STATISTIK

Notizen zur Vorlesung

Anton Wakolbinger

— Wintersemester 2001/02 —

Inhaltsverzeichnis

1 Kann denn das Zufall sein? 4

1.1 Fishers exakter Test 4

1.2 Von Menschen und Schimpansen 6

1.3 Geherzte Babies 7

1.4 Der Rangsummentest von Wilcoxon 9

2 Empirische Verteilungen, Quantile und QQ-Plots 13

3 Schätzen von Lageparametern 17

3.1 Mittelwert und Median 17

3.2 Schätzen des Medians mit Konfidenz 20

3.3 Ein asymptotisches Konfidenzintervall für den Erwartungswert . 21

4 Mehr über Konfidenzintervalle 22

4.1 Zwei Baupläne für Konfidenzintervalle 22

4.2 Bootstrap-Konfidenzintervalle 23

5 Student, Fisher und die schöne normale Welt 26

5.1 Die Student-Statistik 26

5.2 Die Standard-Normalverteilung auf $\mathbb{R}^n$ und der Satz von Fisher . 27

5.3 Die Student-Statistik nochmal 28

5.4 Die Dichte der Student-Verteilung (Fishers geometrische Herleitung) 29

5.5 Die Dichte der χ^2-Verteilung 30

5.6 Student-Konfidenzintervall und Student-Test 31

5.7 Das Shift-Modell in der schönen normalen Welt 32

6 Gepaarte Stichproben **34**

6.1 Der t-Test für gepaarte Stichproben 34

6.2 Der Wilcoxon-Test für gepaarte Stichproben 35

6.2.1 Das Shiftmodell für symmetrische Verteilungen 36

7 Zur Dualität zwischen Tests und Konfidenzintervallen **38**

7.1 Die Dualitätsbeziehung 38

7.2 Wilcoxon-Test und Hodges-Lehmann-Schätzer 39

8 Passen die Daten zur Hypothese oder zur Alternative? **40**

8.1 Tests à la Neyman-Pearson 40

8.2 Das Lemma von Jerzy Neyman und Egon Pearson 41

9 Wie gut passen die Daten zu einer Hypothese? **43**

9.1 Das Konzept der Likelihood 43

9.2 Zwei Beispiele für Maximum-Likelihood-Schätzer 43

9.3 Die Idee des Maximum-Likelihood-Quotiententests 46

10 Wie gut passen beobachtete Häufigkeiten? **49**

10.1 Feste hypothetische Gewichte: Der Satz von Karl Pearson 49

10.2 Der multivariate zentrale Grenzwertsatz 52

10.2.1 Kovarianzmatrizen 52

10.2.2 Normalverteilungen auf $\mathbb{R}^k$ 53

10.2.3 Verteilungskonvergenz von $\mathbb{R}^k$-wertigen Zufallsvariablen . 54

10.3 Der χ^2-Test 55

10.3.1 χ^2-Test auf Unabhängigkeit zweier Merkmale 57

11 Zur Asymptotik der ML-Schätzer **59**

12 Suffizienz und Vollständigkeit **63**

12.1 Bedingte Verteilung und bedingte Erwartung 63

12.2 Suffizienz 64

12.3 Der Satz von Rao-Blackwell 66

12.4 Vollständigkeit 67

13 Die Cramér-Rao-Ungleichung **70**

14 Das Lageproblem für mehrere Stichproben **72**

14.1 Die einfache Varianzanalyse 72

14.2 Paarweise Vergleiche 77

14.3 Der Kruskal-Wallis-Test 81

15 Das normale lineare Modell **84**

15.1 Schätzen von $\vec{\mu}$ und σ^2 84

15.2 Testen von linearen Hypothesen 86

16 Das lineare Modell mit weißem Rauschen **88**

17 Lineare Regression **90**

17.1 Vom linearen Modell zum linearen Regressionsmodell 90

17.2 Lineares Regressionsmodell mit Weißem Rauschen 95

17.3 Normales lineares Regressionsmodell: Konfidenzellipsoid für β . . 96

17.4 Lineare Regression mit Kontrollvariablen 97

18 Versuchspläne **99**

18.1 Ein Beispiel 99

18.2 Zwei Schätzprobleme 100

18.3 Optimale Versuchspläne 101

18.4 Zur Äquivalenz von D- und G-Optimalität 102

18.5 Polynomiale Regression auf $[-1, 1]$ 106

Kapitel 1

Kann denn das Zufall sein?

1.1 Fishers exakter Test

Wir beginnen mit einem Beispiel, das uns mitten in die Welt des statistischen Testens einer Hypothese führt. Es ist dem empfehlenswerten Buch "Statistics" von D. Freedman et al (2nd ed., W.W. Norton, 1991) entnommen.

Eine Botschaft ein- und desselben Inhalts (es ging um den Vergleich des Erfolgs zweier Therapiemethoden, "Operation" und "Bestrahlung") wurde in zwei unterschiedliche Darstellungsweisen verpackt. Von insgesamt 167 Ärzten, die an einer Sommerschule in Harvard teilnahmen, wurden rein zufällig 80 ausgewählt, denen die Botschaft in der Form A vermittelt wurde, die restlichen 87 bekamen die Botschaft in der Form B mitgeteilt. Jeder der Ärzte entschied sich dann im Lichte der Botschaft für die Bevorzugung einer der beiden Therapiemethoden. Das Ergebnis war:

	für Operation	für Bestrahlung	Summe
A	40	40	80
B	73	14	87
Summe	113	54	167

Beobachtung: in der A-Gruppe gibt es ein wesentlich schlechteres Verhältnis "pro Operation" (nämlich 40 : 40) als in der B-Gruppe (nämlich 73 :14). Gibt es also doch eine Beeinflussung der Entscheidung durch die Form ? Ein Skeptiker könnte sagen: "Ach was, selbst unter der Hypothese, dass keine solche Beeinflussung vorliegt, kann ein derartiges Ergebnis durch den reinen Zufall zustande kommen." Um damit umzugehen, stellen wir erst einmal fest: Angenommen, die Form der Botschaft hat keinen Einfluss auf die Meinungsbildung dieser Ärzte. Die Aufteilung der 80 A-Formulare auf die 113 Befürworter von "Operation" und die 54 Befürworter von "Bestrahlung" wäre dann rein zufällig zustande gekommen. Wie wahrscheinlich ist dann eine so extreme Aufteilung wie die beobachteten 40 : 40 ?

Mathematisch ausgedrückt: Wenn aus einer Urne mit 113 roten und 54 blauen Kugeln rein zufällig 80 Kugeln gezogen werden, wie wahrscheinlich ist dann ein so extremes Ergebnis wie das, nur 40 rote Kugeln zu ziehen ?

Das ist ein Fall für die hypergeometrische Verteilung:

$$N = 113 + 54 = 167, \qquad M = 113, \qquad n = 80,$$

$$\mathrm{Ws}_{N,M,n}\left(\{Z = k\}\right) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

$$\mathbf{E}_{N,M,n} Z = n \cdot \frac{M}{N} \, (= 54,1)$$

Die Wahrscheinlichkeit, ein Ergebnis zu erhalten, das mindestens so weit vom Zentrum weg ist wie das beobachtete, ist

$$\mathrm{Ws}_{N,M,n}(Z \leq 40) + \mathrm{Ws}_{N,M,n}(Z > 68)$$

Das Programm R kennt dies als

```
phyper(40,113,54,80) + phyper(68,113,54,80,lower.tail=FALSE)
```

Als Wert ergibt sich $3,0 \cdot 10^{-6}$.

Dies ist ,,der zu den Daten gehörige p-Wert", oder auch das ,,beobachtete Signifikanzniveau", zu dem wir die Hypothese ablehnen dürfen.

Die Interpretation ist: ,,Angenommen die Hypothese trifft zu. Dann tritt ein Ergebnis, das so extrem ist wie das beobachtete, mit Wahrscheinlichkeit $\sim 3.10^{-6}$, also 3 mal in einer Million, auf".

Was wir beschrieben haben, ist auch als Fishers exakter Text bekannt (nach Sir Ronald A. Fisher, ...)

