3,\vartheta}(1) = \frac{1}{2}\,,\ h_{C_4,\vartheta}(1) = 0$$

Bemerkenswert: $h^{(1)}_{C,\vartheta}$ hängen von ϑ ab,
$h_{C,\vartheta}$ hängen nicht von ϑ ab.

Das liegt darin begründet, daß $T(y) = \sum y_i$ suffizient, hingegen $T(y) = y$ nicht suffizient ist.

Bedeutung suffizienter Statistiken

Wir haben gesehen, daß aus einer Statistik g bei Kenntnis einer suffizienten Statistik T eine neue Statistik $k(y) = h(T(y))$ konstruiert werden kann, die der bedingte Erwartungswert von $g(Y)$ ist und unabhängig von ϑ ist. Das hat eine besondere Bedeutung für die Konstruktion von statistischen Entscheidungsfunktionen. Wir nehmen als Beispiel die erwartungstreue Schätzung eines Parameters $\gamma(\vartheta) \in \mathbb{R}^1$.
Es sei g eine erwartungstreue Schätzfunktion für $\gamma(\vartheta)$, d.h.

$$\mathsf{E}_\vartheta g(Y) = \gamma(\vartheta) \qquad \forall \vartheta \in \Theta.$$

Wenn g_1, g_2 erwartungstreue Schätzfunktionen sind, so heißt g_1 nicht schlechter als g_2, falls
$\mathsf{Var}_\vartheta g_1(Y) \le \mathsf{Var}_\vartheta g_2(Y) \quad \forall \vartheta \in \Theta$, sie heißen „besser“, falls $<$ für wenigstens ein ϑ gilt.

Satz 8.5 (Rao-Blackwell) *Es sei T eine suffiziente Statistik für $\vartheta \in \Theta$. Dann gibt es zu jeder erwartungstreuen Schätzfunktion g eines eindimensionalen Parameters $\gamma(\vartheta)$ eine erwartungstreue Schätzfunktion k mit $k(y) = h(T(y))$, nämlich eine von ϑ unabhängige Festlegung des bedingten Erwartungswert $h(t)$ von $g(Y)$ bei gegebenem $T(y) = t$,*

$$h(T(y)) = \mathsf{E}_\cdot\left(g(Y) \mid T(y)\right).$$

Diese Schätzfunktion hat eine gleichmäßig nicht-größere Varianz, d.h.

$$\mathsf{Var}_\vartheta h\left(T(Y)\right) \le \mathsf{Var}_\vartheta g(Y), \qquad \vartheta \in \Theta \tag{10}$$

Ist $\mathsf{Var}_\vartheta g(Y) < \infty$, so steht in (10) das Gleichheitszeichen gdw. $g(y) = h(T(y))$ $[\mathfrak{C}, \mathsf{P}_\vartheta]$.

BEWEIS: Da g erwartungstreu ist, gilt dies auch für $h(T)$, d.h.

$$\begin{aligned}
\mathsf{E}_\vartheta h(T(Y)) &= \mathsf{E}_\vartheta \mathsf{E}_\cdot\big(g(Y) \mid T(Y)\big) = \gamma(\vartheta) \\
\mathsf{Var}_\vartheta g(Y) &= \mathsf{E}_\vartheta\big(g(Y) - \gamma(\vartheta)\big)^2 = \mathsf{E}_\vartheta\big(g(Y) - h(T(Y)) + h(T(Y)) - \gamma(\vartheta)\big)^2 \\
&= \mathsf{E}_\vartheta\big(g(Y) - h(T(Y))\big)^2 + \mathsf{E}_\vartheta\big(h(T(Y)) - \gamma(\vartheta)\big)^2 \\
&+ 2\mathsf{E}_\vartheta\big(g(Y) - h(T(Y))\big)\big(h(T(Y)) - \gamma(\vartheta)\big)
\end{aligned}$$

Nun haben wir mit Satz 8.2

$$\mathsf{E}_\vartheta\big(g(Y)\underbrace{\big(h(T(Y))) - \gamma(\vartheta)\big)}_{T^{-1}(\mathfrak{D})\text{-meßbar}} \mid t\big) = \big(h(t) - \gamma(\vartheta)\big)h(t)\,.$$

Durch Erwartungswertbildung auf beiden Seiten ergibt sich

$$\begin{aligned}
\mathsf{E}_\vartheta\big(g(Y)(h(T(Y))) - \gamma(\vartheta))\big) &= \mathsf{E}_\vartheta h(T(Y)))(h(T(Y)) - \gamma(\vartheta)) \\
\Rightarrow \mathsf{E}_\vartheta\big(g(Y) - h(T(Y))\big)\big(h(T(Y)) - \gamma(\vartheta)\big) &= 0
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \mathsf{Var}_\vartheta g(Y) &= \underbrace{\mathsf{E}_\vartheta\big(h(T(Y)) - \gamma(\vartheta)\big)^2}_{\mathsf{Var}_\vartheta h(T(Y))} + \mathsf{E}_\vartheta\big(g(Y) - h(T(Y))\big)^2 \\
&\geq \mathsf{Var}_\vartheta h(T(Y))\,.
\end{aligned}$$

Wenn nun in (10) das "=" steht, so bedeutet das $\mathsf{E}_\vartheta\big(g(Y) - h(T(Y))\big)^2 = 0$.
$\Rightarrow\ g(y) = h(T(Y))\quad \mathsf{P}_\vartheta$-fast sicher.

□

Damit bedeutet der Satz von Rao-Blackwell, daß aus einer Schätzfunktion g eine bessere konstruiert wird, wenn nicht schon $g(y) = h(T(y))$ gilt.

Beispiel 8.6 Binomialverteilung $Y_1, ..., Y_n$ i.i.d., $\mathsf{P}(Y_i = 1) = \vartheta$
Wir wissen, daß $T(y) = \sum_{i=1}^n y_i$ suffizient ist. Es gilt $\mathsf{P}(Y_i = 1) = \mathsf{E}Y_i = \vartheta$. Sei $g(y) = y_1$.
$\Rightarrow g(Y)$ ist erwartungstreu für ϑ, ebenso ist $\frac{1}{n}\sum_{i=1}^n Y_i$ erwartungstreu für ϑ.
Da T suffizient ist, gilt

$$\mathsf{E}_\vartheta\big(g(Y) \mid T(y)\big) = \mathsf{E}_\vartheta\big(Y_1 \mid T(y)\big) = \mathsf{E}_\cdot\big(Y_1 \mid T(y)\big)\,.$$

Wir berechnen den bedingten Erwartungswert $h(t)$:

$$\begin{aligned}
h(t) = \frac{\sum\limits_{y:\,T(y)=t} y_1 P^Y(y)}{\sum\limits_{y:\,T(y)=t} P^Y(y)} &= \frac{1\cdot P_\vartheta(y_1=1, \sum_{i=2}^n y_i = t-1) + 0\cdot P_\vartheta\left(y_1=0, \sum\limits_{i=2}^n y_i = t\right)}{\binom{n}{t}\vartheta^t(1-\vartheta)^{n-t}} \\
&= \frac{\vartheta\binom{n-1}{t-1}\vartheta^{t-1}(1-\vartheta)^{(n-1)-(t-1)}}{\binom{n}{t}\vartheta^t(1-\vartheta)^{n-t}} = \frac{\binom{n-1}{t-1}}{\binom{n}{t}} = \frac{t}{n}.
\end{aligned}$$

$\Rightarrow h(T(y) = \frac{1}{n}\sum_{i=1}^{n} y_i$ ist erwartungstreu und es gilt

$$\mathsf{Var}_{\vartheta}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \leq \mathsf{Var}_{\vartheta} Y_1$$

und es gilt "<".

Vollständige Statistiken

Die Suffizienz von Statistiken wurde damit motiviert, daß man optimale Schätzfunktionen nur in Abhängigkeit der suffizienten Statistik zu suchen braucht. Außerdem liegt häufig eine wesentlich reduzierte Dimension der suffizienten Statistik im Vergleich zur Originaldimension vor. Eine andere Eigenschaft von Statistiken ist die Vollständigkeit.

Definition 8.5 *Es sei $\mathcal{P} = \{\mathsf{P}_\vartheta : \vartheta \in \Theta\}$ eine Familie von Wahrscheinlichkeitsverteilungen über $(\mathcal{Y}, \mathfrak{C})$. Es sei $T(y)$ eine Statistik mit Werten in $(\mathcal{T}, \mathfrak{D})$. Die Statistik T heißt vollständig für $\vartheta \in \Theta$, wenn für jede $\mathfrak{D}$-meßbare Funktion $s(t)$ mit $\mathsf{E}_\vartheta s(T(Y)) = 0 \quad \forall \vartheta \in \Theta$ die Beziehung*

$$s(t) = 0 \qquad [\mathfrak{D}, \mathcal{P}^T]$$

gilt.

Anschauliche Interpretation: Es sei p_ϑ Dichte von P_ϑ bzgl. eines dominierenden σ-endlichen Maßes μ. Dann gilt

$$\mathsf{E}_\vartheta r(Y) = \int r(y) p_\vartheta(y) \mu(dy) \qquad \forall \vartheta.$$

Wenn daraus folgen soll, daß $r(y) = 0$ ist, so muß p_ϑ für denjenigen Raum, in dem die Funktion r liegen soll, ein vollständiges Funktionensystem sein. Deshalb ist der Begriff Vollständigkeit naheliegend. Die Vollständigkeit ermöglicht Eindeutigkeitsaussagen über erwartungstreue Schätzfunktionen, die nur von der vollständigen Statistik abhängen. Sind nämlich $h(T(y)), h_1(T(y))$ eine erwartungstreue Schätzfunktion für einen Parameter $\gamma(\vartheta)$, so folgt aus der $\mathfrak{D}-$Meßbarkeit von $h(t), h_1(t)$ mit $s(t) = h(t) - h_1(t)$ und $\mathsf{E}_\vartheta s(T(Y)) = 0 \quad \forall \vartheta$ die Beziehung

$$h(t) = h_1(t) \qquad \text{fast überall} \qquad (*)$$

Satz 8.6 (Lehmann-Scheffé) *Sei T eine vollständige und suffiziente Statistik für $\vartheta \in \Theta$. Dann gibt es zu jeder erwartungstreuen Schätzfunktion g eines eindimensionalen Parameters $\gamma(\vartheta)$ eine erwartungstreue Schätzfunktion der Form $h(T(\cdot))$ mit $\mathfrak{D}$-meßbarem h, die unter allen erwartungstreuen Schätzfunktionen eine gleichmäßig kleinste Varianz hat. $h(T(y))$ ist bis auf Nullmengen aus $\mathfrak{F} = T^{-1}(\mathfrak{D})$ bzgl. $\mathcal{P}$ bestimmt.*

BEWEIS: Es gilt $\mathsf{E}_\vartheta g(Y) = \gamma(\vartheta) \quad \forall \vartheta, h(T(y)) = \mathsf{E}.(g(Y) \mid T(y))$. Aus Satz (8.1) folgt: $h(T)$ ist erwartungstreu.
Annahme: Es existieren eine erwartungstreue Schätzfunktion $g_1(y)$ und ein $\vartheta_1 \in \Theta$ mit

$$\mathsf{Var}_{\vartheta_1} g_1(Y) < \mathsf{Var}_{\vartheta_1} h(T(Y)) .$$

So folgt aus dem Satz von Rao-Blackwell, daß $h_1(T) = \mathsf{E}.(g_1(Y) \mid T(y))$ erwartungstreu ist und es gilt $\mathsf{Var}_{\vartheta_1} h_1(T(Y)) \leq \mathsf{Var}_{\vartheta_1} g_1(Y) < \mathsf{Var}_{\vartheta_1} h(T(Y))$. Andererseits sind $h(T(\cdot)), h_1(T(\cdot))$ erwartungstreu. Aus $(*)$ folgt $\mathsf{Var}_{\vartheta_1} h_1(T(Y)) = \mathsf{Var}_{\vartheta_1} h(T(Y)) \Rightarrow$ Widerspruch.

□

Beispiel 8.7 $Y_1, ..., Y_n$ i.i.d., $Y_i \sim \mathsf{B}(1, \vartheta), \vartheta \in (0, 1)$. Dann ist T mit $T(y) = \sum_{i=1}^{n} y_i$ vollständig, denn aus

$$\mathsf{E}_\vartheta s(T(Y)) = \sum_{j=0}^{n} s(j) \binom{n}{j} \vartheta^j (1-\vartheta)^{n-j} = 0 \quad \forall \vartheta$$

folgt $s(j) = 0, j = 0, ..., n$. Da k mit $k(y) = \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ erwartungstreu für ϑ ist, k von y nur über die suffiziente Statistik $T(y)$ abhängt, folgt aus dem Satz von Lehmann-Scheffé, daß $k(y) = \overline{y}$ eine erwartungstreue Schätzfunktion mit minimaler Varianz ist.

Es gilt noch ein wichtiger Satz über Exponentialfamilien. Es sind nämlich in wichtigen Exponentialfamilien die Statistiken $T(y)$ vollständig.

Satz 8.7 *Es sei* $\mathcal{P} = \{\mathsf{P}_\vartheta : \vartheta \in \Theta\}$ *eine k-parametrige Exponentialfamilie in* $T(y)$ *und* $\zeta(\vartheta)$. *Dann ist* T *eine vollständige Statistik für* Θ, *falls* $\{\zeta : \zeta = \zeta(\vartheta), \vartheta \in \Theta\}$ *ein nichtentartetes k-dimensionales Intervall enthält.*

Beispiel 8.8 $Y_1, ..., Y_n$ i.i.d., $Y_i \sim \mathsf{N}(a, \sigma^2), (a, \sigma^2) \in \mathbb{R}^1 \times \mathbb{R}^+$. Die gemeinsame Dichte ist

$$\begin{aligned} p_\vartheta(y) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - a)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \Sigma y_i^2 + \frac{1}{\sigma^2} a \Sigma y_i - \frac{n}{2\sigma^2} a^2\right) \\ \Rightarrow T(y) &= (\Sigma y_i^2, n\overline{y}) \quad \text{suffizient} . \end{aligned}$$

Außerdem ist $\mathcal{P}$ eine 2-parametrige Exponentialfamilie in $T(y)$ und $\zeta(\vartheta) = (-\frac{1}{2\sigma^2}, \frac{na}{\sigma^2})$. Die Komponenten von $\zeta(\vartheta)$ durchlaufen unabhängig voneinander nichtentartete Intervalle. $\Rightarrow$ $T(y)$ ist vollständig und folglich ist $k(y) = \overline{y}$ eine erwartungstreue Schätzfunktion für a mit minimaler Varianz,

$$\tilde{k}(y) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2 = \frac{1}{n-1} (\Sigma y_i^2 - n\overline{y}^2)$$

ist eine erwartungstreue Schätzfunktion für σ^2 mit minimaler Varianz.

9 Regression und Varianzanalyse

Die Untersuchung der Wirkung einiger Einflußgrößen auf eine oder mehrere gemessene Zielgrößen gehört zu den zentralen Aufgaben der angewandten Statistik. Von Interesse sind Mittelwertvergleiche, Schätzung der Mittelwerte und Varianzen. Als statistischen Modellansatz für derartige Probleme wählt man häufig lineare Regressionsmodelle.

Definition 9.1 *Es sei $Y \in \mathbb{R}^n$, $Y \sim \mathsf{P}_\vartheta$ mit $Y - \mathsf{E}_\vartheta Y = \varepsilon$ die Verteilung Q_ϑ von ε und $\mathsf{E}_\vartheta Y$ seien als Funktionen von ϑ unabhängig voneinander. Dann heißt $Y = \mu + \varepsilon$ bei $\mu = \mathsf{E}_\vartheta Y$ Regressionsmodell für Y. Gilt $\mu = X\beta$ für eine $n \times k$-Matrix X, so heißt $Y = X\beta + \epsilon$ lineares Regressionsmodell. Gilt $\mu = X(\beta)$ für $\beta \in \mathbb{R}^k$ und eine nichtlineare Funktion X, so heißt $Y = X(\beta) + \varepsilon$ nichtlineares Regressionsmodell.*

9.1 Das lineare Regressionsmodell

Es sei eine Variable $\tilde{Y}$ gegeben, die mit k Einflußgrößen $x_1, ..., x_k$ in folgender Weise verbunden sein soll:

$$\tilde{Y} = x_1\beta_1 + ... + x_k\beta_k + \tilde{\varepsilon}.$$

Die Regressionskoeffizienten $\beta_1, ..., \beta_k$ sind unbekannt, die zufällige Variable $\tilde{\varepsilon}$ sei ein Fehler mit $\mathsf{E}\tilde{\varepsilon} = 0$. Über die unbekannte Varianz σ^2 von $\tilde{\varepsilon}$ und die Parameter $\beta_1, ..., \beta_k$ sollen statistische Aussagen abgeleitet werden.

Liegen nun für n gegebene k-Tupel der Einstellgrößen $x_1, ..., x_k$ Beobachtungen von $\tilde{Y}$ vor, so haben wir

$$Y_t = x_{t1}\beta_1 + ... + x_{tk}\beta_k + \varepsilon_t, \quad t = 1, ..., n$$

oder in Matrixschreibweise

$$Y = X\beta + \varepsilon.$$

Y ist der Beobachtungsvektor, X die bekannte $n \times k$-Matrix der Einstellgrößen (Versuchsplanmatrix). Es sei $\mathsf{E}\varepsilon = 0, \mathsf{Var}\ \varepsilon = \sigma^2 W$. W sei positiv definit. Auf Gauß (1809) geht die Methode zurück, als Schätzung $\tilde{\beta} = (\tilde{\beta}_1, ..., \tilde{\beta}_k)'$ für β diejenigen Werte zu nehmen, für welche die Summe der Abweichungsquadrate

$$\sum_{t=1}^{n} (y_t - X_{t1}\beta_1 - ... - X_{tk}\beta_k)^2 = \mid y - X\beta \mid^2$$

minimal wird. Es wird auch häufig eine gewichtete Summe der Abweichungsquadrate der Form $\mid y - X\beta \mid_Z^2$ minimiert. Es sei $Z \in \mathcal{M}_n^>$. Lösen wir diesen Fall, so differenziert

man nach β und setzt diesen Ausdruck $= 0$. Man erhält als zu lösendes Gleichungssystem die Normalgleichung (NGL)

$$X'ZX\hat{\beta} = X'ZY.$$

Wenn X den Rang k hat, so ist auch $X'ZX$ nichtsingulär und folglich

$$\hat{\beta} = (X'ZX)^{-1}X'ZY$$

die einzige Lösung. Hat X keinen vollen Rang, so existiert keine eindeutige Lösung der NGL, aber jedes

$$\hat{\beta} = (X'ZX)^{-}X'ZY$$

für eine beliebige verallgemeinerte Inverse $(X'ZX)^{-}$ ist Lösung der NGL und minimiert $\mid Y - X\beta \mid_Z^2$.

Definition 9.2 *Eine Parameterfunktion $F\beta$ für eine $m \times k$-Matrix F heißt schätzbar, falls eine $n \times m$-Matrix L derart existiert, daß $\mathsf{E}L'Y = F\beta$ für alle β gilt.*

Damit ist $F\beta$ schätzbar gdw. ein L derart existiert, daß $L'X = F$ gilt. Man erkennt, daß ein solches L immer existiert, falls X den vollen Rang k hat. Dann sind alle $F\beta$ schätzbar. Offenbar ist $X\beta$ immer schätzbar und man erkennt auch, daß die schätzbaren Parameter genau die $LX\beta$ für eine beliebige Matrix L sind. Der Hintergrund der Schätzbarkeit ist: Wenn zwei Parameter β, $\tilde{\beta}$ mit $X\beta = X\tilde{\beta}$ existieren, so muß auch $F\beta = F\tilde{\beta}$ gelten. Andernfalls könnte der Unterschied zwischen $F\beta$ und $F\tilde{\beta}$ aus den identischen Variablen $X\beta + \varepsilon$ und $X\tilde{\beta} + \varepsilon$ nicht erklärt werden.

Gesucht wird häufig eine beste erwartungstreue Schätzung für $F\beta$. Dabei heißt γ^* bester linearer erwartungstreuer Schätzer für $F\beta$, falls $\gamma^* = \gamma^*(Y)$ linear von Y abhängt,

$\mathsf{E}\gamma^* = F\beta$ für alle β gilt und
$\mathsf{Var}\ \gamma^* \leq \mathsf{Var}\ L'Y$ für alle erwartungstreuen Schätzungen $L'Y$ erfüllt ist.

Bezeichnung: β^* ist BLUE (best linear unbiased estimator)

Satz 9.1 (Gauß–Markov) *Gegeben sei eine schätzbare Funktion $F\beta$. Dann ist $F\hat{\beta} = F(X'W^{-1}X)^{-}X'W^{-1}Y$ für eine beliebige g-Inverse ein bester linearer erwartungstreuer Schätzer (genauer Schätzvariable) für $F\beta$.*

Beweis: Zunächst ist $F\hat{\beta}$ linear in Y. Es gilt $\mathsf{E}F\hat{\beta} = F(X'W^{-1}X)^{-}X'W^{-1}X\beta$. Da $F\beta$ schätzbar ist, existiert ein L derart, daß $L'X = F$. Daher haben wir $F(X'W^{-1}X)^{-}X'W^{-1}X = L'X = F$ und somit $\mathsf{E}F\hat{\beta} = F\beta$ für alle β. Folglich ist $F\hat{\beta}$ erwartungstreu. Es bleibt nun noch z.z., daß für eine beliebige erwartungstreue Schätzung $L'Y$ die Ungleichung

$$\mathsf{Var}\ F\hat{\beta} \leq \mathsf{Var}\ L'Y$$

erfüllt ist. Wir erhalten zunächst $\mathsf{Var}\, F\hat{\beta} = \sigma^2 F(X'W^{-1}X)^- X'W^{-1}X(X'W^{-1}X)^- F = \sigma^2 F(X'W^{-1}X)^+ F'$. Da $L'Y$ erwartungstreu ist, folgt $L'X = F$ und

$$\mathsf{Var}\, L'Y = \sigma^2 L'WL.$$

Schreiben wir $L' = F(X'W^{-1}X)^- X'W^{-1} + [L' - F(X'W^{-1}X)^- X'W^{-1}]$, so folgt

$$L'WL = F(X'W^{-1}X)^+ F' + [L' - F(X'W^{-1}X)^- X'W^{-1}]W[L - W^{-1}X(X'W^{-1}X)^{-'}F']$$
$$\text{wegen} \qquad F(X'W^{-1}X)^- X'[L - W^{-1}X(X'W^{-1}X)^{-'}F'] = 0.$$

Das bedeutet aber

$$L'WL \geq F(X'W^{-1}X)^+ F', \qquad (*)$$

womit die Behauptung bewiesen ist. Außerdem erkennt man, daß in (*) nur dann die Gleichheit gilt, falls

$$L' = F(X'W^{-1}X)^{-'} XW^{-1}$$

und somit $L'Y = F\hat{\beta}$ gilt.

□

Folgerung: Die beste lineare Schätzung $X\hat{\beta}$ für $X\beta$ ist eine (nichtorthogonale) Projektion auf $\mathcal{R}(X)$.
Beweis: Es gilt $X\hat{\beta} = X(X'W^{-1}X)^- X'W^{-1}Y =: T_0Y$ Es gilt $T_0T_0 = T_0$. $\Rightarrow$ T_0 beschreibt Projektion parallel zu $\mathcal{R}(A)$, wenn $X'W^{-1}A = 0$ mit maximalem Rang für A gilt.

Wenn speziell $W = I$ ist, so ist $X\hat{\beta} = P_{\mathcal{R}(X)}Y$ für die orthogonale Projektion $P_{\mathcal{R}(X)}$.

9.1.1 Schätzung unter Restriktionen

Bisher haben wir angenommen, daß alle Komponenten von β beliebige Werte annehmen können. Oftmals sind aber die Komponenten funktionell verbunden. Wir nehmen an, daß diese Zusammenhänge durch ein lineares Gleichungssystem $R\beta = r$ gegeben sind. Hier sind die $v \times k-$Matrix R und $r \in \mathbb{R}^v$ bekannt. Wir nehmen an, daß eine Lösung von $R\beta = r$ existiert. Bezeichne D eine $k \times v$-Matrix, die den Nullraum von R aufspannt, und sei β_0 eine spezielle Lösung von $R\beta = r$, so lassen sich alle Lösungen in der Form

$$\beta = \beta_0 + D\alpha \quad , \quad \alpha \in \mathbb{R}^v$$

darstellen.

Definition 9.3 *Eine Parameterfunktion $F\beta$ für $F: m \times k$ heißt im Regressionsmodell unter den Restriktionen schätzbar, wenn eine $n \times m$-Matrix L und ein $\tilde{l} \in \mathbb{R}^m$ derart existieren, daß $\mathsf{E}(L'Y + \tilde{l}) = F\beta$ für alle β mit $R\beta = r$ gilt.*

Man zeigt unmittelbar, daß $F\beta$ unter den Restriktionen genau dann schätzbar ist, falls ein L mit

$$L'XD = FD$$

existiert. Daraus ergibt sich, daß die Schätzbarkeit nicht von der Wahl einer speziellen Matrix D abhängt. Diese Schätzbarkeit bedeutet, daß für zwei β, $\tilde{\beta} \in \mathbb{R}^k$ mit $R\beta = R\tilde{\beta}$ und $X\beta = X\tilde{\beta}$ stets $F\beta = F\tilde{\beta}$ gilt.

Definition 9.4 *$F\hat{\beta}_R$ heißt beste lineare Schätzung unter den Nebenbedingungen $R\beta = r$, falls $\hat{\beta}_R$ linear von Y abhängt,*

$$\mathsf{E}_\vartheta F\hat{\beta}_R(Y) = F\beta \qquad \forall \beta \text{ mit } R\beta = r,$$
$$\mathsf{Var}_\vartheta(F\hat{\beta}_R(Y)) \leq \mathsf{Var}_\vartheta(TY) \qquad \forall T \text{ mit } \mathsf{E}_\vartheta T(Y) = F\beta \text{ bei } R\beta = r \text{ gilt.}$$

Die Existenz einer derartigen Schätzung gibt der folgende Satz.

Satz 9.2 *Es sei $F\beta$ unter den Restriktionen schätzbar. Dann ist*

$$\hat{\beta}_R = \beta_0 + D(D'X'W^{-1}XD)^- D'X'W^{-1}(Y - X\beta_0)$$

für eine beliebige g-Inverse und beliebige D, β_0, die die geforderten Bedingungen erfüllen, eine erwartungstreue Schätzung unter $R\beta = r$. Außerdem ist $\hat{\beta}_R$ eine Schätzung, die unter allen linearen, unter $R\beta = r$ erwartungstreuen Schätzungen eine gleichmäßig kleinste Kovarianz haben. Es gilt

$$\mathsf{Var}\, F\hat{\beta}_R = \sigma^2 FD(D'X'W^{-1}XD)^- D'F'.$$

Es gilt $\mathsf{Var}\, F\hat{\beta}_R = \sigma^2 L'W^{1/2}PW^{1/2}L$ für die Projektionsmatrix $P = P_{\mathcal{R}(W^{-1/2}XD)}$.

Spezialfälle

1. X habe vollen Rang. Dann existiert für jedes F ein L, so daß die Schätzbarkeitsbedingung $L'XD = FD$ erfüllt ist. Dann gilt die Darstellung

$$\hat{\beta}_R = \hat{\beta} - (X'W^{-1}X)^{-1}R'(R(X'W^{-1}X)^{-1}R')^-[R\hat{\beta} - r].$$

Die Kovarianzmatrix hat dann die Gestalt

$$\begin{aligned}\mathsf{Var}\,\hat{\beta}_R \;=\; & \sigma^2(X'W^{-1}X)^{-1/2} \\ & [I - (X'W^{-1}X)^{-1/2}R'(R(X'W^{-1}X)^{-1}R')^- R(X'W^{-1}X)^{-1/2}](X'W^{-1}X)^{-1/2}.\end{aligned}$$

Für $R \neq 0$ gilt offenbar $\mathsf{Var}\,\hat{\beta}_R \leq \mathsf{Var}\,\hat{\beta}$.

2. X habe keinen vollen Rang. Es gelte $rgX = k - s, rgR = s$ und $rg(X' : R') = k$. Damit betreffen die Restriktionen genau den Raum der nichtschätzbaren linearen Funktionen. In diesem Fall heißen die Gleichungen $R\beta = r$ auch Reparametrisierungsbedingungen. Wir wählen

$$D = (X'W^{-1}X + R'R)^{-1}X'W^{-1}X.$$

Dann gilt $RD = 0$ und es ist $L'XD = FD$ für beliebiges F für ein geeignetes L erfüllt. Damit ist β unter $R\beta = r$ schätzbar. Es gilt

$$\hat{\beta}_R = (X'W^{-1}X + R'R)^{-1}X'W^{-1}(Y - X\beta_0) + \beta_0 \text{ bei } R\beta_0 = r.$$

9.1.2 Schätzen der Varianz

Sei $\hat{\beta} = \hat{\beta}(Y)$. Der Vektor der Residuen ist

$$\hat{\varepsilon} = W^{-1/2}(Y - X\hat{\beta})$$

und ist unabhängig von Mehrdeutigkeiten der BLUE $\hat{\beta}$ eindeutig bestimmt. Schreiben wir

$$\begin{aligned} \hat{\varepsilon} &= W^{-1/2}Y - W^{-1/2}X(X'W^{-1}X)^{-}X'W^{-1}Y \\ &= \left(I - P_{\mathcal{R}(W^{-1/2}X)}\right) W^{-1/2}Y, \end{aligned}$$

so ergibt sich

$$\hat{\varepsilon}'\hat{\varepsilon} = Y'W^{-1/2}\left(I - P_{\mathcal{R}(W^{-1/2}X)}\right) W^{-1/2}Y.$$

Daraus errechnen wir mit $q = rgX$

$$\mathsf{E}\hat{\varepsilon}'\hat{\varepsilon} = \sigma^2 sp\left(I - P_{\mathcal{R}(W^{-1/2}X)}\right) = (n-q)\sigma^2.$$

Wir erhalten das

Ergebnis: Die Schätzung $\hat{\sigma}^2 = \hat{\sigma}^2(Y)$ mit

$$\hat{\sigma}^2 = \frac{1}{n-q}(Y - X\hat{\beta})'W^{-1}(Y - X\hat{\beta})$$

ist erwartungstreu für σ^2.

Liegen nun noch Restriktionen wie im vorigen Abschnitt vor, so ist analog wie oben

$$\hat{\sigma}_R^2 = \frac{1}{n-q_R}(Y - X\hat{\beta}_R)'W^{-1}(Y - X\hat{\beta}_R)$$

für $q_R = rgXD$ erwartungstreu für σ^2.

Beispiel 9.1 $Y_1, \ldots, Y_n$ seien i.i.d. mit $\mathsf{E}Y_i = \beta$, $\mathsf{Var}\ Y_i = \sigma^2$. Dann ist $\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{Y})^2$ erwartungstreu für σ^2.

BEWEIS: Es ist $Y = \beta \cdot \mathbf{1} + \varepsilon$ mit $\mathsf{Var}\ \varepsilon = \sigma^2 I$. Folglich gilt

$$\begin{aligned} \hat{\beta} &= \hat{Y}, \\ \hat{\varepsilon} &= Y - \hat{Y}\mathbf{1} \text{ und} \\ \mid \hat{\varepsilon} \mid^2 &= \sum_{i=1}^{n} (Y_i - \hat{Y})^2. \end{aligned}$$

Wir haben $q = 1$ und folglich ist

$$\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \hat{Y})^2 \qquad \text{erwartungstreu für } \sigma^2.$$

□

Allgemeine Aussagen der Optimalität der Schätzungen $\hat{\beta}$, $\hat{\sigma}^2$ liefert das folgende Beispiel.

Beispiel 9.2 Optimalität von $\hat{\beta}$ und $\hat{\sigma}^2$.
Es liege das normale lineare Regressionsmodell

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathsf{N}(0, \sigma^2 W), \qquad \beta \in \mathbb{R}^k, \qquad \sigma^2 \in \mathbb{R}^+, \qquad rgX = k$$

vor. Dann ist die Dichte bei $\vartheta = (\beta, \sigma^2)$

$$\begin{aligned} p_\vartheta(y) &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}} \mid W \mid^{\frac{1}{2}}} exp\left(-\frac{1}{2\sigma^2}(y - X\beta)' W^{-1} (y - X\beta)\right) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}} \mid W \mid^{\frac{1}{2}}} exp\left(-\frac{s^2}{2\sigma^2} - \frac{1}{2\sigma^2}(\hat{\beta} - \beta)' X' W^{-1} X (\hat{\beta} - \beta)\right) = * \end{aligned}$$

Hier nutzen wir die Bezeichnung $s^2 = s^2(y)$, $\hat{\sigma}^2 = \hat{\sigma}^2(y)$.
Wir haben

$$\begin{aligned} s^2 &= (y - X\hat{\beta})' W^{-1} (y - X\hat{\beta}) \\ &= y' W^{-\frac{1}{2}} P_{\mathcal{N}(X'W^{-\frac{1}{2}})} W^{-\frac{1}{2}} y \\ \hat{\beta}' X' W^{-1} X \hat{\beta} &= y' W^{-\frac{1}{2}} P_{\mathcal{R}(W^{-\frac{1}{2}}X)} W^{-\frac{1}{2}} y, \text{ folglich} \\ s^2 + \hat{\beta}' X' W^{-1} X \hat{\beta} &= y' W^{-1} y. \qquad \Rightarrow \end{aligned}$$

$$* = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}} \mid W \mid^{\frac{1}{2}}} exp\left(-\frac{y' W^{-1} y}{2\sigma^2} + \frac{1}{\sigma^2} \beta' X' W^{-1} X \hat{\beta}\right) e^{-\frac{1}{2\sigma^2}\beta' X' W^{-1} X \beta}.$$

$\Rightarrow$ $\mathcal{P}$ ist $k+1$-parametrige Exponentialfamilie in

$$T(y) = \left(y' W^{-1} y, \hat{\beta}'\right)' \text{ und } \zeta(\vartheta) = \left(-\frac{1}{2\sigma^2}, \frac{1}{\sigma^2}\beta' X' W^{-1} X\right)'$$

unter $rgX = k$. Da die Komponenten $\zeta(\vartheta)$ unabhngig voneinander variieren können, ist $T(y)$ vollständig (und suffizient). Zu Beginn dieses Abschnitts wurde gezeigt, daß $\hat{\sigma}^2 = \frac{1}{n-k}s^2$ erwartungstreu für σ^2 ist. Außerdem ist $\hat{\sigma}^2$ von y nur über $T(y)$ abhängig, denn wir haben

$$\hat{\sigma}^2 = \frac{1}{n-k}\left(Y'W^{-1}Y - \hat{\beta}'X'W^{-1}X\hat{\beta}\right).$$

$\Rightarrow$ $\hat{\sigma}^2$ ist eine erwartungstreue SF für σ^2 mit minimaler Varianz. Sei nun $F\beta$ schätzbar (das ist bei $rgX = k$ keine Bedingung), $a \in \mathbb{R}^m$ sei ein fester Vektor, dann folgt, daß $a'F\beta$ schätzbar ist und der eindimensionale Parameter $a'F\beta$ ist erwartungstreu für $a'F\beta$. Außerdem hängt $a'F\hat{\beta}$ nur über die suffiziente und vollständige Statistik $T(y)$ von y ab, so daß für alle β, σ^2

$$\mathsf{Var}\left(a'F\hat{\beta}\right) \leq \mathsf{Var}\left(f(Y)\right) \qquad (*)$$

für eine beliebige für $a'F\beta$ erwartungstreue SF $f(Y)$ gilt. Wir haben

$$\mathsf{Var}\left(a'F\hat{\beta}\right) = a'\mathsf{Var}\left(F\hat{\beta}\right)a \qquad (**)$$

Wir wollten zeigen, daß $F\hat{\beta}$ die beste erwartungstreue SF für $F\beta$ ist. Nehmen wir an, das stimmt nicht, es existiert also ein $G(Y)$ und ein a derart, daß

$$\begin{aligned} \mathsf{E}_\vartheta G(Y) &= F\beta \qquad \forall\vartheta \text{ und für ein } \vartheta_0, \sigma_0 \\ \mathsf{Var}_{\beta_0,\sigma_0^2}\left(a'G(Y)\right) &< \mathsf{Var}_{\beta_0,\sigma_0^2}\left(a'F\hat{\beta}\right) \end{aligned}$$

Da aber $a'G(Y)$ erwartungstreu für $a'F\beta$ ist, ist das ein Widerspruch zu $(*)$. Damit ist wegen $(**)$ für alle erwartungstreuen SF für $F\beta$ die Matrix $\mathsf{Var}\left(F\hat{\beta}\right)$ minimal.