In R:

```
x <- matrix (c(40,73,40,14), nr=2)
fisher.test(x)
```

(`nr` ist die "number of rows)

Ausgabe:

```
data:x
P-value=2.984e-06
...
odds ratio 0.1938
```

(Die odds ratio ist das beobachtete "Chancenverhältnis" $\frac{40}{40}/\frac{73}{14}$)

Bei dem beschriebenen Beispiel handelt es sich um eine geplante Studie, die mit dem Instrument der Randomisierung arbeitet (rein zufällige Auswahl der 80 Ärzte, die die Form A bekommen).

Nicht immer ist man in dieser schönen Situation. Im nächsten Abschnitt und in den Übungen lernen wir Beispiele von sogenannten Beobachtungsstudien kennen.

Mit einigem Vorbehalt kann man die Frage „Kann das Zufall sein ?" auch guten Gewissens in derartigen Situationen stellen. Zur anregenden Bettlektüre empfohlen seien die ersten beiden Kapitel im oben zitierten Buch von D. Freedman et.al.

1.2 Von Menschen und Schimpansen

Die Ausführungen dieses Abschnitts sind inspiriert von einer Arbeit, bei der Dr. Dirk Metzler (Frankfurt) als Statistiker mitgewirkt hat (I. Ebersberger, D. Metzler, C. Schwarz, S. Pääbo, Genome-wide comparison of DNA sequences between humans and chimpanzees, Am. J. Hum. Genet. 70 (2002), 1490-1497). Es geht darum herauszufinden, ob die Unterschiede in der DNA zwischen Mensch und Schimpansen gleichmässig über die Chromosomen verteilt sind oder nicht.

Aus dem Genom des Schimpansen wurden N DNA-Stücke ($i = 1, \ldots, N$) der Länge L_i (mit L_i zwischen 50 und 500 Basenpaaren) sequenziert,

d.h. es wurden die N entsprechenden Wörter der Länge L_i über dem DNA-Alphabet {A, C, G, T} bestimmt.

Diese wurden dem Wort auf dem jeweils entsprechenden Stück beim Menschen gegenübergestellt.

$M_i \ldots$ # Mutationen auf Stück i (genauer: die Anzahl der Stellen, an denen sich die beiden Wörter unterscheiden)

$\mu_i := M_i/L_i \ldots$ *Mutationsrate* auf Stück i

$c(i) \in \{1, \ldots, 22\} \ldots$ Nummer des menschlichen Chromosoms (genauer: Chromosomenpaars) auf welches Stück i fällt.

(Betrachtet werden nur Stücke, die zu autosomalen Chromosomen, d.h. zu denen außer den Geschlechtschromosomen gehören). Spannende Fragen sind: Wo liegen bevorzugt die Stücke mit hoher Mutationsrate? Warum ist dort die Mutationsrate hoch? Eine bescheidenere Frage ist: Gibt es Chromosomen, auf die eher Stücke mit hoher Mutationsrate fallen? Gibt es in diesem Sinn eine signifikante Variabilitat der Mutationsrate zwischen den Chromosomen?

Formalisieren wir nun diese Frage:

$I_j := \{i | c(i) = j\}$ ... alle zum Chromosom j gehörigen Stücke

$p_j := \frac{\sum_{i \in I_j} M_i}{\sum_{i \in I_j} L_i}$ ... "Mutationsrate auf Chromosom j"

Ist die Variabilität der (p_j) untypisch hoch ?

Ist $p := (p_1, \ldots, p_{22})$ untypisch weit von der Diagonalen entfernt ? Und zwar untypisch in Bezug auf die *Nullhypothese*, dass die Stücke (mit hoher und niedriger Mutationsrate) rein zufällig auf die Chromosomen verteilt sind.

Der euklidische Abstand von p zur Diagonalen ist

$$S(p) := \sqrt{\sum_{j=1}^{22} (p_j - \bar{p})^2} \, .$$

Wir machen uns ein Bild der Verteilung eines *zufälligen* p unter der Nullhypothese. Dazu permutieren wir die $c(i), 1, \ldots, N$, rein zufällig. (Dies entspricht der Vorstellung, dass - unter Respektierung der # I_j - die Stücke rein zufällig auf die Chromosomen verteilt werden). Dadurch entsteht ein zufälliges

$P = (P_1, \ldots P_{22})$. Wir betrachten die Verteilung von $S(P)$. In der Praxis permutieren wir 10000 mal und betrachten die so entstehende "empirische Verteilung".

Es ergab sich, dass keiner der 10000 Werte $S(P^{(1)}), \ldots, S(P^{(10000)})$ größer war als $S(p) \quad (= 0.0045)$.

Wir können also sagen:

Unter der Nullhypothese der rein zufälligen Verteilung der Stücke auf die Chromosomen wird die Wahrscheinlichkeit einer so großen Variabilität der Mutationsrate zwischen den Chromosomen wie der beobachteten auf höchstens 1/10000 geschätzt.

Eine noch kleinere Wahrscheinlichkeit ergibt sich unter der Nullhypothese der rein zufälligen Verteilung aller $M = \sum_{i=1}^{N} M_i$ Mutationsstellen auf die gesamten $L = \sum_{i=1}^{N} L_i$ Stellen. Die Verteilung von $S(P)$ unter der letztgenannten Nullhypothese liegt deutlich links von der Verteilung von $S(P)$ unter der oben diskutierten Nullhypothese der rein zufälligen Verteilung der *Stücke* : das Auseinandernehmen von Stücken mit hoher bzw. niedriger Mutationsrate hat einen glättenden Effekt. Wenn man aber ohnehin davon ausgeht, dass die Mutationsrate von Stück zu Stück unterschiedlich sein kann, ist nicht die rein zufällige Verteilung aller Mutationsstellen, sondern die rein zufällige Verteilung aller Stücke auf die Chromosomen die adäquate Nullhypothese.

1.3 Geherzte Babies

Hat das Hören von (mütterlich klingenden) Herztönen einen Effekt auf die Gewichtszunahme von Neugeborenen ? (vgl. Lee Salk, The Role of the Heartbeat in the Relations between Mother and Infant, Sci.Am., May 1973, p.26-29)

Aus einer Anzahl Neugeborener wurden zufällig einige ausgewählt, die in ihren ersten Lebenstagen ohne Unterbrechung den Ton des Herzschlags eines Erwachsenen hören durften. Die restlichen Babies kamen nicht in diesen Genuss; im übrigen wurden alle Babies gleich behandelt.

Beobachtet wurde die Gewichtszunahme vom ersten bis zum vierten Tag nach der Geburt. Dabei wurde auch nach Gewichtsklasse (leicht, mittel, schwer) unterschieden.

Die (aus dem Diagramm auf p.29 der o.a. Referenz übertragenen) Rohdaten finden Sie unter:
http://www.math.uni-frankfurt.de/ ∼ stoch/wakolbinger/WS2001_2002/Babies.dat

Wir wollen uns auf die mittelschweren Babies konzentrieren und fragen: Wie wahrscheinlich ist eine so viel größere Gewichtszunahme bei den "geherzten" Babies unter der Hypothese, dass das Hören der Herztöne keinen Einfluss hat ?

Wir fragen : Hat die *Behandlung* (Hören der Herztöne) einen *Effekt* (auf die Gewichtszunahme) ?

Gibt es eine signifikante Verschiebung der Gewichtszunahme zwischen der behandelten und der unbehandelten Gruppe ?

Unsere Situation ist die folgende:

Von $m+n$ Individuen werden m rein zufällig ausgewählt, diese bekommen die Behandlung, der Rest ist die "Kontrollgruppe".

Beobachtete Werte:

- $x_1, \ldots, x_m$ in der unbehandelten Gruppe
- $y_1, \ldots, y_n$ in der behandelten Gruppe

Angenommen, die Behandlung hat keinen Effekt. Unter dieser *Nullhypothese* vergleichen wir unsere beobachteten Werte mit dem Ergebnis eines Gedankenexperiments.

Ziehe rein zufällig (ohne Zurücklegen) $Z_1, \ldots, Z_m$ aus der Liste $x_1, \ldots, x_m$, $y_1, \ldots, y_n$.

Berechne die Wahrscheinlichkeit $Ws(\{Z_1 + \ldots + Z_m \leq x_1 + \ldots + x_m\})$. Dies ist i.A. ein mühseliges Abzählproblem.

Zielführender ist eine *Monte-Carlo-Methode*:

Ziehe (z.B.) 10000 mal und notiere die relative Häufigkeit der Erfolge (für das Ereignis $\{Z_1 + \ldots + Z_m \leq x_1 + \ldots + x_m\}$).