9.1.3 Testen von Hypothesen

Wir betrachten das normale lineare Regressionsmodell

$$Y = X\beta + \varepsilon \qquad , \qquad \varepsilon \sim \mathsf{N}(0, \sigma^2 W).$$

Zu prüfen sei die lineare Hypothese
$H : G\beta = g$ gegen
$K : G\beta \neq g$.
G ist eine $m \times k$-Matrix und $G\beta = g$ sei lösbar.

Definition 9.5 *Eine Hypothese $G\beta = g$ heißt testbar im linearen Regressionsmodell, falls $G\beta$ schätzbar ist.*

Wir setzen im folgenden voraus, daß $G\beta = g$ testbar ist. Heuristisch wesentlich ist dann die Differenz $G\hat{\beta} - g$, die zufällig ist. Es muß eine Maßzahl dafür gefunden werden, wann $G\hat{\beta} - g$ groß ist. Es gilt

$$G\hat{\beta} \sim \mathsf{N}\left(G\beta, \sigma^2 GKG'\right) \qquad \text{für } K = (X'W^{-1}X)^+$$

Dann ist

$$(GKG')^{\frac{1}{2}+}GKG'(GKG')^{\frac{1}{2}+} =: V$$

eine orthogonale Projektion und es gilt

$$V = \frac{1}{\sigma^2}\mathsf{Var}_{\beta,\sigma^2}\left(\underbrace{(GKG')^{\frac{1}{2}+}G\hat{\beta}}_{=:Z}\right) \qquad \text{für} \quad \hat{\beta} = \hat{\beta}(Y).$$

Folglich gilt $Z \sim \mathsf{N}((GKG')^{\frac{1}{2}+}G\beta, \sigma^2 V)$. Damit ist $\mid Z \mid^2 = \sum_{i=1}^{n} Z_i^2$ eine Summe von Quadraten normalverteilter Variable mit der Kovarianzmatrix V. Ist $\varkappa = rgV$, so haben wir

$$\mid Z - \mathsf{E}_\vartheta Z \mid^2 \sim \sigma^2\chi^2_\varkappa.$$

Nach Definition hat $\mid Z \mid^2$ eine nichtzentrale χ^2-Verteilung:

$$\mid Z \mid^2 \sim \sigma^2\chi^2_{\varkappa,\psi^2} \qquad \text{für } \psi^2 = \frac{1}{\sigma^2}\beta'G'(GKG')^+G\beta.$$

Da wir an $G\hat{\beta} - g$ interessiert sind, bilden wir

$$M := (GKG')^{\frac{1}{2}+}(G\hat{\beta} - g) \sim \mathsf{N}((GKG')^{\frac{1}{2}+}(G\beta - g), \sigma^2 V),$$

um eine zufällige Variable mit einer idempotenten Kovarianzmatrix zu erhalten. Folglich ist

$$\mid M \mid^2 \sim \sigma^2\chi^2_{\varkappa,\psi_1^2} \qquad \text{für } \psi_1^2 = \frac{1}{\sigma^2}(G\beta - g)'(GKG')^+(G\beta - g).$$

Unter H ist $\psi_1^2 = 0$ und dann ist $\mid M \mid^2$ zentral χ^2-verteilt. Außerdem sind M und $\hat{\sigma}^2$ stochastisch unabhängig, denn $\hat{\beta}$ und $\hat{\sigma}^2$ sind unabhängig.

Getestet werden soll $G\beta = g$. Wir haben unter H
$\mid M \mid^2 = (G\hat{\beta} - r)'(GKG')^+(G\hat{\beta} - r) \sim \sigma^2\chi^2_\varkappa$ für $\varkappa = rg(K^{\frac{1}{2}}G') = rgG$
Unter K gilt

$$\mid M \mid^2 \sim \sigma^2\chi^2_{\varkappa,\psi_1^2}.$$

Außerdem haben wir $s^2 \sim \sigma^2\chi^2_{n-q}$ für $s^2 = (Y - X\hat{\beta})'W^{-1}(Y - X\hat{\beta})$. Also ist die Verteilung von $\frac{|M|^2}{s^2}$ unter H bekannt, es ist eine F-Verteilung:

$$F(Y) := \frac{n-q}{\varkappa} \cdot \frac{\mid M \mid^2}{s^2} = \frac{n-q}{\varkappa} \cdot \frac{(G\hat{\beta} - r)'(GKG')^+(G\hat{\beta} - r)}{Y'W^{-\frac{1}{2}}(I - P_{\mathcal{R}(W^{-\frac{1}{2}}X)})W^{-\frac{1}{2}}Y} \sim F_{\varkappa,n-q}$$

Ergebnis: (F-Test der linearen Hypothese)
Ein Test mit der Irrtumswahrscheinlichkeit 1.Art α für $H: G\beta = g$ gegen $K: G\beta \neq g$ ist gegeben durch

$$\varphi(y) = \begin{cases} 1 & \text{für} \quad F(y) > F_{\varkappa,n-q;\alpha} \\ 0 & \text{sonst} \end{cases}$$

Dieser Test ist ein bester α-Test für H gegen K (in einer Klasse invarianter Tests).

D.h. der kritische Bereich S ist

$$S = \{y : F > F_{\varkappa, n-q;\alpha}\}$$

Falls $y \in S$, so wird H abgelehnt.

$$\begin{aligned} P(K \text{ angenommen} \mid H \text{ richtig}) &= \alpha \\ P(H \text{ angenommen} \mid H \text{ richtig}) &= 1-\alpha \end{aligned}$$

$\Rightarrow$ Eine irrtümliche Annahme von K tritt nur in $\alpha \cdot 100\%$ der möglichen Fälle auf.

Wir betrachten nun den Fall zusätzlicher Restriktionen $R\beta = r$. Da es nicht sinnvoll ist, gegebene Restriktionen zu testen, setzen wir voraus, daß $\mathcal{R}(G)$ und $\mathcal{R}(R)$ disjunkt sind, d.h.

$$rg(R' : G') = rgR + rgG.$$

Ist dann $G\beta$ unter der Restriktion $R\beta = r$ schätzbar, d.h. auch testbar, so setzen wir

$$\begin{aligned} s_H^2 &= (G\hat{\beta}_R - g)'[GD(D'X'W^{-1}XD)^- D'G'](G\hat{\beta}_R - g) \qquad \text{und} \\ s_R^2 &= (Y - X\hat{\beta}_R)'W^{-1}(Y - X\hat{\beta}_R). \end{aligned}$$

Analog wie oben sehen wir, daß

$$F(Y) = \frac{n - q_R}{\varkappa} \cdot \frac{s_H^2}{s_R^2} \qquad \text{für} \quad s_H^2 = s_H^2(Y), s_R^2 = s_R^2(Y),$$

eine F-Verteilung mit $\varkappa = rgGD$ und $q_R = rgXD$ besitzt.

Beispiel 9.3 Vergleich von Apfelgewichten von zwei Dörfern.
Gewichte der Äpfel :

Plantage 1 :	156	146	147	152	147	155
Plantage 2 :	161	152	152	158	151	159

Wir nutzen die Modellansätze

$$\begin{aligned} Y_i &= \rho + \varepsilon_i \qquad i = 1, \ldots, 6 \\ Y_i &= \gamma + \varepsilon_i \qquad i = 7, \ldots, 12 \end{aligned} \quad (*)$$

$$\begin{aligned} \Rightarrow \hat{\rho} &= 150,5 \\ \hat{\gamma} &= 155,5 \end{aligned}$$

Beim Modellansatz $Y_i = \mu + \varepsilon_i \qquad i = 1, \ldots, 12$, bei dem also ein einheitliches Modell für alle Gewichte vorausgesetzt wird, erhalten wir das mittlere Gewicht:

$$\begin{aligned} \frac{1}{12}\sum_{i=1}^{12} y_i &= \frac{1}{12}(6 \cdot 150 + 6 \cdot 155 + 3 + 3) \\ &= 153 \end{aligned}$$

Existiert ein Unterschied im Gewicht der Äpfel der beiden Dörfer? Dann würde man $(*)$ ansetzen und fragen: Gilt $\rho = \gamma$? Als Modell hat man

$$Y = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \rho \\ \gamma \end{pmatrix} + \varepsilon$$

$\rho = \gamma$ bedeutet

$$\begin{aligned} (1 \vdots -1) \begin{pmatrix} \rho \\ \gamma \end{pmatrix} &= 0 \\ G\beta &= 0 \end{aligned}$$

Also lautet die Hypothese $G\beta = 0$, die Alternative $G\beta \neq 0$.
Unter der Hypothese muß β geschätzt werden, anderenfalls auch.
$\Rightarrow$ Schätzung mit und ohne Restriktionen. Anschließend muß ein Vergleich der Schätzungen durchgeführt werden zur Beantwortung der Frage. Hier ist $G\hat{\beta} = -5$. Ist das groß oder klein? Getestet wird

$$H: \ G\beta = 0 \quad \text{gegen} \quad K: \ G\beta \neq 0$$

für $G = (1, -1)$; $\quad n = 12$; $\quad q = 2$. Es ist

$$\beta = \begin{pmatrix} \rho \\ \gamma \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}, \qquad W = I \Rightarrow \quad K = (X'X)^{-1} = \begin{pmatrix} 6 & 0 \\ 0 & 6 \end{pmatrix}^{-1} = \tfrac{1}{6} I$$

$$GKG' = \tfrac{1}{6}(1, -1) \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \tfrac{1}{3}$$

$$V = 1 \quad \Rightarrow \quad \varkappa = 1$$

$$\hat{\beta}(y) = \begin{pmatrix} 150{,}5 \\ 155{,}5 \end{pmatrix} \quad \Rightarrow \quad G\hat{\beta}(y) = -5 \qquad X\hat{\beta}(y) = \begin{pmatrix} 150{,}5 \\ \vdots \\ 150{,}5 \\ 155{,}5 \\ \vdots \\ 155{,}5 \end{pmatrix}$$

$$\begin{aligned} s^2(y) &= y'y - \hat{\beta}' X' X \hat{\beta} \\ &= 281174 - 280983 = 191 \end{aligned}$$

$| M |^2 = (-5) \cdot 3 \cdot (-5) = 75 \qquad \Rightarrow \quad F(y) = \frac{12-2}{1} \cdot \frac{75}{191} = \frac{750}{191} = 3{,}93$
Aus der Tabelle ergibt sich

$$\begin{aligned} F_{1,10;0,05} &= 4{,}96 \\ &\Rightarrow \ H \text{ wird nicht abgelehnt.} \end{aligned}$$

9.1.4 Konfidenzbereiche

Nehmen wir an, daß Y einem normalen linearen Modell genügt und sei $F\beta$ schätzbar. Dann haben wir gezeigt, daß mit Wahrscheinlichkeit $1-\alpha$ die Ungleichung

$$\frac{1}{\varkappa}\cdot\frac{1}{\hat{\sigma}^2}(F\hat{\beta}-F\beta)'(F(X'W^{-1}X)^{-}F')^{-}(F\hat{\beta}-F\beta)<F_{\varkappa,n-q;\alpha}$$

gilt, wobei $\varkappa=rgF, \hat{\sigma}^2=s^2/(n-q)$ gilt. Definieren wir eine Menge $\mathcal{K}(\beta)$ durch

$$\mathcal{K}(\beta)=\left\{z\in\mathcal{R}(F):\frac{1}{\varkappa\cdot\hat{\sigma}^2}(F\beta-z)'(F(X'W^{-1}X)^{-}F')^{-}(F\beta-z)<F_{\varkappa,n-q;\alpha}\right\}, \text{ so ist}$$

$\mathcal{K}(\hat{\beta})$ ein Ellipsoid (möglicherweise entartet) mit dem Mittelpunkt $F\hat{\beta}$. Es folgt

$$\mathsf{P}\left(\mathcal{K}(\hat{\beta})\ni F\beta\right)=1-\alpha. \qquad (1)$$

Die Eigenschaft (1) ist eine wesentliche Charakterisierungsgröße für die sogenannten Konfidenzintervalle.

Definition 9.6 *$\mathcal{C}(y)$ heißt Konfidenzintervall zum Niveau $1-\alpha$ zu vorgegebenem α, wenn*

$$\mathsf{P}_\theta\left(\gamma(\theta)\in\mathcal{C}(y)\right)\geq 1-\alpha \quad \textit{für alle} \quad \theta\in\Theta$$

gilt.

Wenden wir diese Sprechweise auf unser ursprüngliches Problem an, so ist $\mathcal{K}(\hat{\beta})$ ein derartiger Bereich, der $F\beta$ mit der Wahrscheinlichkeit $1-\alpha$ enthält. Folglich ist $\mathcal{K}(\hat{\beta})$ ein Konfidenzbereich für $F\beta$ zum Niveau $1-\alpha$. Allgemein gesagt, stellen die Konfidenzintervalle eine Methode zur Parameteridentifikation dar. Hier sind die Entscheidungen nicht einzelne Punkte, sondern Teilmengen des Parameterraumes (häufig Intervalle, Würfel oder Ellipsoide). Das Problem besteht, will man eine Aussage über einen Parameter $\gamma(\theta)$ machen, in der Angabe einer Funktion $\mathcal{C}$, für deren Werte $\mathcal{C}(y)\subseteq\gamma(\Theta)$ gilt und für die $\mathcal{C}(y)$ den wahren Parameter mit einer vorgegebenen Mindestwahrscheinlichkeit $1-\alpha$ enthält.

Spezialfälle:

1. F sei ein $1\times k$-Zeilenvektor: $F=f'$. Dann ist $\varkappa=1$ und $F_{1,n-q;\alpha}=t^2_{n-q;\alpha/2}$. Das Ellipsoid $\mathcal{K}(\hat{\beta})$ ist dann ein Intervall mit den Grenzen

$$f'\hat{\beta}\pm\hat{\sigma}t_{n-q;\alpha/2}\sqrt{f'(X'W^{-1}X)^{-}f} \qquad (*)$$

2. Betrachten wir für allgemeines F den Konfidenzbereich $\mathcal{K}(\hat{\beta})$. Für jedes $F\beta\in\mathcal{K}(\hat{\beta})$ und jedes $c\in\mathbb{R}^m$ gilt:

$$\inf_{F\beta\in\mathcal{K}(\hat{\beta})}c'F\beta\leq c'F\beta\leq\sup_{F\beta\in\mathcal{K}(\hat{\beta})}c'F\beta.$$

Wegen (1) hatten wir

$$\mathsf{P}\left(\mathcal{K}(\hat{\beta}) \ni F\beta\right) = 1-\alpha$$

und somit

$$\mathsf{P}\left(\inf_{F\beta\in\mathcal{K}(\hat{\beta})} c'F\beta \le c'F\beta \le \sup_{F\beta\in\mathcal{K}(\hat{\beta})} c'F\beta \quad \text{für beliebiges } c\right) \ge 1-\alpha$$

Hilfssatz 9.3 *Es gilt*

$$\begin{aligned}
\sup_{F\beta\in\mathcal{K}(\hat{\beta})} c'F\beta &= c'F\hat{\beta} + \left(c'F\left(X'W^{-1}X\right)^{+}F'c\cdot\varkappa\cdot F_{\varkappa,n-q;\alpha}\cdot\hat{\sigma}^2\right)^{\frac{1}{2}} \\
\inf_{F\beta\in\mathcal{K}(\hat{\beta})} c'F\beta &= c'F\hat{\beta} - \left(c'F\left(X'W^{-1}X\right)^{+}F'c\cdot\varkappa\cdot F_{\varkappa,n-q;\alpha}\cdot\hat{\sigma}^2\right)^{\frac{1}{2}}
\end{aligned}$$

BEWEIS: Setzt man $L = F\left(X'W^{-1}X\right)^{-}F'$, so haben wir

$$\begin{aligned}
& (F\hat{\beta}-F\beta)'L^{+}(F\hat{\beta}-F\beta) && \le \varkappa\hat{\sigma}^2 F_{\varkappa,n-q;\alpha} \\
\Leftrightarrow\quad & (L^{+})^{\frac{1}{2}}(F\hat{\beta}-F\beta)(F\hat{\beta}-F\beta)'(L^{+})^{\frac{1}{2}} && \le \varkappa\hat{\sigma}^2 F_{\varkappa,n-q;\alpha}\cdot I \\
& \text{es liegt } F\hat{\beta}-F\beta \quad \text{in } \mathcal{R}(L) \quad \text{mit Wk 1.} \\
& \text{denn } \sigma^2 L = \mathsf{Var}\, F\hat{\beta} \\
& \text{und für eine ZG } V \text{gilt: } V - \mathsf{E}V \in \mathcal{R}(\mathsf{Var}\, V) \\
\Leftrightarrow\quad & (F\hat{\beta}-F\beta)(F\hat{\beta}-F\beta)' && \le \varkappa\hat{\sigma}^2 F_{\varkappa,n-q;\alpha}\cdot L \\
\Rightarrow\quad & \underbrace{c'(F\hat{\beta}-F\beta)(F\hat{\beta}-F\beta)'c}_{|c'(F\hat{\beta}-F\beta)|^2} && \le \varkappa\hat{\sigma}^2 F_{\varkappa,n-q;\alpha}\cdot c'Lc \\
\Rightarrow\quad & \sup_{F\beta\in\mathcal{K}(\hat{\beta})} \mid c'(F\hat{\beta}-F\beta)\mid^2 && = \varkappa\hat{\sigma}^2 F_{\varkappa,n-q;\alpha}\cdot c'Lc
\end{aligned}$$

Das führt direkt zur Behauptung.

□

Das führt uns nun zu folgenden Aussagen:

1.

$$\begin{aligned}
\text{Setzt man } g &= \frac{1}{\sqrt{f'(X'W^{-1}X)'f}}f, \text{ so} \\
\text{gilt } \mathsf{Var}\, g'\hat{\beta} &= \frac{1}{\sqrt{f'(X'W^{-1}X)^{-1}f}}\cdot\sigma^2\cdot f'(X'W^{-1}X)'f = \sigma^2 \\
&\Rightarrow \quad (1) \quad \text{gibt Konfidenzintervall für } g'\beta \quad \text{für ein festes } g. \\
\text{Länge} &= 2\hat{\sigma}\sqrt{F_{1,n-q;\alpha}}
\end{aligned}$$

2.

$$\begin{aligned} \text{Setzt man } g &= \frac{1}{\sqrt{c'F(X'W^{-1}X)'F'c}}F'c \\ \text{so ist } \mathsf{Var}\, g'\hat{\beta} &= \hat{\sigma}^2 \\ &\Rightarrow \quad (**) \quad \text{gibt Konfidenzintervall für bel. } g = F'c \\ \text{Länge} &= 2\hat{\sigma}\sqrt{\varkappa}\sqrt{F_{\varkappa,n-q;\alpha}} \end{aligned}$$

Setzen wir jetzt $c'F = f'$, so gibt (2.) ein Konfidenzintervall für $g'\beta$, genau wie wir in (1.) ein derartiges hatten. Es gilt stets

$$\varkappa F_{\varkappa,n-q;\alpha} > F_{1,n-q;\alpha},$$

so daß das Konfidenzintervall unter (2.) stets größer ist, aber dieselbe Überdeckungswahrscheinlichkeit $1-\alpha$ nur einhält. Während (1.) nur für ein f gilt, gilt (2.) simultan für alle c bzw. alle $f = F'c \in \mathcal{R}(F')$. Ist F speziell so beschaffen, daß es den Raum aller schätzbaren Funktionen aufspannt, d.h. $\mathcal{R}(F') = \mathcal{R}(X')$ und deshalb $\varkappa = q$ gilt, so sind durch

$$f'\hat{\beta} \pm \hat{\sigma}\left[qF_{q,n-q;\alpha}f'(X'W^{-1}X)^{-}f\right]^{\frac{1}{2}}$$

simultane Konfidenzbereiche für alle schätzbaren Funktionen $f'\beta$ gegeben. Wir haben damit den folgenden Satz bewiesen:

Satz 9.4 *Es gilt*

$$\mathsf{P}\Big(f'\hat{\beta} - \hat{\sigma}\left[qF_{q,n-q;\alpha}f'(X'W^{-1}X)^{-}f\right]^{1/2} \le f'\beta \le f'\hat{\beta} + \hat{\sigma}\left[qF_{q,n-q;\alpha}f'(X'W^{-1}X)^{-}f\right]^{\frac{1}{2}}$$
$$\textit{für alle schätzbaren} \quad f\Big) = 1-\alpha.$$

Beispiel 9.4 Konfidenzbereiche für Regressionsgeraden

In einem Gleichgewichts–Radioimmunoassay kann der Zusammenhang zwischen der gebundenen Radioaktivität R und der Ligand–Konzentration C durch

$$ln\frac{R}{1-R} = \alpha - \gamma \ln C + \tilde{\varepsilon}$$

beschrieben werden. Die zu einigen Standards C beobachteten Werte sind $\left(\tilde{y} = ln\frac{R}{1-R}, x = lnC\right)$:

$\tilde{y}$	2.92	2.87	2.34	1.79	1.68	.98	.81	.66	.14
x	-1.39	-1.04	-.69	-.35	0	.35	.69	1.04	1.39

Führen wir die Scheinvariable $x_{t1} = 1, t = 1, ..., 9$ ein, so haben wir ein lineares Regressionsmodell (wir setzen Unabhängigkeit voraus) und es gilt

$$X'X = \begin{pmatrix} 9 & 0 \\ 0 & 7.22 \end{pmatrix} \quad \text{und} \quad \hat{\beta} = \begin{pmatrix} \hat{\alpha} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} 1.58 \\ 1.04 \end{pmatrix}, \hat{\sigma} = 0.157.$$

Da X vollen Rang hat, sind für beliebiges x die Parameter $d-\gamma x$ schätzbar. Die beiden Konfidenzintervallarten können nun berechnet werden. Es gilt $n=9$, $q=2$

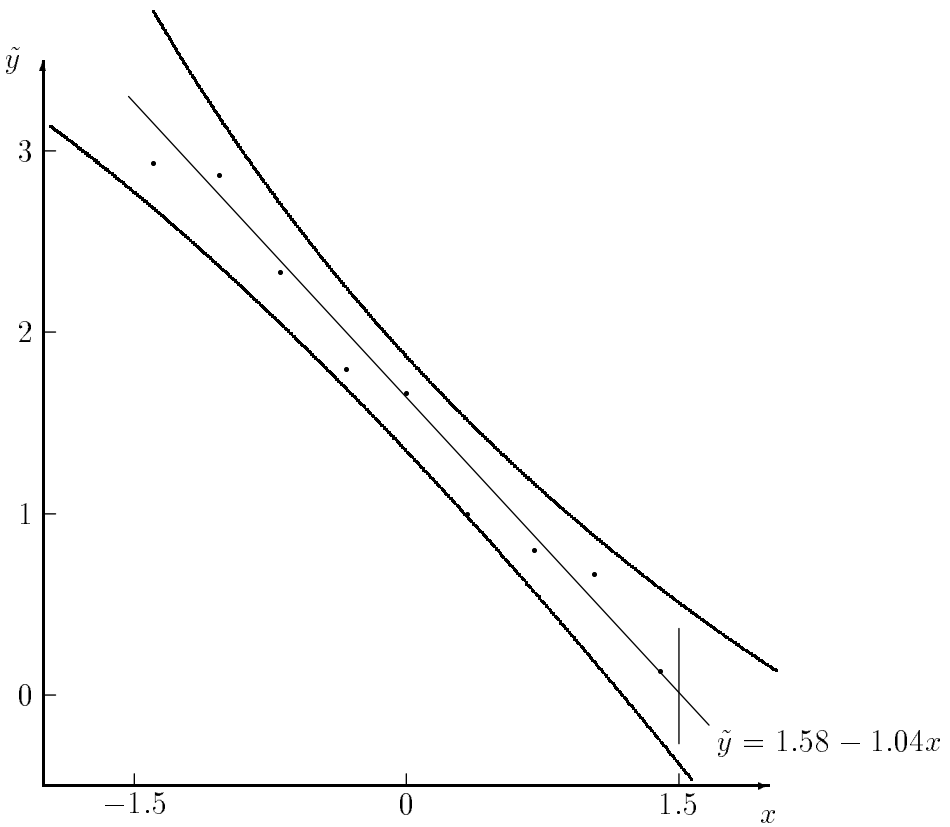
1.Art: Konfidenzintervall ist

$$\begin{aligned} & \hat{\alpha}-\hat{\gamma}x \pm \hat{\sigma}\sqrt{F_{1,7;0.01}}\sqrt{(1\vdots -x)(X'X)^{-1}\begin{pmatrix}1\\-x\end{pmatrix}} \\ = \;& \hat{\alpha}-\hat{\gamma}x \pm \hat{\sigma}\sqrt{F_{1,7;0.01}}\sqrt{0.111+0.1385x^2} \\ = \;& \hat{\alpha}-\hat{\gamma}x \pm \hat{\sigma}3.5\sqrt{0.111+0.1385x^2} \end{aligned}$$

2.Art: Konfidenzintervall (simultanes Konfidenzintervall) ist

$$\begin{aligned} & \hat{\alpha}-\hat{\gamma}x \pm \hat{\sigma}\sqrt{F_{2,7;0.01}}\sqrt{0.111+0.1385x^2}\sqrt{2} \\ = \;& \hat{\alpha}-\hat{\gamma}x \pm \hat{\sigma}4.37\sqrt{0.111+0.1385x^2}\sqrt{2} \end{aligned}$$

Die simultanen Konfidenzgrenzen für die schätzbaren Parameter $\alpha-\gamma x$, d.h. für die Regressionsgerade, sind in der Abbildung gegeben. In $x_*=1.5$ ist zum Vergleich auch das Konfidenzintervall für $\alpha-\gamma x_*$ eingezeichnet. $\left(t^2_{1,7;.01}=12.25, F_{2,7;.01}=9.55\right)$.



Wir können auch für σ^2 einen Konfidenzbereich angeben. Im normalen linearen Modell gilt

$$\frac{s^2}{\sigma^2} \sim \chi^2_{n-q}.$$

Wir hatten $\hat{\sigma}^2 = \frac{s^2}{n-q}$ gesetzt, so daß $\frac{\hat{\sigma}^2}{\sigma^2}(n-q) \sim \chi^2_{n-q}$ gilt. Folglich ergeben sich die drei Beziehungen:

$$\begin{aligned}
\mathsf{P}\left(\frac{n-q}{\sigma^2}\hat{\sigma}^2 \leq \chi^2_{n-q;\alpha}\right) &= 1-\alpha \\
\mathsf{P}\left(\frac{n-q}{\sigma^2}\hat{\sigma}^2 \geq \chi^2_{n-q;1-\alpha}\right) &= 1-\alpha \\
\mathsf{P}\left(\chi^2_{n-q;1-\frac{\alpha}{2}} \leq (n-q)\frac{\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{n-q;\frac{\alpha}{2}}\right) &= 1-\alpha
\end{aligned}$$

Das bedeutet, daß folgendes gilt:

$$\begin{aligned}
\mathsf{P}\left(\frac{(n-q)\hat{\sigma}^2}{\chi^2_{n-q;\alpha}} \leq \sigma^2\right) &= 1-\alpha \\
\mathsf{P}\left(\frac{(n-q)\hat{\sigma}^2}{\chi^2_{n-q;1-\alpha}} \geq \sigma^2\right) &= 1-\alpha \\
\mathsf{P}\left(\frac{(n-q)\hat{\sigma}^2}{\chi^2_{n-q;\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-q)\hat{\sigma}^2}{\chi^2_{n-q;1-\frac{\alpha}{2}}}\right) &= 1-\alpha.
\end{aligned}$$

Dann sind die Intervalle

$$\left(\frac{(n-q)\hat{\sigma}^2}{\chi^2_{n-q;\alpha}}\ ,\ \infty\right), \left(0\ ,\ \frac{(n-q)\hat{\sigma}^2}{\chi^2_{n-q;1-\alpha}}\right), \left(\frac{(n-q)\hat{\sigma}^2}{\chi^2_{n-q;\frac{\alpha}{2}}}, \frac{(n-q)\hat{\sigma}^2}{\chi^2_{n-q;1-\frac{\alpha}{2}}}\right)$$

Konfidenzintervalle für σ^2 zum Niveau $1-\alpha$. Die Längen der Intervalle sind ∞ und $\left(\frac{(n-q)\hat{\sigma}^2}{\chi^2_{n-q;1-\alpha}}\right)$ beziehungsweise $\hat{\sigma}^2(n-q)\left[\frac{1}{\chi^2_{n-q;1-\frac{\alpha}{2}}} - \frac{1}{\chi^2_{n-q;\frac{\alpha}{2}}}\right]$.

9.2 Varianzanalysemodelle

Definition 9.7 *Regressionsmodelle, in denen die Versuchsplanmatrix X nur* 0 *oder* 1 *als Komponenten hat, heißen Varianzanalysemodelle.*

Das bedeutet, daß nur gewisse Parameter auf ein Y_i Einfluß haben. Im allgemeinen wählt man für Varianzanalysemodelle spezielle, explizite Darstellungen, aus denen der konkrete Versuchsplan abgelesen werden kann.

9.2.1 Kreuzklassifikationsmodelle

Wir beginnen mit einem Beispiel.

Beispiel 9.5 Herstellung von Graphitformteilen.

Für Messungen des elektrischen Widerstandes von Graphitformteilen liegen acht Stichproben vom Umfang vier vor. Die Stichproben werden bei zwei verschiedenen Tiegelhöhen (Messungen 1 bis 4 bzw. 5 bis 8) bzw. Endtemperaturen des Graphitierungsprozesses (jeweils 1,5; 2,6; 3,7; 4,8 haben gleiche Endtemperatur) gemessen.

Einzelmessungen des elektrischen Widerstandes:

	Einzelwerte				Mittelwert
t	$y_{t\nu}$				$y_{t.}$
1	64.8	63.2	61.7	65.2	63.725
2	65.7	63.1	62.8	66.3	64.475
3	63.9	62.5	59.9	64.9	62.800
4	65.2	63.2	62.1	65.9	64.100
5	60.2	58.9	56.3	60.6	59.000
6	64.5	62.9	60.1	66.0	63.375
7	61.9	60.5	59.2	60.4	60.500
8	62.5	62.1	60.1	62.6	61.825

Stichproben entstehen hier durch verschiedene Behandlungen. Die $y_{t.}$ sind verschieden. Es fragt sich, ob die verschiedenen Behandlungen, d.h. die verschiedenen Tiegelhöhen oder Temperaturen zu unterschiedlichen Widerständen führen.

Dann gibt es I verschiedene Tiegelhöhen, J verschiedene Endtemperaturen. Die zu messenden Werte des Widerstandes sind Realisierungen von Zufallsgrößen $Y_1, ..., Y_n$. Wir wollen ein der Aufgabenstellung angepaßtes Modell herleiten. Dieses Beispiel soll zur Illustration eines allgemeinen Modells dienen.

Man nimmt an, daß die Einflußgrößen (hier Tiegelhöhen, Temperaturen) die Zielgröße beeinflussen. Diese Einstellgrößen heißen *Faktoren*, deren mögliche Einstellungen heißen *Stufen* der Faktoren. Die zur i-ten Stufe des Faktors 1, zur j-ten Stufe des Faktors 2,..., zur l-ten Stufe des r- ten Faktor gemessenen Werte sollen Realisierungen einer Zufallsgröße $Y_{ij...l}$ sein.

Definition 9.8 *Es seien $Y_{ij...l}$ unabhängige Zufallsgrößen. Werden die einzelnen Stufen der r Faktoren unabhängig voneinander in der Versuchanstellung durchlaufen, so heißt das Modell*

$$Y_{ij...l} = f_{ij...l} + \varepsilon_{ij...l}, \tag{11}$$
$$i = 1, ..., I;\ j = 1, ..., J;\ ...\ l = 1, ..., L; \nu = 1, ..., n_{ij...l}$$

r-faktorielles Kreuzklassifikationsmodell. Falls $n_{ij...l} = \tilde{n}$ für alle $i, j, ..., l$, so heißt das Modell balanciert, andernfalls unbalanciert.