In unserem Beispiel der geherzten und ungeherzten Babies geht das mit R so:

```
daten<-read.table("Babies.dat")
```

Aus dem Textfile Babies.dat wird dadurch die $N \times 2$-Matrix *daten*.

```
babies<-split(daten[[1]],daten[[2]])
```

Weil die 2. Spalte von *daten* 6 verschiedene Eintragungen hatte (10,11,20,21,30,31, wobei z.B. 10 für "leicht und unbehandelt" steht), besteht *babies* aus 6 (eben mit diesen Eintragungen indizierten) Datensätzen.

```
b20<-babies$"20"
```

ist der Vektor der Gewichtszunahme der mittelschweren unbehandelten Kinder.

```
b21<-babies$"21"
 b2<-c(b20,b21)


length(b20)
```

gibt die Anzahl (44) der mittelschweren unbehandelten Kinder aus.

```
v<-c(Null)
```

Vor Beginn der Schleife wird der Variablen namens v erst einmal der leere Vektor zugewiesen.

```
for ( i in 1:10000 ) {v[i] <- sum(sample(b2, 44))}
```

10000 mal werden (rein zufällig, ohne zurücklegen) 44 Komponenten aus den $44 + 45$ Komponenten von $b2$ herausgegriffen, aufsummiert und als neue Komponente von v eingetragen.

```
hist(v)
```

gibt ein Histogramm von v aus.

```
sort(v)
```

gibt die aufsteigend sortierten Komponenten von v aus.

Nur einer der 10000 Werke war kleiner als

```
sum(b20).
```

1.4 Der Rangsummentest von Wilcoxon

Die Verteilung der im vorigen Abschnitt besprochenen Teststatistik $Z_1 + \ldots + Z_m$ unter der Nullhypothese hängt nicht nur von m und n, sondern auch von den Werten $x_1, \ldots, x_m$, $y_1, \ldots, y_n$ ab.

Eine elegante Möglichkeit, die Abhängigkeit von den Werten loszuwerden, bietet die sogenannte *Rangsummenstatistik* von Wilcoxon. Prinzip: Ersetze die Werte durch ihre Ränge.

Beispiele

a)

Population	5	4	10	8
Ränge	2	1	4	3

b)

Population	5	4	10	4
Ränge	3	1,5	4	1,5

Wir nehmen zur Vereinfachung an: $x_1, \ldots, x_m$, $y_1, \ldots, y_n$ sind verschieden. Die *beobachtete Rangsumme* der x_i ist $\sum_{i=1}^{m} \text{Rang}(x_i)$, wobei $\text{Rang}(x_i)$ den Rang von x_i in der Gesamtstichprobe $x_1, \ldots, y_n$ angibt.

Bemerkung 1: Unter der Nullhypothese des "rein zufälligen Ziehens" ist

$$(\text{Rang}(Z_1), \ldots, \text{Rang}(Z_m)) \stackrel{\text{d}}{=} (R_1, \ldots, R_m)$$

wobei die R_i rein zufällige Züge aus $\{1, 2, \ldots, m+n\}$ ohne Zurücklegen sind.

$$\mathcal{S} := S_{m,n} := \sum_{i=1}^{m} R_i$$

Was läßt sich über die Verteilung von $\mathcal{S}$ aussagen ?

(Erwartungswert, Varianz; asymptotische Verteilung für große m, n ?)

$$\mathbf{E}R_i = \sum_{k=1}^{m+n} k \frac{1}{m+n} = \frac{(m+n)(m+n+1)}{2} \frac{1}{m+n} = \frac{m+n+1}{2}$$

$$\mathbf{E}\mathcal{S} = m\mathbf{E}R_1 = \frac{m(m+n+1)}{2}$$

Die Varianz von S läßt sich elegant berechnen durch eine schlaue Darstellung (der Verteilung) von $\mathcal{S}$, die dann auch hilfreich ist für Einsicht in die asymptotische Verteilung bei große m, n.

Bemerkung 2: Seien $U_1, \ldots, U_m$, $V_1, \ldots, V_n$ unabhängig und gleichverteilt auf [0,1]. Sei $\text{Rang}(U_i) :=$ Rang von U_i in $\{U_1, \ldots U_m, V_1, \ldots, V_n\}$. Dann ist

$$(R_1, \ldots, R_m) \stackrel{\text{d}}{=} (\text{Rang}(U_1), \ldots, \text{Rang}(U_m))$$

Nun gilt aber:

$$\text{Rang}(U_i) = \sum_{k=1}^{m} I_{\{U_i \geq U_k\}} + \sum_{j=1}^{n} I_{\{U_i > V_j\}} \quad \text{fast sicher}$$

(Dabei ist der erste Summand der rechten Seite die Anzahl der Punkte, die U_i gegen die heimische Mannschaft, d.h. die $U_1, \ldots, U_m$ macht, und der zweite Summand die Anzahl der Punkte, die U_i gegen die gegnerische Mannschaft, d.h. die $V_1, \ldots, V_n\}$ macht.)

Daraus ergibt sich:

$$\tilde{S} := \sum_{i=1}^{m} \text{Rang}(U_i) = \frac{m(m+1)}{2} + \sum_{i=1}^{m} \sum_{j=1}^{n} I_{\{U_i > V_j\}} \text{ fast sicher}$$

$$\mathbf{E}\tilde{S} = \frac{m(m+1)}{2} + \frac{1}{2}mn = m\frac{1}{2}(m+n+1) \qquad \text{(das hatten wir schon!)}$$

$$\begin{aligned}\operatorname{Var}\tilde{S} &= \operatorname{Var}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{\{U_i>V_j\}} \\ &= \sum_{(i,j)}\operatorname{Var}\left(I_{\{U_i>V_j\}}\right) + \sum_{(i_1,j_1)\neq((i_2,j_2)} \operatorname{Cov}(I_{\{U_{i_1}>V_{j_1}\}}, I_{\{U_{i_2}>V_{j_2}\}})\end{aligned}$$

Für $i_1 \neq i_2$ und $j_1 \neq j_2$ sind $I_{\{U_{i_1}>V_{j_1}\}}$ und $I_{\{U_{i_2}>V_{j_2}\}}$ unabhhängig, also $\operatorname{Cov}(\ldots) = 0$.

Für $i_1 = i_2 =: i$ und $j_1 \neq j_2$ ist

$$\operatorname{Cov}(I_{\{U_i>V_{j_1}\}}, I_{\{U_i>V_{j_2}\}}) = \mathbf{E}[I_{\{U_i>V_{j_1}\}} I_{\{U_i>V_{j_2}\}}] - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Insgesamt also:

$$\operatorname{Var}\tilde{S} = mn\frac{1}{4} + mn(n-1)\frac{1}{12} + nm(m-1)\frac{1}{12} = \frac{mn(m+n+1)}{12}$$

Für große m, n gilt für die überwiegende Anzahl von Paaren $((i_1, j_1), (i_2, j_2))$, dass $i_1 \neq i_2$ und $j_1 \neq j_2$ ist, und somit $I_{\{U_{i_1}>V_{j_1}\}}$ unabhängig von $I_{\{U_{i_2}>V_{j_2}\}}$ ist.

Dies macht das folgende Resultat plausibel, welches wir hier nicht beweisen:

Satz 1 (Hoeffding 1949)**.**

Für $m, n \to \infty$ gilt: $\frac{S_{m,n} - \mathbf{E}S_{m,n}}{\sqrt{\operatorname{Var} S_{m,n}}} \to \mathcal{N}(0,1)$ *in Verteilung*

Beispiel: die Baby-Daten aus dem vorigen Abschnitt

```
wilcox.test(b20,b21)
```

Als Output gibt R:

```
data:  b20 and b21
W = 539.5, p-value = 0.0002182
alternative hypothesis: true mu is not equal to 0
```

R gibt den Wert W der sogenannten Mann-Whitney-Statistik an:

$$W = \sum_{i=1}^{n} \text{Rang} x_i - \frac{m(m+1)}{2} \quad (= \sum_{i=1}^{m}\sum_{j=1}^{n} I_{\{x_i>y_j\}} \text{ im Fall keiner Bindungen})$$

Achtung: Der Wilcoxon-Test prüft die Nullhypothese der Gleichheit zweier Verteilungen gegen die *Alternative*, dass die eine gegenüber der anderen verschoben ist. Er ist nicht sensibel gegen andere Arten von Alternativen.

Beispiel: $X_1, \ldots, X_{100}$ unabhängig $\mathcal{N}(0,1)$ - verteilt

$Y_1, \ldots, Y_{100}$ unabhängig $\frac{1}{2}(\mathcal{N}(2,1) + \mathcal{N}(-2,1))$ verteilt

Die Verteilung von Y entsteht durch ein doppelt-stochastisches Experiment: wirf zuerst eine faire Münze und wähle je nach Ergebnis Y entweder aus $\mathcal{N}(2,1)$ oder aus $\mathcal{N}(-2,1)$.

```
x<-rnorm(100)
```

liefert einen Vektor aus 100 standard-normalverteilte Zufallszahlen und weist sie der Variablen x zu

```
y<-rnorm(100)+(4*rbinom(100,1,0.5)-2)
```

Der Befehl 4*rbinom(100,1,0.5)-2 liefert einen Vektor aus 100 unabhängigen, auf $\{-2, +2\}$ gleichverteilten Zufallsvariablen.

Wir machen uns ein Bild von x und y mittels eines Dotplots in dem wir - jeweils leicht verwackelt (jittered) - x über dem Abszissenwert 1 und y über dem Abszissenwert 2 abtragen.

```
b<-c(x,y)
a<- c(rnorm(100,1,0.02), rnorm(100,2,0.02))
plot(a,b, axes=F, xlim=c(0,4), xlab="" , ylab="" )
```

Obwohl x und y definitiv nicht aus demselben Topf kommen, schlägt der Wilcoxon-Test alles andere als Alarm:

```
wilcox.test(x,y)
```

```
W = 5151, p-value = 0.7131
```

Kapitel 2

Empirische Verteilungen, Quantile und QQ-Plots

Mit einer Stichprobe $\mathfrak{x} = (x_1, \ldots, x_m)$ assoziieren wir deren *empirische Verteilung*

$$\nu(\mathfrak{x}) := \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i}.$$

Die empirische Verteilung legt somit in jeden der Punkte x_i die Masse $\frac{1}{m}$; kommt ein Wert in der Stichprobe mit Vielfachheit k vor, so hat dieser Wert die Masse $\frac{k}{m}$.

Verteilungen auf der reellen Achse lassen sich auch durch ihre Verteilungsfunktionen charakterisieren. Zur Erinnerung:
Die zu einer Verteilung ν auf $\mathbb{R}$ gehörige *Verteilungsfunktion* ist

$$F_\nu(q) := \nu((-\infty, q)).$$

Beispiel: $\nu = \text{Binom}(1, 0.5) = \frac{1}{2}(\delta_0 + \delta_1)$:
$F_\nu = 0 \cdot 1_{(-\infty,0]} + 1/2 \cdot 1_{(0,1]} + 1_{(1,\infty)}$.

R bevorzugt bei der Definition der Verteilungsfunktion die "rechtsstetige Variante":

$$p\nu(q) := \nu((-\infty, q])$$

Beispiel: $\nu = \text{Binom}(4, 0.5)$:
Die Auswertung der Funktion $p\nu$ an der Stelle q bekommen wir durch

```
pbinom(q,4,0.5).
```

Der Befehl pbinom(0,4,0.5) liefert die Ausgabe 0.0625, was nichts anderes ist als 1/2^4. (Demgegenüber hat die linksstetige Variante F_ν an der Stelle 0 den Wert 0 - aber immerhin, Sprungstellen und Sprunghöhen von F_ν und $p\nu$ stimmen überein, und das ist die Hauptsache.) Werten wir unsere Verteilungsfunktion, sagen wir, zwischen -1 und 5 an einem feinen Gitter (etwa der Maschenweite 0.01) aus. Den Vektor der Gitterpunkte bekommen wir durch

```
q <- seq(-1,5,0.01)
```

(dies ist eine Folge mit Schrittweite 0.01 zwischen -1 und 5.) Den Plot erhalten wir durch

```
plot(q,pbinom(q,4,0.5), type="s")
```

Der Plottyp s steht hier für (eine spezielle Art von) Stufenfunktion: verbinde zwei Punkte, indem du zuerst horizontal nach rechts und dann vertikal nach oben gehst.

Ist $\mathfrak{X} = (x_1, \ldots, x_m)$ eine Stichprobe mit Werten auf der reellen Achse, dann nennen wir $F_{\mathfrak{X}} := F_{\nu(\mathfrak{X})}$ die *empirische Verteilungsfunktion* von $\mathfrak{X}$.

Im Klartext:

$$F_{\mathfrak{X}} := \frac{1}{m} \cdot \#\{i | x_i < q\}.$$

Machen wir uns ein Bild von einer empirischen Verteilungsfunktion im folgenden einfachen Beispiel einer reellwertigen Stichprobe des Umfangs 4:

```
x <- c(4.2, 5.1, 2.0, 9.4)
```

Zum Plotten der "empirical distribution function" (ecdf) müssen wir zuerst eine Bibliothek aufrufen, die etwas von Stufenfunktionen weiß:

```
library(stepfun)
```

```
plot(ecdf(x))
```

Nicht zuletzt für den Vergleich von Verteilungen au der reellen Achse spielt die *inverse Verteilungsfunktion* (auch *Quantilfunktion* genannt) eine prominente Rolle. Sie ist definiert als

$$q\nu(p) := \sup\{q | F_\nu(q) < p\}, \quad 0 < p \leq 1,$$
$$q\nu(0) := \lim_{p \to 0} q\nu(p) = \inf\{q | F_\nu(q) > 0\}.$$

Beispiel: $\nu = \text{Binom}(1, 0.5)$: $q\nu = 0 \cdot 1_{[0,1/2]} + 1_{(1/2,1]}$.

Im Fall vieler gängiger Verteilungen nimmt uns R die Arbeit der Berechnung der Quantile ab. Überzeugen wir uns erst mal, dass das 0% -Quantil der Binomial(4,0.5)-Verteilung gleich 0 ist. In der Tat erhält man dies als Ausgabe nach dem Befehl

```
qbinom(0,4,0.5)
```

Bei Wert $p = 6.25\%$ springt diese Quantilfunktion von 0 auf 1. Machen wir uns wieder ein Bild! Unser Gitter ist jetzt

```
p <- seq(0,1,0.01)
```

Der Plotbefehl ist

```
plot(p,qbinom(p,4,0.5), type="s")
```

Als nächstes machen wir uns ein Bild von der *empirischen Quantilfunktion* der zu Anfang betrachteten Stichprobe

```
x <- c(4.2, 5.1, 2.0, 9.4)
```

Dies leistet der Befehl

```
i<- 0:4; ; plot(i/4,c(min(x),sort(x)),type="S")
```

Der Vektor $sort(x) = (x_{(i)})_{i=1,\ldots,m}$ ist dabei der Vektor der aufsteigend geordneten x_i.
Im Plottyp S werden zwei Punkte verbunden, indem man zuerst vertikal nach oben und dann horizontal nach rechts geht.

Für die Berechnung der empirischen Quantile hält R auch die Funktion quantile bereit. quantile ist allerding keine Sprungfunktion, sondern es interpoliert die m Punkte $((i-1)/(m-1), x_{(i)})$, $i = 1, \ldots, m$, linear. In der Tat liefern die beiden Befehle

```
i<- 0:3; ; plot(i/3, ; sort(x), ; type="l")
```

und

```
p <- seq(0,1,0.01);plot(p, ; quantile(x,p), ; type="l")
```

im wesentlichen dasselbe Ergebnis.

Wie vergleicht man nun zwei empirische Quantilfunktionen? Wenn die Stichprobenumfänge gleich sind, ist dies ein leichtes: man trägt $sort(x)$ gegen $sort(y)$ ab. Sind die Stichprobenumfänge m und n nicht gleich, so bietet sich lineare Interpolation an: man könnte (etwa für $m < n$) $sort(x)$ gegen $quantile(y, i/(m-1)$ abtragen, mit

```
i<- 0:(m-1).
```

Also etwa für y <- c(3.5, 1, 2, 8, 7):

```
i<- 0:(length(x)-1); plot(sort(x), quantile(y,i/(length(x)-1)))
```

Genau das tut R auch auf Knopfdruck, genauer: auf den Befehl

```
qqplot(x,y)
```

Betrachten wir ein paar QQ-Plots:

```
x<-rexp(100,1)*(2*rbinom(100,1,0.5)-1)
 z<-rnorm(100)
 qqplot(x,z)
```

(Die Dichte der zweiseitigen Standard-Exponentialverteilung ist $\frac{1}{2}e^{-|x|}, x \in \mathbb{R}$).

```
 z<-rnorm(100)
 y<-rcauchy(100)
 qqplot(y,z)
```

(Die Dichte der Standard-Cauchyverteilung ist $\frac{1}{\pi(1+x^2)}, x \in \mathbb{R}$).