Im Beispiel gilt $r = 2, I = 2, J = 4, \tilde{n} = 4$, so daß zu unserem Beispiel ein zweifaktorielles balanciertes Kreuzklassifikationsmodell gehört. Um die Schreibweise

zu vereinfachen, nehmen wir im folgenden $r = 2$ an. Das ist keine prinzipielle Einschränkung. Wir setzen voraus, daß in (11) die Zufallsgrößen ε_{ij} den Erwartungswert 0 und die Varianz σ^2 haben, also

$$\mathsf{E}\varepsilon_{ij} = 0, \mathsf{E}\varepsilon_{ij}^2 = \sigma^2.$$

Wir setzen

$$\begin{aligned} f_{..} &= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij}, \\ f_{i.} &= \frac{1}{J}\sum_{j} f_{ij}, \quad i = 1, ..., I, \\ f_{.j} &= \frac{1}{I}\sum_{i} f_{ij}, \quad j = 1, ..., J. \end{aligned}$$

Dann hat der Faktor 1 keinen Einfluß, wenn

$$f_{i.} - f_{..} = 0 \quad \text{für alle } i = 1, ..., I.$$

Entsprechend hat der Faktor 2 keinen Einfluß, wenn

$$f_{.j} - f_{..} = 0 \quad \text{für alle } j = 1, ..., J.$$

Gewöhnlich bezeichnet man $\mu = f_{..}$ als Gesamtmittel,

$$\alpha_i = f_{i.} - f_{..}, \quad i = 1, ..., I$$

und

$$\beta_j = f_{.j} - f_{..}, \quad j = 1, ..., J$$

als (Haupt)Effekte der entsprechenden Faktoren sowie

$$\gamma_{ij} = f_{ij} - f_{i.} - f_{.j} + f_{..} (= (\alpha\beta)_{ij})$$

als Effekte 2. Ordnung oder Wechselwirkungen. Dann haben wir als Modell

$$Y_{ij\nu} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij\nu}, \quad i = 1, ..., I; j = 1, ..., J; \nu = 1, ...\tilde{n}, \tag{12}$$

denn es gilt

$$Y_{ij\nu} = f_{..} + (f_{i.} - f..) + (f_{.j} - f_{..}) + (f_{ij} - f_{i.} - f_{.j} + f_{..}) + \varepsilon_{ij\nu}$$

Dieses Modell ist eine andere Darstellung von (11), ist aber inhaltlich besser zu interpretieren. Deshalb findet man meist (12) als Darstellung der 2-faktoriellen Kreuzklassifikation. Die formulierte Frage, ob die Faktoren 1 oder 2 einen Einfluß haben, wird nun durch die Bedingungen
$\alpha_1 = ... = \alpha_I = 0, \quad \beta_1 = ... = \beta_J = 0$ beschrieben.
Wir schreiben dieses Modell als Regresssionsmodell.

$$y = (y_{111}, ..., y_{11s}, y_{121}, ..., y_{IJ\tilde{n}})',$$
$$\beta = (\mu, \alpha_1, ..., \alpha_I, \beta_1, ..., \beta_J, \gamma_{11}, ..., \gamma_{IJ})'$$

$$X = \begin{bmatrix}
1_{\tilde{n}} & 1_{\tilde{n}}0_{\tilde{n}}...0_{\tilde{n}} & 1_{\tilde{n}}0_{\tilde{n}}...0_{\tilde{n}} & 1_{\tilde{n}}0_{\tilde{n}} & ...0_{\tilde{n}} \\
. & . . \quad . & 0_{\tilde{n}} \; . \quad . & 0_{\tilde{n}}1_{\tilde{n}} & \\
. & . . \quad . & . \quad . \quad . & & \\
 & & . \quad . \; 0_{\tilde{n}} & & \\
1_{\tilde{n}} & 1_{\tilde{n}}0_{\tilde{n}}...0_{\tilde{n}} & 1_{\tilde{n}}...0_{\tilde{n}}1_{\tilde{n}} & & \\
\vdots & \vdots & \vdots & & \\
1_{\tilde{n}} & 0_{\tilde{n}}...0_{\tilde{n}}1_{\tilde{n}} & 1_{\tilde{n}}0_{\tilde{n}}...0_{\tilde{n}} & & \\
. & . \quad . . & 0_{\tilde{n}} \; . \quad . & & \\
. & . \quad . . & . \quad . \quad . & & \\
 & & . \quad . \; 0_{\tilde{n}} & & 1_{\tilde{n}}0_{\tilde{n}} \\
1_{\tilde{n}} & 0_{\tilde{n}}...0_{\tilde{n}}1_{\tilde{n}} & 0_{\tilde{n}}...0_{\tilde{n}}1_{\tilde{n}} & 0_{\tilde{n}} & ...0_{\tilde{n}}1_{\tilde{n}} \\
 & \underbrace{\qquad}_{I} & \underbrace{\qquad}_{J} & \underbrace{\qquad\qquad}_{IJ} &
\end{bmatrix}$$

Dann haben wir die Darstellung

$$Y = X\beta + \varepsilon.$$

X besteht nur aus Nullen und Einsen. Man erkennt unmittelbar, daß $rgX = IJ$ gilt, d.h. die $\tilde{n}IJ \times (1 + I + J + IJ)$-Matrix X besitzt keinen vollen Rang. Um die Parameter eindeutig schätzen zu können, stellt man Restriktionen an diese Parameter. Heuristisch sinnvoll sind die Reparametrisierungsbedingungen

$$\sum_i \alpha_i = 0, \sum_j \beta_j = 0,$$
$$\sum_j \gamma_{ij} = 0, \quad i = 1, ..., I;$$
$$\sum_i \gamma_{ij} = 0, \quad j = 1, ..., J.$$

Schreiben wir diese in der Form $R\beta = 0$ für eine $(I + J + 2) \times (1 + I + J + IJ)$-Matrix, wobei hier

$$R = \begin{bmatrix}
0 \; 1...1 \; 0 \; ....0 & & 0 & \\
0 \; ....0 \; 1 \; 1...1 & & & \\
 & 1 \; 1 \; \ldots \; 1 & 0 \; . \; . \; . \; . & . \; . \; . \; . \; 0 \\
 & & \ddots & \\
 & 0 \; . \; \ldots \; . & . \; . \; . & 1 \; 1 \; \ldots \; 1 \\
 & 1 \; 0 \; \ldots \; 0 & 1 \; 0 & 1 \; 0 \; \ldots \; 0 \\
 & 0 \; . & . & . \\
 & \vdots \quad . & . & . \\
 & 0 \; . \; \ldots \; 1 & 0 \; . \; 1 & 0 \; . \; \ldots \; 1
\end{bmatrix}$$

gilt, überzeugt man sich leicht, daß $rgR = I + J + 1$ und $rg(X' : R') = IJ + I + J + 1$ ist. Damit können wir die früheren Ergebnisse anwenden. Als eindeutige KQ-Schätzung für β ergibt sich

$$\hat{\beta}_R\,(y) = (X'X + R'R)^{-1}X'y.$$

In diesem Fall können wir eine explizite Darstellung der Komponenten $\hat{\beta}_R$ angeben. Führen wir die Bezeichnungen

$$\begin{aligned}
y_{ij.} &= \frac{1}{\tilde{n}}\sum_{\nu} y_{ij\nu}, \quad i = 1, ..., I; j = 1, ..., J,\\
y_{i..} &= \frac{1}{\tilde{n}J}\sum_{j}\sum_{\nu} y_{ij\nu}, \quad i = 1, ..., I,\\
y_{.j.} &= \frac{1}{\tilde{n}I}\sum_{i}\sum_{\nu} y_{ij\nu}, \quad j = 1, ..., J,\\
y_{...} &= \frac{1}{\tilde{n}IJ}\sum_{i}\sum_{j}\sum_{\nu} y_{ij\nu},
\end{aligned}$$

ein. Wir wählen die Darstellung

$$y_{ij\nu} = y_{...} + (y_{i..} - y_{...}) + (y_{.j.} - y_{...}) + (y_{ij.} - y_{i..} - y_{.j.} + y_{...}) + (y_{ij\nu} - y_{ij.}).$$

Dadurch werden die folgenden Schätzer für die Parameter suggeriert:

$$\begin{aligned}
\hat{\mu} &= Y_{...},\\
\hat{\alpha}_i &= Y_{i..} - Y_{...}, \quad i = 1, ..., I\\
\hat{\beta}_j &= Y_{.j.} - Y_{...}, \quad j = 1, ..., J,\\
\hat{\gamma}_{ij} &= Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}, \quad i = 1, ..., I; j = 1, ..., J.
\end{aligned}$$

Diese Schätzungen sind erwartungstreu, wie man beispielsweise aus

$$\begin{aligned}
\mathsf{E}\,\hat{\alpha}_i &= \mathsf{E}(Y_{i..} - Y_{...}) = \frac{1}{J}\sum_{j}\mathsf{E}Y_{ij.} - \frac{1}{IJ}\sum_{i}\sum_{j}\mathsf{E}Y_{ij.}\\
&= \frac{1}{J}\sum_{j}(\mu + \alpha_i + \beta_j + \gamma_{ij}) - \frac{1}{IJ}\sum_{i}\sum_{j}(\mu + \alpha_i + \beta_j + \gamma_{ij}) = \alpha_i
\end{aligned}$$

unter Benutzung der Restriktionen sieht. Sie erfüllen auch nach Definition der $Y_{i..}, Y_{.j.}$ und $Y_{...}$ die Reparametrisierungsbedingungen, d.h. $\sum_i \hat{\alpha}_i = 0$ usw. Wegen

$$\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} \equiv Y_{ij.}, \quad i = 1, ..., I; j = 1, ..., J,$$

und

$$\sum_{\nu}(Y_{ij\nu} - Y_{ij.})^2 = \min_{\xi_{ij}}\sum_{\nu}(Y_{ij\nu} - \xi_{ij})^2$$

stellen die $\hat{\mu}, \hat{\alpha}_i, \hat{\beta}_j, \hat{\gamma}_{ij}$ Kleinste-Quadrate–Schätzungen dar, d.h. sie minimieren die Summe der Abweichungsquadrate

$$\sum_i \sum_j \sum_\nu (Y_{ij\nu} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2.$$

Wir interessieren uns für die Gesamtheit der schätzbaren Parameter. Dazu gehören beispielsweise

$$\alpha_1 - \alpha_\cdot + \gamma_{1\cdot} - \gamma_{\cdot\cdot}, ..., \alpha_I - \alpha_\cdot + \gamma_{I\cdot}\gamma_{\cdot\cdot},$$

wobei $\alpha_\cdot = \frac{1}{I}\sum_i \alpha_i, \gamma_{1\cdot} = \frac{1}{J}\sum_j \gamma_{1j}$ usw., und alle aus diesen gebildeten Linearkombinationen, die sogenannten **Kontraste**. Die Schätzbarkeit ergibt sich, indem wir hier

$$F = \begin{bmatrix} 0 & \frac{I-1}{I} & \frac{-1}{I} & \dots & \frac{-1}{I} & 0\dots0 & \frac{I-1}{IJ} & \dots & \frac{I-1}{IJ} & \dots & \frac{-1}{IJ} & \dots & \frac{-1}{IJ} \\ 0 & \frac{-1}{I} & \frac{I-1}{I} & \dots & \frac{-1}{I} & 0\dots0 & \frac{-1}{IJ} & \dots & \frac{-1}{IJ} & \dots & \frac{I-1}{IJ} & \dots & \frac{-1}{IJ} \\ \vdots & & & & & & & & & & & & \\ 0 & \frac{-1}{I} & \frac{-1}{I} & \dots & \frac{I-1}{I} & 0\dots0 & \frac{-1}{IJ} & \dots & \frac{-1}{IJ} & \dots & \frac{-1}{IJ} & \dots & \frac{I-1}{IJ} \end{bmatrix}$$

wählen und sehen, daß die Schätzbarkeitsbedingung erfüllt ist.

Bildet man $F\hat{\beta}$, d.h. $F(X'X)^- X'Y$, so erhält man schließlich $(Y_{1\cdot\cdot} - Y_{\cdot\cdot\cdot}, ..., Y_{I\cdot\cdot} - Y_{\cdot\cdot\cdot})$ als eindeutigen Kleinste–Quadrate–Schätzer für $(\alpha_1 - \alpha_\cdot + \gamma_{1\cdot}\gamma_{\cdot\cdot}, ..., \alpha_I - \alpha_\cdot + \gamma_{I\cdot} - \gamma_{\cdot\cdot})$. Entsprechende Aussagen ergeben sich auch bei Ersetzung von α durch β. Dagegen sind z.B. die Funktionen $\sum_i \alpha_i, \sum_j \beta_j$ usw. nicht schätzbar.

Satz 9.5 *Die IJ linearen Funktionen*

$$\mu + \alpha_i + \beta_j + \gamma_{ij}$$

sind schätzbar und bilden eine Basis aller schätzbaren linearen Funktionen.

Die Größen $\gamma_{ij} - \gamma_{i\cdot} - \gamma_{\cdot j} + \gamma_{\cdot\cdot}$ sind daher schätzbar. Weiter bemerken wir, daß

$$\begin{aligned} \mathsf{E}\sum_\nu (Y_{ij\nu} - Y_{ij\cdot})^2 &= \mathsf{E}\sum_\nu (\varepsilon_{ij\nu} - \varepsilon_{ij\cdot})^2 \\ &= \sum_\nu \mathsf{E}\varepsilon_{ij\nu}^2 - \tilde{n}\mathsf{E}\varepsilon_{ij\cdot}^2 = \tilde{n}\sigma^2 - \sigma^2 = (\tilde{n}-1)\sigma^2. \end{aligned}$$

gilt. Daher ist für $\tilde{n} > 1$

$$s^2 = \frac{1}{IJ(\tilde{n}-1)} \sum_i \sum_j \sum_\nu (Y_{ij\nu} - Y_{ij\cdot})^2$$

eine erwartungstreue Schätzung von σ^2.

Zerlegung der Varianz

Aus der Darstellung

$$y_{ij\nu} = y_{...} + (y_{i..} - y_{...}) + (y_{.j.} - y_{...}) + (y_{ij.} - y_{i..} - y_{.j.} + y_{...}) + (y_{ij\nu} - y_{ij.})$$

erhält man sofort die Zerlegung der quadratischen Abweichung

$$\sum_i \sum_j \sum_\nu (y_{ij\nu} - y_{...})^2 = \tilde{n} J \sum_i (y_{i..} - y_{...})^2 + \tilde{n} I \sum_j (y_{.j.} - y_{...})^2$$
$$+\tilde{n} \sum_j \sum_j (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2 + \sum_i \sum_j \sum_\nu (y_{ij\nu} - y_{ij.})^2.$$

Es sei hier ausdrücklich vermerkt, daß diese Zerlegung darauf beruht, daß für alle Stufen der Faktoren die gleiche Anzahl $\tilde{n}$ von Beobachtungen vorliegt.

Es ist in der Literatur üblich, den ersten Faktor kurz mit A, den zweiten mit B und die Wechselwirkungen mit AB zu bezeichnen, und die entsprechenden Summen der Quadrate durch

$$\begin{aligned} SQ_A &= \tilde{n} J \sum_i (Y_{i..} - Y_{...})^2, \\ SQ_B &= \tilde{n} I \sum_j (Y_{.j.} - Y_{...})^2, \\ SQ_{AB} &= \tilde{n} \sum_i \sum_j (Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2 \end{aligned}$$

abkürzen. Für die Summe der Quadrate der Residuen wird

$$SQ_R = \sum_i \sum_j \sum_\nu (Y_{ij\nu} - Y_{ij.})^2$$

geschrieben. Daraus folgt mit

$$\begin{aligned} SQ_G &= \sum_i \sum_j \sum_\nu (Y_{ij\nu} - Y_{...})^2 \quad \text{die Zerlegung} \\ SQ_G &= SQ_A + SQ_B + SQ_{AB} + SQ_R. \end{aligned}$$

Seien nun die Fehler $\varepsilon_{ij\nu}, i = 1, ..., I; j = 1, ..., J; \nu = 1, ..., \tilde{n}$, normalverteilt. Wir können mit $\varepsilon_{ij\nu} \sim \mathsf{N}(0, \sigma^2)$

$$\begin{aligned} \varepsilon_{ij\nu} &\sim \mathsf{N}(0, \sigma^2) \\ Y_A &= \sqrt{\tilde{n} J}(Y_{1..}, ..., Y_{I..})' \qquad \text{und} \\ P_A &= I_I - \mathbf{1}_I \mathbf{1}_I' \cdot \frac{1}{I} \end{aligned}$$

die Summe der Quadrate als auch

$$SQ_A = Y_A' P_A Y_A$$

schreiben. Da P_A idempotent und vom Rang $I - 1$ ist, und

$$\begin{aligned} Y_A &\sim N_I(\mu_A, \sigma^2 I) \qquad \text{mit} \\ \mu_A &= \sqrt{\tilde{n} J}(f_{1.}, ..., f_{I.})' \end{aligned}$$

gilt, hat $\frac{1}{\sigma^2} \cdot SQ_A$ eine χ^2-Verteilung mit $I-1$ Freiheitsgraden und dem Nichtzentralitätsparameter

$$\psi_A^2 = \tilde{n} J \sum_i \frac{\alpha_i^2}{\sigma^2}.$$

Analoge Aussagen lassen sich auch für die anderen Quadratsummen ableiten und sind in der Varianztafel dargestellt.

Varianztafel für die zweifache Kreuzklassifikation

Variationsursache	SQ	Freiheitsgrade	Nichtzentralitätsparameter
Faktor A	SQ_A	$I-1$	$\tilde{n} J \sum_i \frac{\alpha_i^2}{\sigma^2}$
Faktor B	SQ_B	$J-1$	$\tilde{n} I \sum_j \frac{\beta_j^2}{\sigma^2}$
Wechselwirkung-AB	SQ_{AB}	$(I-1)(J-1)$	$\tilde{n} \sum_i \sum_j \frac{\gamma_{ij}^2}{\sigma^2}$
Rest	SQ_R	$(\tilde{n}-1)IJ$	0

Durch Ausmultiplizieren kann man sich davon überzeugen, daß die Kovarianzen von $(Y_{i..} - Y_{...})$ und $(Y_{.j.} - Y_{...})$ oder $(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})$ usw. Null sind. Daraus folgt, daß die entsprechenden Vektoren bzw. die quadratischen Formen SQ_A, SQ_B, SQ_{AB} und SQ_R in der Varianztafel unabhängig verteilt sind. Somit besitzen die Quotienten $SQ_A/SQ_R, SQ_B/SQ_R$ und SQ_{AB}/SQ_R - normiert mit den entsprechenden Freiheitsgraden - jeweils eine F-Verteilung.

Beispiel 9.6 Die aus dem Beispiel 9.5 stammenden Daten sind entsprechend ihrer Zugehörigkeit zu den Stufen der einzelnen Faktoren angeordnet.