Kapitel 3

Schätzen von Lageparametern

3.1 Mittelwert und Median

Wie soll man die "Lage" von Verteilungen beschreiben? Durch den Mittelwert? Oder durch den Median?

Bleiben wir erst einmal bei symmetrischen Verteilungen. Dann fallen Mittelwert und Median mit dem Symmetriezentrum zusammen.

Angenommen, $X_1, X_2, \ldots, X_n$ sind unabhänggig und identisch verteilt mit Verteilungsdichte $g_\theta(x)$, wobei

$g_\theta(\cdot)$ symmetrisch um $\theta \in \mathbb{R}$ ist.

Aufgabe: Schätze den Lageparameter θ aus $\mathfrak{X} = (X_1, \ldots, X_n)$ möglichst gut.

Natürliche Kandidaten für Schätzer sind

$$\text{mean}(\mathfrak{X}) := \frac{1}{n}(X_1 + \ldots + X_n)$$

$$\text{median}(\mathfrak{X}) := \begin{cases} X_{(k+1)}, & \text{für } n = 2k+1 \\ \frac{1}{2}(X_{(k)} + X_{(k+1)}), & \text{für } n = 2k \end{cases}$$

Betrachten wir zwei Beispiele:

1. $g_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\theta}{\sigma})^2}$,
2. $g_\theta(x) = \frac{\alpha}{2} e^{-\alpha|x-\theta|}$.

A ... Stichprobenmittel von 100 unabhängig $\mathcal{N}(0,1)$ verteilter Zufallsvariablen,
B ... Stichprobenmedian von 100 unabhängig $\mathcal{N}(0,1)$ verteilter Zufallsvariablen,
a ... 1000 unabhängige Realisierungen von A,
b ... 1000 unabhängige Realisierungen von B.

Analog für zweiseitige Exp(1)-verteilte Zufallsvariablen: V, W, v, w

Es ergaben sich folgende Verhältnisse

$\frac{\text{Var } a}{\text{Var } b} = 0.618$

$\frac{\text{Var } v}{\text{Var } w} = 1,65$

D.h. im Fall der Normalverteilung streut das Stichprobenmittel weniger, und im Fall der zweiseitigen Exponentialverteilung mehr als der Stichprobenmedian!

Theoretische Überlegungen:

1. Zur Asymptotik von mean:
   Nach dem zentralen Grenzwertsatz gilt (mit $\sigma^2 := \text{Var}(X_i)$):
   $$\sqrt{n}(\text{mean}(\mathfrak{X}) - \theta) \to \mathcal{N}\big(0, \sigma^2\big) \quad \text{in Verteilung.}$$

2. Zur Asymptotik von median:
   Wir geben hier ein heuristisches Argument, das in den Übungen rigoros gemacht wird. Dazu betrachten wir erst einen Spezialfall:
   $U_1, U_2, \ldots$ Uniform(0,1)-verteilt,
   $M_n := \text{median}(U_1, \ldots, U_n)$,
   $K_n :=$ Anzahl der U_i, die in $[0, \frac{1}{2}]$ fallen,
   $\mathcal{L}(K_n) = \text{Binom}(n, \frac{1}{2})$
   Für große n ist $\frac{1}{n}K_n - \frac{1}{2}$ asymptotisch so verteilt wie $M_n - \frac{1}{2}$ (beachte: die Distanz zwischen $U_{(i)}$ und $U_{(i+1)}$ ist im Mittel ungefähr $\frac{1}{n}$)
   Nach Moivre-Laplace gilt also:
   $$\sqrt{n}\left(M_n - \frac{1}{2}\right) \to \mathcal{N}\left(0, \frac{1}{4}\right)$$
   Nun zum allgemeinen Fall: Sei ν eine Verteilung auf $\mathbb{R}$ mit Verteilungsfunktion F. Ist U gleichverteilt auf $[0,1]$, dann ist $X := F^{-1}(U)$ gemäß ν verteilt.
   Wie ist $F^{-1}(M_n)$ asymptotisch verteilt ?
   Die asymptotische Verteilung von M_n wird um
   $$\frac{d}{dq}F^{-1}(q)\bigg|_{q=\frac{1}{2}} = \frac{1}{F'(F^{-1}(\frac{1}{2}))}$$
   gestreckt. Mit $\tilde{\mu} := F^{-1}(\frac{1}{2}) =:$ Median von ν und $f := F' \equiv$ Dichte von ν ergibt sich:
   $$\sqrt{n}(F^{-1}(M_n) - \tilde{\mu}) \overset{n\to\infty}{\longrightarrow} \mathcal{N}\left(0, \frac{1}{4f(\tilde{\mu})^2}\right)$$

Kehren wir zurück zu unserem Beispiel:

Die Varianz der zweiseitigen Standard-Exponentialverteilung ist 2, ihre Dichte bei 0 ist $\frac{1}{2}$.

Die Dichte der Standard-Normalverteilung bei 0 ist $\frac{1}{\sqrt{2\pi}}$

Also:

ν / as.Var.	2seitig Exp	$\mathcal{N}(0,1)$
mean	2	1
median	$\frac{1}{4}2^2 = 1$	$\frac{1}{4} \cdot 2\pi = \frac{\pi}{2}$

Die asymptotischen Verhältnisse $\frac{\text{Var (mean)}}{\text{Var (median)}}$ sind also

- 2 für die zweiseitige Exponentialverteilung,
- $2/\pi \approx 0.637$ für die Normalverteilung.

Wir werden gleich sehen:

Die beiden Beispiele "Normalverteilung" und "Zweiseitige Exponentialverteilung" waren gerade so ausgewählt, dass Stichprobenmittel und Stichprobenmedian jeweils der sogenannte Maximum-Likelihood-Schätzer für den Lageparameter war.

Aufgabe: Bestimme zu den Daten $\mathfrak{x} = (x_1, \ldots, x_n)$ ein $\hat{\theta} = \hat{\theta}(\mathfrak{x})$ so, dass die Dichte $f_{\hat{\theta}}(\mathfrak{x})$ maximal unter allen $f_\theta(\mathfrak{x})$ wird.

Wegen der vorausgesetzten Unabhängigkeit der $X_1, \ldots, X_n$ ist

$$f_\theta(\mathfrak{x}) = \prod_{i=1}^{n} g_\theta(x_i).$$

Wegen der strikten Monotonie des Logarithmus ist die Aufgabe äquivalent zu

$$\sum_{i=1}^{n} \log g_\theta(x_i) = \max!$$

Beispiel 1 (Normalverteilung) $g_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\theta}{\sigma})^2}$

$$\begin{aligned} & \text{const} - \sum_{i=1}^{n} \frac{1}{2}(\frac{x_i - \theta}{\sigma})^2 = \max! \\ \Leftrightarrow \quad & \sum_{i=1}^{n} (x_i - \theta)^2 = \min! \\ \Leftrightarrow \quad & \theta = \frac{1}{n} \sum_{i=1}^{n} x_i \end{aligned}$$

Beispiel 2 (Zweiseitige Exponentialverteilung) $g_\theta = \frac{\alpha}{2} e^{-\alpha|x-\theta|}$

$$
\begin{aligned}
& \text{const} - \sum_{i=1}^{n} \alpha |x_i - \theta| = \max ! \\
\Leftrightarrow \quad & \sum_{i=1}^{n} |x_i - \theta| = \min ! \\
\Leftrightarrow \quad & \text{links von } \theta \text{ liegen genau so viele der } x_i \text{ wie rechts von } \theta \\
\Leftrightarrow \quad & \theta \text{ ist } \textit{ein Median} \text{ von } (x_1, \ldots, x_n)
\end{aligned}
$$

3.2 Schätzen des Medians mit Konfidenz

Sei ν eine beliebige Verteilung auf $\mathbb{R}$.
$\tilde{\mu}$ heißt *ein Median* von ν: $\Longleftrightarrow$

$$\nu\left((-\infty, \tilde{\mu})\right) \leq \frac{1}{2} \quad \text{und} \quad \nu\left((\tilde{\mu}, \infty)\right) \leq \frac{1}{2}$$

Sei $\mathfrak{X} = (X_1 \ldots, X_n)$ mit unabhängigen, gemäß ν verteilten $X_1 \ldots, X_n$.

$\tilde{\mu}(\nu)$ sei ein Median von ν.

Gesucht: Ein Intervall $I = I(\mathfrak{X})$ mit

$$\mathrm{Ws}_\nu\left(\{I(\mathfrak{X}) \ni \tilde{\mu}(\nu)\}\right) \geq 0.95$$

$I(\mathfrak{X})$ heißt dann ein 95% *Konfidenzintervall* für $\tilde{\mu}(\nu)$.

Beispiel:

$$
\begin{aligned}
& n = 10 \\
& I(\mathfrak{X}) := [X_{(1)}, X_{(10)}] = [\min(\mathfrak{X}), \max(\mathfrak{X})] \\
& \mathrm{Ws}_\nu\left(I(\mathfrak{X}) \not\ni \tilde{\mu}(\nu)\right) = \\
& = \mathrm{Ws}_\nu(\text{alle } X_1, \ldots, X_{10} \text{ sind} < \tilde{\mu}) + \mathrm{Ws}_\nu(\text{alle } X_1, \ldots, X_{10} \text{ sind} > \tilde{\mu}) \\
& \leq \left(\frac{1}{2}\right)^{10} + \left(\frac{1}{2}\right)^{10} = \frac{1}{29} = \frac{1}{512} \approx 0.002
\end{aligned}
$$

Und wie wär's mit $I(\mathfrak{X}) := [X_{(2)}, X_{(9)}]$:

$$
\begin{aligned}
& \mathrm{Ws}_\nu\left(I(\mathfrak{X}) \not\ni \tilde{\mu}(\nu)\right) \\
= \quad & \mathrm{Ws}_\nu\left(\text{höchstens eines der } X_i \text{ ist} \leq \tilde{\mu}(\nu)\right) \\
& + \mathrm{Ws}_\nu\left(\text{höchstens eines der } X_i \text{ ist} \geq \tilde{\mu}(\nu)\right) \\
= \quad & 2 \ \mathrm{pbinom}(1, 10, 0.5) \\
= \quad & 2 \cdot \left(\binom{10}{0} + \binom{10}{1}\right) \frac{1}{2^{10}} = 11 \cdot 0.002 = 0.022.
\end{aligned}
$$

Für allgemeines n und Niveau $1 - \alpha$ (statt 0.95):
Bestimme l (möglichst groß) so, dass 2 pbinom $(l, n, 0.5) \leq \alpha$
$\Longleftrightarrow l = \mathrm{qbinom}\ (\alpha/2, n, 0.5)$.

3.3 Ein asymptotisches Konfidenzintervall für den Erwartungswert

Im vorigen Abschnitt haben wir zu jeder Stichprobengröße n ein Konfidenzintervall angegeben, das für den Median *jeder beliebigen* Verteilung ν gut war.

Funktioniert so etwas auch für den Erwartungswert?
M.a.W.: Existiert (für festes n) ein $I = I(\mathfrak{X})$ mit $\quad \mathrm{Ws}_\nu(I(\mathfrak{X}) \ni \mu) \geq 0.95$
für alle Verteilungen ν mit Erwartungswert μ?

Die Antwort ist nein. Sei nämlich ν_0 fest gewählt mit
Erwartungswert μ. $\quad X$ habe Verteilung ν_0.

$$X^\varepsilon := \begin{cases} X & \text{mit Wahrscheinlichkeit } 1-\varepsilon \\ 1/\varepsilon^2 & \text{mit Wahrscheinlichkeit } \varepsilon \end{cases}$$

$$\nu^\varepsilon := \mathcal{L}(X^\varepsilon) \quad , \quad \mu^\varepsilon := \mathbf{E}X^\varepsilon \overset{\varepsilon\to 0}{\longrightarrow} \infty$$

Für jedes von $\mathfrak{X} = (X_1, \ldots, X_n)$ abhängige endliche Intervall $I(\mathfrak{X})$ gilt dann:

$$\mathrm{Ws}\left(I(X_1,\ldots,X_n) = I(X_1^\varepsilon,\ldots,X_n^\varepsilon)\right) \overset{\varepsilon\to 0}{\longrightarrow} 1$$

und $\quad \mathrm{Ws}(I(X_1,\ldots,X_n) \ni \mu_\varepsilon) \overset{\varepsilon\to 0}{\longrightarrow} 0$

somit: $\mathrm{Ws}(I(X_1^\varepsilon,\ldots,X_n^\varepsilon) \ni \mu_\varepsilon) \overset{\varepsilon\to 0}{\longrightarrow} 0,$

also kann I kein Konfidenzintervall für alle Verteilungen mit endlichem Erwartungswert sein.

Es gibt allerdings einen approximativen Ersatz für große n:

Wir wissen: Für jedes ν (mit Erwartungswert μ und Varianz ($\sigma^2 < \infty$) und $X_1, X_2 \ldots$ unabhängig und identisch verteilt gemäß ν gilt nach dem Zentralen Grenzwertsatz

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1) \text{ in Verteilung.}$$

Also gilt für großes n:

$$\mathrm{Ws}_\nu\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \in [\mathrm{qnorm}(0.025), \mathrm{qnorm}(0.975)]\right) \approx 0.95$$

$$\iff \mathrm{Ws}_\nu\left(\bar{X}_n - \mu \in \left[\sigma\frac{1}{\sqrt{n}}\mathrm{qnorm}(0.025), \sigma\frac{1}{\sqrt{n}}\mathrm{qnorm}(0.975)\right]\right) \approx 0.95$$

$$\iff \mathrm{Ws}_\nu\Bigg(\mu \in \underbrace{\left[\bar{X}_n - \sigma\frac{1}{\sqrt{n}}\mathrm{qnorm}(0.975), \bar{X}_n - \sigma\frac{1}{\sqrt{n}}\mathrm{qnorm}(0.025)\right]}_{=: I(\mathfrak{X}_n)}\Bigg) \approx 0.95$$

In diesem Sinn ist $I(\mathfrak{X}_n)$ ein asymptotisches 95%-Konfidenzintervall für μ.

Bemerkung: Ist $\nu = \mathcal{N}(\theta, \sigma^2)$ (mit bekanntem σ^2), so ist $I(\mathfrak{X}_n)$ sogar ein exaktes 95%-Konfidenzintervall für θ.

Kapitel 4

Mehr über Konfidenzintervalle

4.1 Zwei Baupläne für Konfidenzintervalle

V sei eine Familie von Verteilungen
(z.B. alle Verteilungen auf $\mathbb{R}$
oder: alle Normalverteilungen auf $\mathbb{R}$ mit fester Varianz σ^2,
oder: alle verschobenen Verteilungen $\nu := \nu_0(. - \theta)$ mit festem ν_0).

$\theta := m(\nu)$ sei ein reellwertiges Merkmal der Verteilung ν (z.B. Median, Erwartungswert, Korrelationskoeffizient, ...)
$X_1, \ldots, X_n$ unabhängig und identisch verteilt gemäß ν , $\mathfrak{X} := (X_1, \ldots, X_n)$
$L(\mathfrak{X})$ sei ein Schätzer für θ.

Annahme 1. Angenommen wir kennen die Verteilung π von $L(\mathfrak{X}) - \theta$, und diese hängt nicht von θ ab. Dann können wir uns durch den folgenden Ansatz leicht ein (1-α)-Konfidenzintervall für θ beschaffen:

$$\begin{aligned} I(\mathfrak{X}) &:= [L(\mathfrak{X}) - a, L(\mathfrak{X}) - b] \\ I(\mathfrak{X}) \not\ni \theta &\iff L(\mathfrak{X}) - a > \theta \text{ oder } L(\mathfrak{X}) - b < \theta \\ &\iff L(\mathfrak{X}) - \theta > a \text{ oder } L(\mathfrak{X}) - \theta < b \end{aligned}$$

Also ist mit $a := q\pi(1 - \frac{\alpha}{2}), b := q\pi(\frac{\alpha}{2})$ $I(\mathfrak{X})$ ein (1-α)-Konfidenzintervall für θ.