	Faktor B							
Faktor A	1. Stufe		2. Stufe		3. Stufe		4. Stufe	
1. Stufe	64.8	63.2	65.7	63.1	63.9	62.5	65.2	63.2
	61.7	65.2	62.8	66.3	59.9	64.9	62.1	65.9
2. Stufe	60.2	58.9	64.5	62.9	61.9	60.5	62.5	62.1
	58.3	60.6	60.1	66.0	59.2	60.1	60.1	62.6

Die Varianzanalyse–Tafel zu diesen Werten ist dann folgendermaßen gegeben:

Variationsursache	SQ	Freiheitsgrade	F-Wert
Faktor A	54.08	1	16.473
Faktor B	34.07	3	3.459
Wechselwirkung-AB	13.92	3	1.413
Rest	78.79	24	
Gesamt	180.88	31	

Zum Prüfen der Hypothese

$$H : \alpha_1 = 0 \quad \text{gegen} K : \alpha_1 \neq 0$$

bilden wir die F-Statistik

$$\frac{24}{1}\frac{SQ_A}{SQ_R} = 16.473.$$

Dieser Wert ist größer als der entsprechende 5%-Signifikanzpunkt 4.26. Wir können somit bei diesem Signifikanzniveau die Hypothese, daß die Effekte des ersten Faktors verschwinden, ablehnen.

Für den zweiten Faktor erhalten wir $24SQ_B/3SQ_R = 3.459$. Da $F_{3.24;0.05} = 3.01$ ist, lehnen wir auch die Hypothese der Gleichheit der Effekte $\beta_1, ..., \beta_4$ ab. Schließlich ist $24SQ_{AB} = 3SQ_R = 1.41$, und somit sind die Wechselwirkungen nicht signifikant von Null verschieden.

Klassifikation ohne Wechselwirkung

Das 2-faktorielle Modell stellt nicht in jeder Situation, in der zwei Einflußfaktoren kombiniert werden, die einzig sinnvolle Beschreibung der Zusammenhänge dar.

Es sei aus sachlogischen oder anderen Gründen bekannt, daß der Unterschied, den zwei verschiedene Stufen des Faktors A bei der gleichen Stufe j des Faktors B bewirken, unabhängig von j ist, genauer sei

$$f_{ij} - f_{i'j} = f_{ij'} - f_{i'j'}, \quad i, i' = 1, ..., I, j, j' = 1, ..., J.$$

Summation über i' und j' ergibt

$$f_{ij} - f_{i.} - f_{.j} + f_{..} = 0, \quad i = 1, ..., I; j = 1, ..., J,$$

d.h., es sind keine Wechselwirkungen vorhanden. Wir wählen die Schreibweise

$$\begin{aligned} Y_{ij\nu} &= \mu + \alpha_i + \beta_j + \varepsilon_{ij\nu}, \quad \nu = 1, ..., \tilde{n}; i = 1, ..., I; j = 1, ..., J; \\ \sum_i \alpha_i &= \sum_j \beta_j = 0. \end{aligned}$$

Alle obigen Ausführungen zum Modell mit Wechselwirkungen lassen sich entsprechend auf dieses Modell übertragen. Wir bemerken, daß in der Varianztafel jetzt SQ_{AB} zentral χ^2 verteilt ist, so daß bei der Bildung von Teststatistiken anstelle von SQ_R die Größe $SQ_R + SQ_{AB}$ mit entsprechend mehr Freiheitsgraden treten kann.

Während bisher die Bedingung $\tilde{n} > 1$ für die Existenz eines von Null verschiedenen SQ_R und damit für die Konstruktion der F-Tests notwendig war, ergeben sich jetzt auch für $\tilde{n} = 1$ sinnvolle Resultate, vorausgesetzt $I, J > 1$.

K-faktorielle Klassifikation

Ganz analoge Überlegungen wie im 2-faktoriellen Modell können getroffen werden, wenn man es mit der kombinierten Wirkung von K Faktoren zu tun hat. Für $K = 3$ beispielsweise erhält man

$$Y_{ijl\nu} \equiv \mu + \alpha_i + \beta_j + \gamma_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{il} + (\beta\gamma)_{jl} + (\alpha\beta\gamma)_{ijl} + \varepsilon_{ijl\nu}$$
$$i = 1, ..., I; j = 1, ..., J; l = 1, ..., L; \nu = 1, ..., \tilde{n}$$

mit den Reparametrisierungsbedingungen

$$\begin{aligned}
\sum_i \alpha_i &= \sum_j \beta_j = \sum_l \gamma_l = 0, \\
\sum_j (\alpha\beta)_{ij} &= 0, \quad i = 1, ..., I, \\
\sum_i (\alpha\beta)_{ij} &= 0, \quad j = 1, ..., J, \\
&\vdots \\
\sum_i (\alpha\beta\gamma)_{ijl} &= 0, \quad j = 1, ..., J; l = 1, ..., L.
\end{aligned}$$

Diese Reparametrisierungsbedingungen sind äquivalent zu

$$\begin{aligned}
\mu &= f_{...} \\
\alpha_i &= f_{i..} - f_{...} \\
&\vdots \\
(\alpha\beta\gamma)_{ijl} &= f_{ijl} - f_{ij.} - f_{i.l} - f_{.jl} + f_{i..} + f_{.j.} + f_{..l} - f_{...}
\end{aligned}$$

und

$$\begin{aligned}
f_{ij.} &= \frac{1}{L} \sum_l f_{ijl}, \quad i = 1, ..., I; j = 1, ..., J, \\
&\vdots \\
f_{i..} &= \frac{1}{JL} \sum_j \sum_l f_{ijl}, \quad i = 1, ..., I, \\
f_{...} &= \frac{1}{IJL} \sum_i \sum_j \sum_l f_{ijl.}
\end{aligned}$$

Schreibt man $(\alpha\beta\gamma)_{ijl}$ in der Form

$$(\alpha\beta\gamma)_{ijl} = f_{ijl} - (\alpha\beta)_{ij} - (\alpha\gamma)_{il}(\beta\gamma)_{jl} - \alpha_i - \beta_j - \gamma_l - \mu,$$

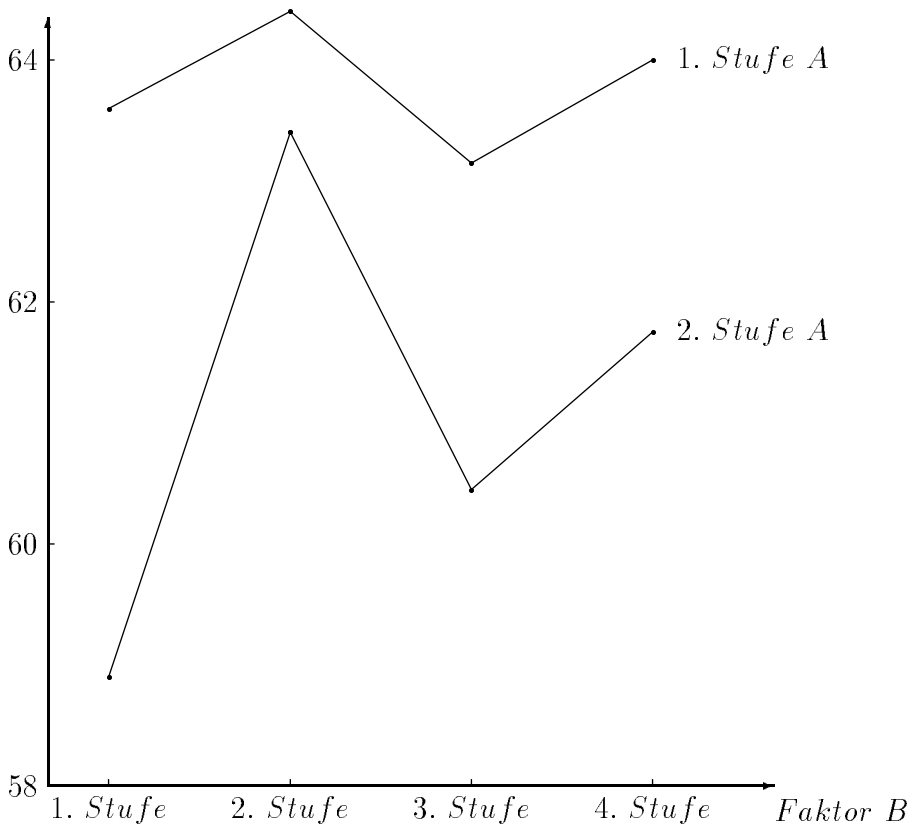
so wird vielleicht die Bedeutung der Effekte 3. Ordnung klarer. Die Zerlegung der Varianz in unabhängige Summen von Quadraten kann bei der balancierten K-faktoriellen Kreuzklassifikation in ähnlicher Weise, wie sie für die 2fache Klassifikation angegeben wurde, durchgeführt werden. In Tabelle 1 ist sie für $K = 3$ angegeben.

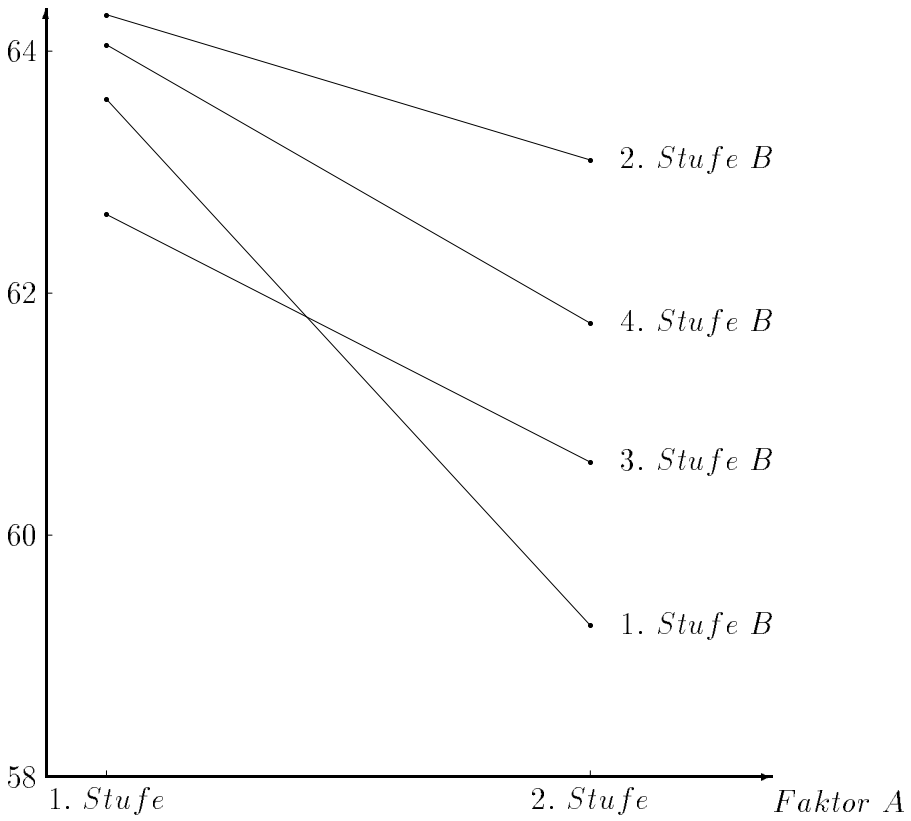
Variationsursache	SQ	Freiheitsgrade	N
Faktor A	$SQ_A = \tilde{n}JL\sum_i (Y_{i\ldots} - Y_{\ldots\cdot})^2$	$I-1$	$\tilde{n}$
Faktor B	$SQ_B = \tilde{n}IL\sum_j (Y_{\cdot j\cdot} - Y_{\ldots\cdot})^2$	$J-1$	$\tilde{n}$
Faktor C	$SQ_C = \tilde{n}IJ\sum_l (Y_{\cdot\cdot l} - Y_{\ldots\cdot})^2$	$L-1$	$\tilde{n}$
Wechselwirkung AB	$SQ_{AB} = \tilde{n}\sum_i\sum_j (Y_{ij\cdot} - Y_{i\ldots} - Y_{\cdot j\cdot} + Y_{\ldots\cdot})^2$	$(I-1)(J-1)$	$\tilde{n}$
Wechselwirkung AC	$SQ_{AC} = \tilde{n}\sum_i\sum_l (Y_{i\cdot l} - Y_{i\ldots} - Y_{\cdot\cdot l} + Y_{\ldots\cdot})^2$	$(I-1)(L-1)$	$\tilde{n}$
Wechselwirkung BC	$SQ_{BC} = \tilde{n}\sum_j\sum_l (Y_{\cdot jl} - Y_{\cdot j\cdot} - Y_{\cdot\cdot l} + Y_{\ldots\cdot})^2$	$(J-1)(L-1)$	$\tilde{n}$
Wechselwirkung ABC	$SQ_{ABC} = \tilde{n}\sum_i\sum_j\sum_l (Y_{ijl} - Y_{ij\cdot} - Y_{i\cdot l} - Y_{\cdot jl} + Y_{i\ldots} + Y_{\cdot j\cdot} + Y_{\cdot\cdot l} - Y_{\ldots\cdot})^2$	$(I-1)(J-1)(L-1)$	$\tilde{n}$
Rest	$SQ_R = \sum_i\sum_j\sum_l\sum_\nu (Y_{ijl\nu} - Y_{ijl\cdot})^2$	$(\tilde{n}-1)IJL$	0
Gesamt	$SQ_G = \sum_i\sum_j\sum_l\sum_\nu (Y_{ijl\nu} - Y_{\ldots\cdot})^2$	$IJL-1$	

Genau wie im Falle der 2fachen Klassifikation kann auch für allgemeines K vorgegangen werden, wenn man Modelle betrachtet, in denen Effekte bestimmter Ordnungen nicht vorhanden sind.

Interpretation der Effekte

Eine einfache grafische Methode zur Beurteilung möglicher Wechselwirkungen in 2faktoriellen Modellen ist durch I Polygonzüge gegeben, welche die Mittelwerte $y_{i1.}, ..., y_{iJ.}$ verbinden, $i = 1, ..., I$.





Ungefähre Parallelität der Kurven deutet auf geringe oder verschwindende Wechselwirkungen hin. Ob eine optisch erkennbare Abweichung von der Parallelität statistisch signifikant ist, kann durch den F-Test der Varianzanalyse ermittelt werden:
$F = (\tilde{n} - 1)IJSQ_{AB}/[(I-1)(J-1)SQ_R]$.

Für den Fall der Ablehnung der entsprechenden Null–Hypothesen kann eine Betrachtung der Grafiken mitunter wesentliche Hinweise auf die Struktur dieser Wechselwirkungen und damit auf die Art des Einflusses der verschiedenen Stufen geben.

9.2.2 Hierarchische Effekte

Zur Analyse der Gewichtszunahme durch zwei verschiedene Fütterungstechnologien mögen von acht Ställen mit gleich vielen Versuchstieren je vier eine dieser Technologien verwenden. Wir können, in dem wir die zu jeder Technologie gehörenden Ställe jeweils von eins bis vier numerieren, wieder

$$Y_{ij\nu} = f_{ij} + \varepsilon_{ij\nu}$$

als Modell ansetzen und, ausgehend von der gegebenen Zerlegung

$$f_{ij} = f_{..} + (f_{i.} - f_{..}) + (f_{.j} - f_{..}) + (f_{ij} - f_i - f_{.j} + f_{..}),$$

die Schätzungen und Varianzzerlegungen im Kreuzklassifikations-Modell anwenden.

Dieses Vorgehen ist aber wenig befriedigend, da z.B. $f_{.1} - f_{..}$ der "Effekt des 1. Stalls" keinen sinnvoll interpretierbaren oder interessierenden Parameter darstellt. Diese oder eine ähnlichen Problemstellung, in der eine Stufe des zweiten Faktors nicht mit allen, sondern nur mit einer Stufe des ersten Faktors kombiniert werden kann, ist die Zerlegung

$$f_{ij} \equiv f_{..} + (f_{i.} - f_{..}) + (f_{ij} - f_{i.})$$

und entsprechend

$$Y_{ij\nu} = f_{..} + (f_{i.} - f_{..}) + (f_{ij} - f_{i.}) + \varepsilon_{ij\nu}, \quad \nu = 1, ..., \tilde{n}; i = 1, ..., I; j = 1, ..., J, (*)$$

besser angepaßt. Es ist üblich, mit $\mu = f_{..}$ das Gesamtmittel, mit

$$\alpha_i = f_{i.} - f_{..}, \quad i = 1, ..., I,$$

den Effekt der i-ten Stufe des ersten Faktors und mit

$$\beta_{ij} = f_{ij} - f_{i.}, \quad i = 1, ..., I; j = 1, ..., J,$$

den Effekt der j-ten Stufe des zweiten Faktors innerhalb der i-ten Stufe des ersten Faktors zu bezeichnen, so daß (*) in der Form

$$Y_{ij\nu} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ij\nu}, \quad \nu = 1, ..., \tilde{n}; i = 1, ..., I; j = 1, ..., J, \qquad (**)$$

geschrieben werden kann.

Wie bei der Kreuzklassifikation sind, solange die $f_{..}$ und $f_{i.}$ nicht festgelegt werden, μ, α_i und β_{ij} auch nicht eindeutig bestimmt. Eindeutigkeit wird erreicht durch die Reparametrisierungsbedingungen

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_{ij} = 0, \quad i = 1, ..., I.$$

Ein derartiges Modell (**) mit den Reparametrisierungsbedingungen heißt balanciertes, zweifach hierarchisches Klassifikationsmodell.

Schätzungen

Ähnlich wie bei der Kreuzklassifikation sind

$$\begin{aligned} \hat{\mu} &= Y_{...} \\ \hat{\alpha}_i &= Y_{i..} - Y_{...}, \quad i = 1, ..., I, \\ \hat{\beta}_{ij} &= Y_{ij.} - Y_{i..}, \quad i = 1, ..., I; j = 1, ..., J \end{aligned}$$

und

$$s^2 = \frac{1}{(\tilde{n} - 1)IJ} \sum_i \sum_j \sum_\nu (Y_{ij\nu} - Y_{ij.})^2$$

erwartungstreue Schätzungen der entsprechenden Parameter, erfüllen die Reparametrisierungsgleichungen und ergeben sich auch als Kleinste-Quadrate–Schätzungen für die Summe der Abweichungsquadrate

$$\sum_i \sum_j \sum_\nu (Y_{ij\nu} - \mu - \alpha_i - \beta_{ij})^2.$$

Zerlegung der Varianz

Die Darstellung

$$y_{ij\nu} \equiv y_{...} + (y_{i..} - y_{...}) + (y_{ij.} - y_{i..}) + (y_{ij\nu} - y_{ij.})$$

führt zur Varianzanalyse–Tafel.