Die Annahme 1 trifft übrigens zu im Beispiel
$\mathbf{V} = \{\mathcal{N}(\theta, \sigma) \,|\theta \in \mathbb{R}\}$ (mit festem σ^2)
denn $\mathcal{L}(\bar{X}_n - \theta) = \mathcal{N}\big(0, \frac{1}{n}\sigma^2\big)$.

Annahme 2. Sei $S(\mathfrak{X})$ irgendeine andere Statistik.
Angenommen, wir kennen die Verteilung ρ von $\frac{L(\mathfrak{X})-\theta}{S(\mathfrak{X})}$,

und diese hängt nicht von θ ab.
Ansatz: (*) $I(\mathfrak{X}) := [L(\mathfrak{X}) - a\mathcal{S}(\mathfrak{X}), L(\mathfrak{X}) - b\mathcal{S}(\mathfrak{X})]$

$$I(\mathfrak{X}) \not\ni \theta \Longleftrightarrow L(\mathfrak{X}) - a\mathcal{S}(\mathfrak{X}) > \theta \text{ oder } L(\mathfrak{X}) - b\mathcal{S}(\mathfrak{X}) < \theta$$
$$\Longleftrightarrow \frac{L(\mathfrak{X}) - \theta}{\mathcal{S}(\mathfrak{X})} > a \text{ oder } \frac{L(\mathfrak{X}) - \theta}{\mathcal{S}(\mathfrak{X})} < b$$

Mit $a := q\rho(1 - \frac{\alpha}{2})$ und $b := q\rho(\frac{\alpha}{2})$
ist somit $I(\mathfrak{X})$ ein $(1 - \alpha)$-Konfidenzintervall für θ.

Wie wir in Abschnitt 14 sehen werden, ist die Annahme 2 erfüllt im Beispiel

$$\mathbf{V} = \{\mathcal{N}(\theta, \sigma) \,|\theta \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$$

mit $L(\mathfrak{X}) := \text{mean}(\mathfrak{X}), \quad S(\mathfrak{X}) := \text{sd}(\mathfrak{X}) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}.$

Dann wird auch klar, warum ein Konfidenzintervall der Form (*) manchmal als *studentisiert* bezeichnet wird.

4.2 Bootstrap-Konfidenzintervalle

Wir gehen aus von der im Beginn des vorigen Abschnitts beschriebenen Situation.

Wie kommt man an die Verteilung von $L(\mathfrak{X}) - m(\nu)$ (bzw. von $\frac{L(\mathfrak{X}) - m(\nu)}{S(\mathfrak{X})}$) heran, wenn man über ν erst mal nichts weiß?

Alles was man in der Hand hat, ist eine Realisierung $\mathfrak{x}$ von $\mathfrak{X} = (X_1, \ldots, X_n)$, und die zugehörige empirische Verteilung $\nu_{\mathfrak{x}}$.

Bootstrap-Idee: Lass $\nu_{\mathfrak{x}}$ die Rolle von ν übernehmen

ν	$\nu_{\mathfrak{x}}$
$\mathfrak{X} = (X_1, \ldots, X_n)$ mit X_i u.i.v. gemäß ν	$\mathfrak{X}^* = (X_1^*, \ldots, X_n^*)$ mit X_i^* u.i.v. gemäß $\nu_{\mathfrak{x}}$

$\mathfrak{X}^*$ entsteht also durch n-maliges Ziehen mit Zurücklegen aus $\mathfrak{x} = (x_1, \ldots, x_n)$.

Hoffnung: $\pi := \mathcal{L}(L(\mathfrak{X}) - m(\nu))$ wird gut approximiert durch

$$\pi_{\mathfrak{x}} := \mathcal{L}(L(\mathfrak{X}^*) - m(\nu_{\mathfrak{x}}))$$

.

In der Praxis wird $\pi_{\mathfrak{x}}$ nicht theoretisch berechnet, sondern durch Monte-Carlo-Simulation angenähert.

Z.B.: Beschaffe eine Realisierung $(\mathfrak{x}_1^*, \ldots, \mathfrak{x}_B^*)$ von B unabhängige Kopien von $\mathfrak{X}^*$ (z.B: $B = 1000$) (indem du jeweils n-mal mit Zurücklegen aus $\mathfrak{x}$ ziehst).

Ersetze die $(1 - \frac{\alpha}{2})$ und $\frac{\alpha}{2}$-Quantile von π durch die entsprechenden empirischen Quantile a^* und b^* von $(L(\mathfrak{x}_1^*) - m(\nu_{\mathfrak{x}}), \ldots, L(\mathfrak{x}_B^*) - m(\nu_{\mathfrak{x}}))$.

$$I(\mathfrak{x}) := [L(\mathfrak{x}) - a^* \,, L(\mathfrak{x}) - b^*]$$

heißt dann *basic bootstrap* $(1-\alpha)$-Konfidenzintervall für $m(\nu)$.

Wie funktioniert das in R? Weil dort das Stichprobenziehen bequem ist, kann man's leicht durchführen.

Beispiel: $m(\nu) \ldots$ Mittelwert , $L(\mathfrak{x}) \ldots$ mean$(\mathfrak{x})$, $\mathfrak{x} = (x_1, \ldots, x_n)$

```
v<-c()
for(j in 1:B)
    v[j]<-mean(sample(x,replace=T))
a<-quantile(v-mean(x),1-alpha/2)
b<-quantile(v-mean(x), alpha/2)
I<-c(mean(x)-a, mean(x)-b)
```

Oder auf Knopfdruck mit den beiden R-Routinen boot und boot.ci:

```
library(boot)
bootinfo<-boot(x, function(y,j) mean[y(j)], R=B)
ci<-boot.ci(bootinfo, conf=1-alpha, type = "basic")
```

Beispiel: $\mathfrak{X} = (X_1, \ldots, X_{101})$ ist Realisierung einer Stichprobe aus

$$\nu = \frac{1}{4}\mathcal{N}(-2,1) + \frac{3}{4}\mathcal{N}(2,1)$$

Siehe dazu

www.math.uni-frankfurt/˜stoch/Wakolbinger/Skripte/FP5.11.pdf
(es gibt dazu auch einen Link von der Homepage der Vorlesung).

Wir wenden uns jetzt dem Ansatz 2) aus dem vorigen Abschnitt zu. Die Hoffnung ist nun, dass

$$\rho := \mathcal{L}\left(\frac{L(\mathfrak{X}) - m(\nu)}{S(\mathfrak{X})}\right)$$

gut approximiert wird durch

$$\rho_{\mathfrak{x}} := \mathcal{L}\left(\frac{L(\mathfrak{X}^*) - m(\nu_{\mathfrak{x}})}{S(\mathfrak{X}^*)}\right).$$

Wieder wird $\rho_{\mathfrak{x}}$ durch Monte-Carlo-Simulation angenähert:

Sei $(\mathfrak{x}_1^*, \ldots, \mathfrak{x}_B^*)$ eine Realisierung von B unabhängigen Kopien von $\mathfrak{X}^*$. Ersetze die $(1-\frac{\alpha}{2})$ und $\frac{\alpha}{2}$-Quantile von ρ durch die entsprechenden empirischen Quantile a^* und b^* von

$$((L(\mathfrak{x}_1^*) - m(\nu_{\mathfrak{x}}))/S(\mathfrak{x}_1^*), \ldots, (L(\mathfrak{x}_B^*) - m(\nu_{\mathfrak{x}}))/S(\mathfrak{x}_B^*))$$

$$I(\mathfrak{x}) := [L(\mathfrak{x}) - a^*\mathcal{S}(\mathfrak{x}) \,, L(\mathfrak{x}) - b^*\mathcal{S}(\mathfrak{x})]$$

heißt dann *studentisiertes bootstrap* (1-α)-Konfidenzintervall für $m(\nu)$.

In unserem Beispiel ist ein natürlicher Kandidat für $S(\mathfrak{x})$ die Stichprobenstreuung

$$\mathrm{sd}(\mathfrak{x}) = \sqrt{\frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}$$

Eine 3. Möglichkeit ist das asymptotische $(1-\alpha)$-Konfidenzintervall, also z.B. für $\alpha = 0.05$

$$I(\mathfrak{x}) = \left[\mathrm{mean}(\mathfrak{x}) - 1.96\frac{1}{\sqrt{n}}\ \mathrm{sd}(\mathfrak{x})\,, \mathrm{mean}(\mathfrak{x}) + 1.96\frac{1}{\sqrt{n}}\ \mathrm{sd}(\mathfrak{x})\right]$$

Halten diese Konfidenzintervalle denn auch einigermaßen ihre nominelle Überdeckungswahrscheinlichkeit ein? Wir sehen uns das an unserem Beispiel an, indem wir 1000 Stichproben aus ν ziehen und notieren, wie oft das jeweilige Intervall den wahren Parameter überdeckt. Siehe dazu die o.a. Web-Seite.