Varianztafel für die balancierte zweifach hierarchische Klassifikation

Variationsursache	SQ	Freiheits-grade	Nichtzentralität von SQ/σ^2
Faktor A	$SQ_A = \tilde{n}J\sum_i (Y_{i..} - Y_{...})^2$	$I-1$	$\tilde{n}J\sum_i \alpha_i^2/\sigma^2$
Faktor B in A	$SQ_B = \sum_i \sum_j (Y_{ij.} - Y_{i..})^2$	$I(J-1)$	$\tilde{n}\sum_i \sum_j \beta_{ij}^2/\sigma^2$
Rest	$SQ_R = \sum_i \sum_j \sum_\nu (Y_{ij\nu} - Y_{ij.})^2$	$(\tilde{n}-1)IJ$	0
Gesamt	$SQ_G = \sum_i \sum_j \sum_\nu (Y_{ij\nu} - Y_{...})^2$	$\tilde{n}IJ-1$	

Hier haben wir unter Annahme normalverteilter Fehler unabhängig χ^2-verteilte Quadratsummen. Um deutlicher zu machen, daß der zweite Faktor B dem Faktor A hierarchisch untergeordnet ist, sind anstelle SQ_B auch die Bezeichnungsweisen $SQ_{B:A}$ oder $SQ_{B<A}$ üblich.

K-fach hierarchische Klassifikation

Die Verallgemeinerung eines balancierten hierarchischen Modells auf K einander hierarchisch untergeordnete Faktoren ist

$$Y_{ij...l\nu} = \mu + \alpha_i + \beta_{ij} + ... + \delta_{ij...l} + \varepsilon_{ij...l\nu},$$
$$\nu = 1, ..., \tilde{n}; i = 1, ..., I; ...; l = 1, ..., L,$$

mit den Reparametrisierungsbedingungen

$$
\begin{array}{lcl}
\sum_i \alpha_i & = & 0 \\
\sum_j \beta_{ij} & = & 0, \quad i = 1, ..., I, \\
 & . & \\
 & . & \\
 & . & \\
\sum_l \delta_{ij...l} & = & 0, \quad i = 1, ..., I; j = 1, ..., J, ...
\end{array}
$$

Die entsprechenden Schätzungen und die Zerlegung der Varianz ergeben sich völlig analog.

9.2.3 Gemischte Klassifikationstypen

In dem zu Beginn des vorigen Abschnitts aufgeführten Beispiel mögen die Versuchstiere neben Fütterungsregime und Stall noch nach einem weiteren Merkmal, dem Geschlecht, klassifiziert sein. Für diese Situation der dreifachen Klassifikation, die weder eine hierarchische noch eine Kreuzklassifikation darstellt ist die folgende Zerlegung der Mittelwerte

$$
\begin{array}{lcl}
f_{ijl} & \equiv & f_{...} + (f_{i..} - f_{...}) + (f_{ij.} - f_{i..}) + (f_{..l} - f_{...}) \\
 & + & (f_{i.l} - f_{i..} - f_{i.l} + f_{...}) + (f_{ijl} - f_{ij.} - f_{i.l} + f_{i..})
\end{array}
$$

sinnvoll und interpretierbar. Bei passender Definition der entsprechenden Parameter bzw. Reparametrisierungsbedingungen können wir

$$
\begin{array}{l}
Y_{ijl\nu} = \mu + \alpha_i + \beta_{ij} + \gamma_l + (\alpha\gamma)_{il} + (\beta\gamma)_{ijl} + \varepsilon_{ijl\nu} \\
\quad i = 1, ..., I; j = 1, ..., J; l = 1, ..., L; \nu = 1, ..., \tilde{n},
\end{array}
$$

schreiben. Die dazugehörige Varianzanalyse ist unten beschrieben.

Welches gemischte Modell man bei gegebenen Faktoren ansetzt, hängt von den konkreten Umständen und der praktischen Situation ab.

Varianztafel für die gemischte Klassifikation

Variations-ursache	SQ	Freiheits-grade	Nichtzentralität von SQ/σ^2
Faktor A	$SQ_A = \tilde{n}JL\sum_i (Y_{i\cdots} - Y_{\cdots})^2$	$I-1$	$\tilde{n}JL\sum_i \alpha_i^2/\sigma^2$
Faktor B in A	$SQ_B = \tilde{n}L\sum_i\sum_j (Y_{ij\cdot} - Y_{i\cdots})^2$	$I(J-1)$	$\tilde{n}L\sum_i\sum_j \beta_{ij}^2/\sigma^2$
Faktor C	$SQ_C = \tilde{n}JJ\sum_l (Y_{\cdot\cdot l\cdot} - Y_{\cdots\cdot})^2$	$L-1$	$\tilde{n}IJ\sum_l \gamma_l^2/\sigma^2$
Wechsel-wirkung AC	$\tilde{n}J\sum_i\sum_l (Y_{i\cdot l\cdot} - Y_{i\cdots} - Y_{\cdot\cdot l\cdot} + Y_{\cdots\cdot})^2$	$(I-1)(L-1)$	$\tilde{n}J\sum_i\sum_l (\alpha\gamma)_{il}^2/\sigma^2$
Wechsel-wirkung BC in A	$\tilde{n}\sum_i\sum_j\sum_l (Y_{ijl\cdot} - Y_{ij\cdot} - Y_{i\cdot l\cdot} + Y_{i\cdots})^2$	$I(J-1)(L-1)$	$\tilde{n}\sum_i\sum_j\sum_l (\beta\gamma)_{ijl}^2/\sigma^2$
Rest	$SQ_R = \sum_i\sum_j\sum_l\sum_\nu (Y_{ijl\nu} - Y_{ijl\cdot})^2$	$(\tilde{n}-1)IJL$	0
Gesamt	$SQ_G = \sum_i\sum_j\sum_l\sum_\nu (Y_{ijl\nu} - Y_{\cdots\cdot})^2$	$\tilde{n}IJL-1$	

Ausgehend von inhaltlichen Überlegungen kommt man über Zerlegungen wie in den vorangegangenen balancierten Modellen ohne größere Schwierigkeiten zu den entsprechenden Schätzungen und Varianzanalysen. Es sei noch erwähnt, daß beispielsweise für drei Faktoren eine Kreuzklassifikation oft durch $A \times B \times C$ symbolisiert wird, eine hierarchische durch $C \prec B \prec A$ und die gemischte Klassifiktion durch $(B \prec A) \times C$. Die entsprechenden Erweiterungen sind naheliegend.

Die bisher betrachteten Modelle hatten die Besonderheit, daß in jeder Zelle, d.h. für jede Kombination der Stufen, gleich viele Beobachtungen vorliegen. Das ist aber im allgemeinen nicht erfüllt. Man muß mit unterschiedlichen Besetzungszahlen n_{ij} arbeiten:

$$Y_{ij\nu} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij\nu}, \quad \nu = 1, ..., n_{ij}$$

wäre eine unbalancierte 2-faktorielle Kreuzklassifikation. Für mehrfaktorielle und hierarchische Modelle ist das analog. Durch die Unbalanciertheit verändern sich die statistischen Eigenschaften der heuristisch naheliegenden Statistiken. Die Reparametrisierungsbedingungen hängen von den Besetzungszahlen im allgemeinen ab (siehe z.B. Läuter/Pincus S. 94ff.).

9.3 Multiple Mittelwertsvergleiche

Betrachten wir wie anfangs k normalverteilte Stichproben

$$\begin{array}{cccc} y_{11}, & \cdots & , y_{1n_1} \\ & \vdots & \\ y_{k1} & , ..., & y_{kn_k} \end{array}$$

mit gleichen Varianzen. Dabei gilt $y_{ij} = Y_{ij}(\omega)$ und $Y_{ij} \sim N(\mu_i, \sigma^2), j = 1, ..., n_i$; $i = 1, ..., k$. Sei $n = n_1 + ... + n_k$. Betrachten wir speziell $k = 2$ und untersuchen das Testproblem

$$\begin{aligned} H : \mu_1 &= \mu_2 \\ \text{gegen } K : \mu_1 &\neq \mu_2. \end{aligned}$$

Zum Testen erinnern wir uns an die allgemeine lineare Hypothese:
Es sei $Y \sim \mathsf{N}(X\beta, \sigma^2 W)$ und getestet werden soll

$$\begin{aligned} H : \ G\beta &= g \quad \text{gegen} \\ K : \ G\beta &\neq g. \end{aligned}$$

Die Lösung liefert der F-Test der linearen Hypothese. Dazu erinnern wir uns an die Bezeichnungen

$$\begin{aligned} \mid M \mid^2 &= (G\hat{\beta} - g)'(GKG')^{-}(G\hat{\beta} - g) \sim \sigma^2 \chi^2_{\varkappa} \quad \text{für } \varkappa = rgG \ , K = (X'W^{-1}X)^{+} \\ s^2 &= Y'W^{\frac{-1}{2}}(I - P_{\mathcal{R}(W^{-1}X)})W^{\frac{-1}{2}}Y \sim \sigma^2 \chi^2_{n-q} \quad q = rgX \end{aligned}$$

Gilt speziell

$$\begin{aligned} Y_{11}, ..., Y_{1n_1} &\sim \mathsf{N}(\mu_1, \sigma^2) \qquad i.i.d. \\ Y_{21}, ..., Y_{2n_2} &\sim \mathsf{N}(\mu_2, \sigma^2) \qquad i.i.d. \end{aligned}$$

so haben wir

$$\begin{aligned} Y &= (Y_{11}, ..., Y_{1n_1}, Y_{21}, ..., Y_{2n_2})' \\ Y &= X\beta + \varepsilon \ , \qquad \varepsilon \sim \mathsf{N}(0, \sigma^2 I) \\ & \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad , n = n_1 + n_2 \\ \Rightarrow \ & X = \begin{pmatrix} 1_{n_1} & 0 \\ 0 & 1_{n_2} \end{pmatrix} \qquad \text{mit } rgX = 2 = q. \end{aligned}$$

Es soll speziell folgendes Problem getestet werden:

$$\begin{aligned} H : \ \mu_1 - \mu_2 &= 0 \qquad \mu_1 - \mu_2 = (1, -1)\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = G\beta \\ K : \ \mu_1 - \mu_2 &\neq 0 \\ \Rightarrow g = 0 \qquad & , \varkappa = 1. \end{aligned}$$

Wir hatten

$$K = (X'W^{-1}X)^{+} = (X'X)^{+} = (X'X)^{-1} = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix},$$

$$GKG' = (1, -1)\begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{1}{n_1} + \frac{1}{n_2} = \frac{n_1 + n_2}{n_1 n_2},$$

$(GKG')^- = \frac{n_1 n_2}{n_1+n_2}$, $\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix} \begin{pmatrix} \sum_i Y_{1i} \\ \sum_i Y_{2i} \end{pmatrix} = \begin{pmatrix} Y_{1.} \\ Y_{2.} \end{pmatrix}$

$G\hat{\beta} = Y_{1.} - Y_{2.}$

Außerdem gilt:

$s^2 = Y'(I - P_{\mathcal{R}(X)})Y$. $\mathcal{R}(X)$ hat die Basis $e_1 = \frac{1}{\sqrt{n_1}} \begin{pmatrix} 1_{n_1} \\ 0 \end{pmatrix}$, $e_2 = \frac{1}{\sqrt{n_2}} \begin{pmatrix} 0 \\ 1_{n_2} \end{pmatrix}$

Es gilt somit:

$$\begin{aligned} P_{\mathcal{R}(X)} &= e_1 e_1' + e_2 e_2' = \frac{1}{n_1} \begin{pmatrix} 11' & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{n_2} \begin{pmatrix} 0 & 0 \\ 0 & 11' \end{pmatrix}, \\ s^2 &= \sum y_i^2 - \frac{1}{n_1} \left(\sum y_{1i} \right)^2 - \frac{1}{n_2} \left(\sum y_{2i} \right)^2 \\ &= \sum y_{1i}^2 - n_1 y_{1.}^2 + \sum y_{2i}^2 - n_2 y_{2.}^2 \\ &= \sum (y_{1i} - y_{1.})^2 + \sum (y_{2i} - y_{2.})^2 \\ \Rightarrow F &= \frac{n-2}{1} \frac{\mid M \mid^2}{s^2} \quad \sim F_{1,n-2} \end{aligned}$$

Aus der Theorie folgt also: Die Hypothese wird abgelehnt, falls

$$\begin{aligned} \frac{n_1 n_2}{n_1 + n_2} \frac{(y_{1.} - y_{2.})^2}{\hat{\sigma}^2} &> F_{1,n_1+n_2-2;\alpha} \qquad \text{gilt, oder äquivalent} \\ \frac{\mid y_{1.} - y_{2.} \mid}{\hat{\sigma}} &> \underbrace{\sqrt{\frac{n_1+n_2}{n_1 n_2}} \sqrt{F_{1,n_1+n_2-2;\alpha}}}_{\varphi} \end{aligned}$$

$\Rightarrow$ H wird angenommen, falls $-\varphi \leq \frac{y_{1.} - y_{2.}}{\hat{\sigma}} \leq \varphi$ gilt. Die Verteilung von $\frac{Y_{1.} - Y_{2.}}{\hat{\sigma}}$ gehört zu den Standardverteilungen.

Definition 9.9 (t-Verteilung) *Es sei* $Z \sim \mathsf{N}(a, \sigma^2)$, $U^2 \sim \sigma^2 \chi_m^2$ *unabhängig. Dann heißt* $\frac{Z}{\frac{1}{\sqrt{m}} U}$ *t-verteilt mit* m *Freiheitsgraden.*

$$\frac{Z}{\sqrt{\frac{1}{m} U^2}} \sim t_{m;\frac{a^2}{\sigma^2}}$$

Bei $a = 0$ *schreiben wir* $t_{m;0} = t_m$. *Dies ist die zentrale t-Verteilung. Für* $a \neq 0$ *ist sie nichtzentral.*

Dichte der zentralen t-Verteilung:

$$\begin{aligned} \mathsf{P}\left(\frac{Z}{\sqrt{\frac{1}{m} U^2}} < t \right) &= K_n \int_{-\infty}^{t} \left(1 + \frac{x^2}{n} \right)^{-\frac{n+1}{2}} dx \\ \text{für} \qquad K_n &= \frac{1}{\sqrt{\pi m}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \end{aligned}$$

$\Rightarrow$ H wird angenommen falls

$$-\sqrt{\frac{n_1+n_2}{n_1 n_2}}t_{m;\frac{\alpha}{2}} < \frac{y_{1.}-y_{2.}}{\hat{\sigma}} < \sqrt{\frac{n_1+n_2}{n_1 n_2}}t_{m;\frac{\alpha}{2}}$$

Was macht man bei mehr als zwei Populationen (Verteilungen)?
Seien die folgenden unabhängigen Variablen gegeben:

$$\begin{array}{l} Y_{11}, \quad \ldots \quad , Y_{1n_1} \\ \qquad \vdots \\ Y_{k1}, \quad \ldots \quad , Y_{kn_k} \end{array} \qquad \text{mit} \quad Y_{ij} \sim \mathsf{N}(\mu_i, \sigma^2).$$

Für $H: \quad \mu_1 = \ldots = \mu_k \quad$ gegen $\quad K:$ nicht H
können wir H äquivalent umschreiben:

$$\begin{aligned} H: \mu_1 - \mu_. &= 0 \\ &\vdots \\ \mu_k - \mu_. &= 0 \qquad \text{bzw.} \\ G\beta = 0 &= \begin{pmatrix} k-1 & \ldots & -1 \\ \vdots & \ddots & \\ -1 & \ldots & k-1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} \end{aligned}$$

$$\text{Sei } X = \begin{pmatrix} 1_{n_1} & & 0 \\ & \ddots & \\ 0 & & 1_{n_k} \end{pmatrix} \qquad K = \begin{pmatrix} \frac{1}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n_{n_k}} \end{pmatrix}$$

$\Rightarrow \quad rgX = k, rgG = k-1$. Als Test ergibt sich dann:

Satz 9.6 *Für H gegen K ist der Test*

$$\begin{aligned} \varphi(y) &= \begin{cases} 1 & \text{für} \quad F > F_{k-1,n-k;\alpha} \\ 0 & \text{sonst} \end{cases} \\ \text{für} \quad F &= \frac{1}{k-1}\sum_{i=1}^{k} \frac{n_i}{\hat{\sigma}^2}(y_{i.}-y_{..})^2 \end{aligned}$$

ein bester invarianter α-Test.

Beim Testen ist die Interpretation einfach für $k = 2$. Dann sind die Mittelwerte voneinander verschieden. Bei $k \geq 3$ ist das schwieriger. Bei Ablehnung der Hypothese ist nicht zusehen, welche Mittelwertsdifferenzen voneinander verschieden sind und welche nicht. Wir haben hier zusammengesetzte Hypothesen.

Definition 9.10 *Tests, die im Falle der Ablehnung der Hypothese Aussagen darüber machen, welche Teilhypothesen abzulehnen sind, heißen multiple Tests.*

Hervorhebenswerte multiple Tests sind der multiple t-Test, der Scheffé-Test, der Tukey-Test sowie der Bonferroni-Test. Alle diese Tests reagieren auf unterschiedliche Abweichungen besonders empfindlich.

9.3.1 1. Multipler t–Test

Die älteste und einfachste Methode zum Testen ist die Anwendung des t–Testes für alle Paare $y_{i.}$ und $y_{j.}$. Dabei sind μ_i und μ_j verschieden, falls

$$\mid y_{i.} - y_{j.} \mid > \sqrt{\frac{n_i + n_j}{n_i n_j}} \hat{\sigma} t_{n-k;\frac{\alpha}{2}}$$

gilt. Hierbei ist $\hat{sigma}^2$ die aus den k Gruppen bestimmte Schätzung der Varianz.

Die Irrtumswahrscheinlichkeit 1. Art heißt versuchsbezogenes Risiko. Sie ist die Wahrscheinlichkeit, mindestens eines der Paare μ_i, μ_j als signifikant verschieden zu erkennen, obwohl H richtig ist. Die Wahrscheinlichkeit α heißt vergleichsbezogenes Risiko. Es gibt im allgemeinen keine Beziehung zwischen beiden Risiken. Es zeigt sich, daß das versuchsbezogene Risiko um ein Vielfaches größer als α sein kann. So ist zum Beispiel bei $\alpha = .05$ das versuchsbezogene Risiko bei $k = 3, 4,$ bzw. 5 gleich 0.12, 0.20, 0.30

Beispiel 9.7 Bei den Grafitformteilen war $n = 32,\ k = 8,\ n_i = 4.$
Wir hatten die Grenzdifferenz

$$\sqrt{\frac{n_i + n_j}{n_i n_j}} \hat{\sigma} \underbrace{\mid t \mid_{n-k;0.05}}_{t_{n-k;0.025}} = \sqrt{\frac{l}{n_i}} \hat{\sigma} \mid t \mid_{24;0.05}$$

$$= \sqrt{\frac{1}{2}} \cdot 1.81 \cdot 2.06 = 2.64$$

$\Rightarrow$ $y_{5.} - y_{8.},\ y_{5.} - y_{3.},\ y_{5.} - y_{1.},\ y_{5.} - y_{4.},\ y_{5.} - y_{2.},\ y_{7.} - y_{6.},\ y_{5.} - y_{6.}$
$y_{7.} - y_{1.},\ y_{7.} - y_{4.},\ y_{7.} - y_{2.},\ y_{8.} - y_{2.}$ sind kleiner als 2.64.
Die Wahrscheinlichkeit, daß sich unter den auf diese Weise als signifikant ausgewählten Differenzen fälschlicherweise einige befinden, die aus den gleichen Verteilungen stammen ist sehr hoch.
$\Rightarrow$ Wegen des hohen versuchsbezogenen Risikos ist diese Aussage nicht sehr zuverlässig. Das liegt an der teilweisen Addition der falschen Ablehnungsgebiete der Teilhypothesen.

9.3.2 Der Scheffé–Test

Bei den simultanen Konfidenzintervallen hatten wir für schätzbares $F\beta$ im linearen Modell $Y = X\beta + \varepsilon$ ein simultanes Konfidenzintervall für $c'F\beta$ erhalten:
$\mathsf{P}\Big(c'F\hat{\beta} - \hat{\sigma}\,[c'F(X'W^{-1}X)^{-}F'c\varkappa F_{\varkappa,n-q;\alpha}]^{\frac{1}{2}} \leq c'F\beta \leq$

$$c'F\hat{\beta} + \hat{\sigma}\,[c'F(X'W^{-1}X)^{-}F'c\varkappa F_{\varkappa,n-q;\alpha}]^{\frac{1}{2}} \quad \text{für alle } c\Big) = 1-\alpha, \qquad \varkappa = rgF$$

Wir wollen alle Mittelwertdifferenzen $\mu_i - \mu_j$ schätzen. Die Matrix F, die zu allen diesen Kontrasten gehört hat den Rang $k-1$.

$$\text{bei } k=3: \quad F = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}$$

$\Rightarrow$ Für alle Kontraste haben wir ein simultanes Konfidenzintervall. Es ist:

$$W = I; \qquad X = \begin{pmatrix} 1_{n_1} & & 0 \\ & \ddots & \\ 0 & & 1_{n_k} \end{pmatrix} \quad \Rightarrow (X'X)^{-1} = \begin{pmatrix} \frac{1}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n_k} \end{pmatrix}$$

$c'F = k$-Vektor mit einer 1, einer -1, sonst 0, falls c nur Nullen außer an einer Stelle eine Eins enthält

$$\Rightarrow \; c'F(X'X)^{-1}F'c = \frac{1}{n_i} + \frac{1}{n_j} \quad \text{für } 1,-1 \text{ an } i\text{-ter und } j\text{-ter Stelle}$$
$$= \frac{n_i + n_j}{n_i n_j}$$

$$\Rightarrow \; \mathsf{P}\Bigg(y_{i.} - y_{j.} - \hat{\sigma}\left[\tfrac{n_i+n_j}{n_i n_j}(k-1)F_{k-1,n-q;\alpha}\right]^{\frac{1}{2}} \le \mu_i - \mu_j \le$$
$$y_{i.} - y_{j.} + \hat{\sigma}\left[\tfrac{n_i+n_j}{n_i n_j}(k-1)F_{k-1,n-q;\alpha}\right]^{\frac{1}{2}} \quad \text{für alle } i,j\Bigg) \ge 1-\alpha$$

Das bedeutet: Unter H gelten H_{ij} , d.h. $\mu_i = \mu_j \;\; \forall ij$. Folglich

$$\mathsf{P}\Bigg(0 \in \underbrace{y_{i.} - y_{j.} \pm \hat{\sigma}\sqrt{\frac{n_i+n_j}{n_i n_j}}\sqrt{(k-1)F_{k-1,n-q;\alpha}}}_{\text{Intervall } I_{ij}} \;\; \forall ij\Bigg) \ge 1-\alpha$$

$\Rightarrow$ Lehnt man H_{ij} ab, falls $0 \notin I_{ij}$, so ist die Irrtumswk. 1. Art $\le \alpha$, denn unter H gilt:
$\mathsf{P}\big(0 \in I_{ij} \;\forall ij\big) \ge 1-\alpha$.
Dieser Test, bei dem man H ablehnt falls ein H_{ij} abgelehnt wird, heißt Scheffé-Test.

Beispiel 9.8 Im Beispiel der Grafitformteilen wird H_{ij} abgelehnt, falls

$$|\, y_{i.} - y_{j.} \,| \ge \hat{\sigma}\sqrt{\tfrac{1}{2}}\sqrt{7}\sqrt{F_{7,24;\alpha}} = 1.81 \cdot 1.97 \cdot \sqrt{2.4} = 5.24 \;\text{ ist.}$$

Im Vergleich zum multiplen t-Test erhalten wir:
Es wird nur H_{25} abgelehnt.

9.3.3 Der Tukey–Test

Die simultanen Konfidenzintervalle für die Mittelwertsdifferenzen sind konservativ. Für $n_1 = \ldots = n_k$ bedeutet das beim Scheffé-Test:

$$\mathsf{P}\left(0 \in \left[y_{i.} - y_{j.} - \hat{\sigma}\frac{\sqrt{2}}{\sqrt{n_1}}\sqrt{(k-1)F_{k-1,n-k;\alpha}}\ ,\ \ y_{i.} - y_{j.} + \hat{\sigma}\frac{\sqrt{2}}{\sqrt{n_1}}\sqrt{(k-1)F_{k-1,n-k;\alpha}}\right]\right) > 1-\alpha$$

Die Länge des Intervalls ist: $2\frac{\sqrt{2}}{\sqrt{n_1}}\sqrt{(k-1)F_{k-1,n-k;\alpha}}\hat{\sigma}$.
Eine Möglichkeit, zu einem α-Test zu kommen, besteht in der Benutzung des Intervalls

$$\left[y_{i.} - y_{j.} - \hat{\sigma}\frac{\sqrt{2}}{\sqrt{n_1}}c_\alpha, y_{i.} - y_{j.} + \hat{\sigma}\frac{\sqrt{2}}{\sqrt{n_1}}c_\alpha\right] \qquad \text{für } c_\alpha < \sqrt{(k-1)F_{k-1,n-k;\alpha}}.$$

c_α hängt dann von n_i und k ab.

Wir wissen

$$\begin{aligned} & \sqrt{n_1}\frac{|y_{i.} - y_{j.} - (\mu_i - \mu_j)|}{\hat{\sigma}} && \leq \sqrt{2}c_\alpha \quad \forall ij \\ \Leftrightarrow & \\ & \max_{i,j}\sqrt{n_1}\frac{|y_{i.} - y_{j.} - (\mu_i - \mu_j)|}{\hat{\sigma}} && \leq \sqrt{2}c_\alpha \qquad (1) \end{aligned}$$

Ein simultanes Konfidenzintervall für die $\mu_i - \mu_j$ ist gegeben durch

$$\left[y_{i.} - y_{j.} - \frac{\hat{\sigma}}{\sqrt{n_1}}\sqrt{2}c_\alpha\ ,\ \ y_{i.} - y_{j.} + \frac{\hat{\sigma}}{\sqrt{n_1}}\sqrt{2}c_\alpha\right]$$

wenn (1) erfüllt ist. Nun ist

$$\begin{aligned} & Y_{i.} - Y_{j.} - (\mu_i - \mu_j) && \sim \mathsf{N}(0, 2\cdot\frac{1}{n_1}\sigma^2) \\ \Rightarrow\ & \frac{\sqrt{n_1}}{\sqrt{2}}\left(Y_{i.} - Y_{j.} - (\mu_i - \mu_j)\right) && \sim \mathsf{N}(0,\sigma^2) \\ \Rightarrow\ & \max_{i,j}\frac{\sqrt{n_1}|y_{i.} - y_{j.} - (\mu_i - \mu_j)|}{\sqrt{2}\quad\hat{\sigma}} && \leq c_\alpha\ , \end{aligned}$$

falls c_α der α-Signifikanzpunkt einer Zufallsgröße, die man sich als Spannweite von k unabhängigen t-verteilten Variablen vorstellen kann.

$\Rightarrow$ Bezeichnung: Methode der studentisierten Spannweite (Variationsbreite).

Verfahren: H_{ij} wird abgelehnt, falls gilt:

$$\frac{\sqrt{n_1}\ |\ y_{i.} - y_{j.}\ |}{\hat{\sigma}} > \sqrt{2}c_\alpha$$

In der Tabelle ist $q_{k,n-k;\alpha} = \sqrt{2}c_\alpha$ $(n-k=\nu)$
Dieser Test für H gegen K heißt Tukey-Test.

Beispiel 9.9 Bei den Grafitformteilen haben wir $n_i = 4,\ n = 32,\ k = 8$

$\frac{\hat{\sigma}}{\sqrt{n_i}}q_{8,24;0.05} = 1.81 \cdot 0.5 \cdot 4.68 = 4.24$

Signifikante Unterschiede ergeben sich für: $H_{56},\ H_{51},\ H_{54},\ H_{52}$.
$\Rightarrow$ vier Differenzen sind ungleich 0.

9.3.4 Bonferroni-Tests

Eine sehr einfache Methode zur Konstruktion multipler Tests geht auf R.A.Fischer zurück. H ist zusammengesetzt aus den Teilhypothesen $H_{1,2}, ..., H_{k-1,k}$
Umnummerierung zu: $H'_1, ..., H'_L$.
Sei $t_i(y)$ eine geeignete Teststatistik zum Testen von H'_i, und es sei die Verteilung von $t_i(Y)$ unter H'_i bekannt.
Sei $c_{i,\frac{\alpha}{L}}$ ihr $\frac{\alpha}{L}$-Signifikanzpunkt.
d.h. $\mathsf{P}_{H'_i}(t_i > c_{i,\frac{\alpha}{L}}) \leq \frac{\alpha}{L}$.
Wegen der Bonferroni-Ungleichung

$$\mathsf{P}\left(A_1 \cup \ldots \cup A_L\right) \leq \sum_{l=1}^{L} \mathsf{P}(A_l)$$

erhalten wir mit $A_i = \{t_i > c_{i;\frac{\alpha}{2}}\}$

$$\mathsf{P}_{H'_1 \ldots H'_L}\left(t_i \geq c_{i;\frac{\alpha}{2}} \text{ für mind. ein } i\right) \leq L \cdot \frac{\alpha}{L} = \alpha$$

$\Rightarrow$ Lehnt man H'_i ab gdw. $t_i > c_{i;\frac{\alpha}{2}}$, so hält dieser Bonferroni-Test das Signifikanzniveau unter H ein. Außerdem haben wir: Für jede Kombination $H'_{i_1}, \ldots, H'_{i_m}$ gilt:

$$\mathsf{P}_{H'_{i_1},\ldots,H'_{i_m}}\left(t_{i_j} > c_{i_j;\frac{\alpha}{2}} \text{ für mind. ein } i_j \ , \quad j = 1, \ldots, m\right) \leq m\frac{\alpha}{L} < \alpha$$

$\Rightarrow$ Die Wk aus einer beliebigen Gruppe gültiger Teilhypothesen eine oder mehrere irrtümlich abzulehnen ist immer kleiner als α. Alle diese Aussagen gelten ohne Rücksicht auf gegenseitige Abhängigkeit von $t_1, \ldots, t_L$. Der Test ist einfach, aber er erkennt auch nur wenige Differenzen.

Beispiel 9.10 Bei den Grafitformteilen haben wir

$$\frac{k}{2}(k-1) = L; \quad \text{d.h. } L = 4 \cdot 7 = 28$$

$$| \, y_{i.} - y_{j.} \, | > \sqrt{\frac{n_i + n_j}{n_i n_j}}\hat{\sigma} \underbrace{| \, t \, |_{24;0.0017}}_{3.4} = 4.79$$

$\Rightarrow$ Signifikant abgelehnt werden $H_{52}, \ H_{54}$

Hier wird also sehr selten abgelehnt.
Eine Verbesserung des Bonferroni-Tests (Erhöhung der Trennschärfe) stellt der Bonferroni-Holm-Test da:
H wird angenommen, falls alle H'_l angenommen werden. Die Ablehnung von H erfolgt anders.

1. Diejenige Teilhypothese H'_{l_1} wird abgelehnt, für die gilt: $t_{l_1} > c_{l_1;\frac{\alpha}{L}}$
   Gibt es mehrere, so wird diejenige mit dem grösten t-Wert abgelehnt.

2. Unter den verbleibenden $L-1$ Teilhypothesen wird diejenige abgelehnt, für die gilt: $t_{l_2} > c_{l_2;\frac{\alpha}{L-1}}$ usw.

Sobald im r-ten Schritt kein $t_{j_r} > c_{j_r;\frac{\alpha}{L-r+1}}$ existiert, werden alle Teilhypothesen außer $H'_{l_1}, \ldots, H'_{l_{r-1}}$ angenommen.

Satz 9.7 *Der Bonferroni-Holm-Test ist ein α-Test und ist trennschärfer als der ursprüngliche Bonferroni-Test.*

BEWEIS:

$$\begin{aligned} \mathsf{P}_{H'_1 \ldots H'_L}\big(\underbrace{H'_l \text{ ablehnen für wenigstens ein } l}_{t_l > c_{l;\frac{\alpha}{L}}}\big) &\leq \alpha \\ \mathsf{P}_{H'_{l_1} \ldots H'_{l_m}}\big(\text{mind. ein } H'_{l_i} \text{ wird abgelehnt}\big) &\leq \\ \leq \sum_{j=1}^{m} \mathsf{P}_{H'_{l_1} \ldots H'_{l_m}}\big(t_{l_j} > c_{l_j;\frac{\alpha}{m}}\big) &\leq m\frac{\alpha}{m} = \alpha \end{aligned}$$

Der Bonferroni-Holm-Test lehnt die einzelnen Teilhypothesen häufiger ab als der ursprüngliche Bonferroni-Test.

□

Beispiel 9.11

$$\begin{aligned} H_{rs} : t_i(y) &= \sqrt{\frac{n_r n_s}{n_r + n_s}} \frac{\mid y_{r.} - y_{s.} \mid}{\hat{\sigma}} \\ \text{Verteilung:} &\quad \text{t-Verteilung mit 24 Freiheitsgraden} \\ t_{rs}(y) &= \sqrt{2}\frac{\mid y_{r.} - y_{s.} \mid}{\hat{\sigma}} = 0.7813 \cdot \mid y_{r.} - y_{s.} \mid \end{aligned}$$

Wir haben aus der Tabelle die Werte

$$\begin{aligned} \mid t \mid_{24;0.0008} &= 3.47 \\ \mid t \mid_{24;0.001} &= 2.80 \end{aligned}$$

Es gilt

$$\begin{array}{lllll} H_{25} : 4.28 & H_{45} & : & 3.98 & H_{15} : 3.69 \\ H_{56} : 3.42 & H_{27} & : & 3.10 & H_{35} : 2.95 \end{array}$$

Und genau diese werden abgelehnt.

10 Literatur

Bauer, H.: Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie. Berlin 1968.
Devore, J., R. Peck: Statistics. Belmont 1993.
Fisz, M.: Wahrscheinlichkeitsrechnung und Mathematische Statistik. Berlin 1989.
Humak, K.M.S.: Statistische Methoden der Modellbildung. Berlin 1977.
Krengel, U.: Einführung in die Wahrscheinlichkeitstheorie und Statistik. Vieweg 1991.
Läuter, H., R. Pincus: Mathematisch-statistische Datenanalyse. Berlin 1989.
Lukasz, E.: Characteristic functions. New York 1960.
Pitman, J.: Probability. New York 1993.
Rao, C.R.:Linear Statistical Inference and its applications. Wiley 1973.
Schmetterer, L.: Einführung in die Mathematische Statistik. Wien 1966.
Witting, H.: Mathematische Statistik. Stuttgart, Teubner 1978.