Kapitel 5

Student, Fisher und die schöne normale Welt

5.1 Die Student-Statistik

Wir hatten gesehen, wie man sich Kenntnis über die Verteilung von $\frac{L(\mathfrak{X})-m(\nu)}{S(\mathfrak{X})}$ für die Konstruktion eines Konfidenzintervalls für $m(\nu)$ zunutze machen kann.

Wir fassen nun speziell ins Auge:

$$\nu = \mathcal{N}(\mu, \sigma^2), m(\nu) = \mu, L(\mathfrak{X}) := \text{mean}(\mathfrak{X}) =: \bar{X},$$

$$S(\mathfrak{X}) := \text{sd}(\mathfrak{X}) = \sqrt{\frac{1}{n-1}\left((X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2\right)} =: s\ .$$

Die Zufallsvariable $T := \frac{\sqrt{n}\bar{X}-\mu}{s}$ wird dann (zu Ehren von William Gosset, der unter dem Pseudonym "Student" publizierte) als Student-Statistik (oder auch t-Statistik) bezeichnet.

Es ist ein kleines Wunder, dass die Verteilung von T nicht von μ und σ^2 abhängt und sehr schön ausgerechnet werden kann.

Klar wird dies durch die folgende geometrische Interpretation:

$$D := \{(x, \ldots, x)^T \in \mathbb{R}^n | x \in \mathbb{R}\} =: \textit{Diagonale} \text{ von } \mathbb{R}^n$$

(ist 1-dimensionaler Teilraum von $\mathbb{R}^n$)

$(\bar{X}, \ldots, \bar{X}) = \bar{X}(1, \ldots, 1)$ ist die Projektion[1] von $\mathfrak{X}$ auf D
(Symbol: $\mathcal{P}_D\mathfrak{X} = \bar{X}(1, \ldots, 1)$)

In der Tat gilt für alle $\alpha(1, \ldots, 1) \in D$:

$$\langle \mathfrak{X} - \bar{X}(1, \ldots, 1), \alpha(1, \ldots, 1)\rangle = \alpha(\sum_{i=1}^{n} X_i - n\bar{X}) = 0$$

[1] Im folgenden verstehen wir unter *Projektion* stets eine *Orthogonalprojektion*.

Daraus folgt: Die Projektion von $\mathfrak{X}$ auf das orthogonale Komplement $D^\perp$ zu D ist

$$P_{D^\perp}\mathfrak{X} = \mathfrak{X} - P_D\mathfrak{X} = (X_1 - \bar{X}, \ldots, X_n - \bar{X}).$$

Sind $X_1, \ldots, X_n$ unabh. $\mathcal{N}(\mu, \sigma^2)$, so sind $Z_{1,\ldots,}Z_n$

$$(\text{mit } Z_i := \frac{X_i - \mu}{\sigma})$$

unabhängig $\mathcal{N}(0,1)$ verteilt. Mit $\mathfrak{Z} = (Z_{1,} \ldots, Z_n)$ lässt sich die Student-Statistik somit schreiben als

$$T = \frac{\sqrt{n}\bar{Z}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (Z_i - \bar{Z})^2}} = \frac{\sqrt{n}\bar{Z}}{\frac{1}{\sqrt{n-1}}\|\mathcal{P}_{D^\perp}\mathfrak{Z}\|}$$

(mit $\| \ \| :=$ euklidische Länge).

5.2 Die Standard-Normalverteilung auf $\mathbb{R}^n$ und der Satz von Fisher

Definition: $\mathfrak{Z} = (Z_{1,} \ldots, Z_n)^T$ heißt **standard-normalverteilt auf** $\mathbb{R}^n :\Leftrightarrow$ $Z_{1,} \ldots, Z_n$ sind unabhängig, standard-normalverteilt in $\mathbb{R}$.

Die Verteilungsdichte von $\mathfrak{Z}$ ist

$$\varphi_n(z) = \prod_{i=n}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{\|z\|^2}{2}} \ ,$$

sie ist also rotationssymmetrisch!

Wie ist das Normquadrat einer Projektion von $\mathfrak{Z}$ verteilt?
Wie ist das Normquadrat von $\mathfrak{Z}$ selbst verteilt?

Lemma 1: (Über die orthogonale Invarianz der Standard-Normalverteilung)

$\mathfrak{Z}$ sei standard-normalverteilt auf $\mathbb{R}^n$

a) Die $n \times n$ Matrix M sei orthogonal (d.h. $M \cdot M^T = I$). Dann ist $\mathfrak{Y} = M \cdot \mathfrak{Z}$ wieder standard-normalverteilt auf $\mathbb{R}^n$.

b) Zu jeder ON-Basis von $\mathbb{R}^n$ sind die Koordinaten von $\mathfrak{Z}$ wieder standard-normalverteilt.

Beweis:

a) Die Verteilungsdichte $g(\mathfrak{y})$ von $\mathfrak{Y}$ ist

$$g(\mathfrak{y}) = \frac{1}{|\det M|} \varphi_n(M^{-1}\mathfrak{y}).$$

Nun ist aber $|\det M| = 1$ und

$$\begin{aligned}\|M^{-1}\mathfrak{y}\|^2 &= \mathfrak{y}^T(M^{-1})^T M^{-1}\mathfrak{y} \\ &= \mathfrak{y}^T(MM^T)^{-1}\mathfrak{y} = \|\mathfrak{y}\|^2,\end{aligned}$$

also $g(\mathfrak{y}) = \varphi_n(\mathfrak{y})$.

b) $e_1, \ldots, e_n$ sei eine ON-Basis von $\mathbb{R}^n$. Für $\mathfrak{Y} := (e_1^T\mathfrak{Z}, \ldots, e_n^T\mathfrak{Z})^T$ gilt: $\mathfrak{Y} = M\mathfrak{Z}$, mit der orthogonalen Matrix $M = (e_1, \ldots, e_n)^T$. Die Behauptung folgt nun aus a).

Definition: $\chi^2(n) :=$ Verteilung von $Z_1^2 + \ldots + Z_n^2$

Satz von Fisher: $\mathfrak{Z}$ sei standard-normalverteilt in $\mathbb{R}^n$, H sei ein k-dimensionaler Teilraum von $\mathbb{R}^n$.

Dann gilt:

a) $\|\mathcal{P}_H\mathfrak{Z}\|^2$ ist $\chi^2(k)$ verteilt.

b) $\mathcal{P}_H\mathfrak{Z}$ und $\mathcal{P}_{H^\perp}\mathfrak{Z}$ sind unabhängig.

Beweis: $(e_1, \ldots, e_k)$ sei eine ON-Basis von H, die wir zu einer ON-Basis $B = (e_1, \ldots, e_n)$ von $\mathbb{R}^n$ ergänzen.

$W_1, \ldots, W_k, W_{k+1}, \ldots, W_n$ seien die Koordinaten von $\mathfrak{Z}$ in B. Dann gilt:

$$\mathcal{P}_H\mathfrak{Z} = \sum_{i=1}^{k} e_i W_i \quad , \quad \mathcal{P}_{H^\perp}\mathfrak{Z} = \sum_{i=k+1}^{n} e_i W_i \quad , \quad \|\mathcal{P}_H\mathfrak{Z}\|^2 = \sum_{i=1}^{k} W_i^2 .$$

Also folgt die Behauptung sofort aus Lemma 1.

5.3 Die Student-Statistik nochmal

Aus dem Satz von Fisher ergibt sich sofort das
Korollar $X_1, \ldots, X_n$ seien unabhängig und $\mathcal{N}(\mu, \sigma^2)$ verteilt,

$$Z_i := \frac{X_i - \mu}{\sigma} \quad , \quad \mathfrak{Z} = (Z_1, \ldots, Z_n)^T$$

Dann ist

$$\mathcal{P}_D\mathfrak{Z} = \left(\frac{\bar{X} - \mu}{\sigma}\right)(1, \ldots, 1)^T$$

und damit

$$\begin{aligned}\|\mathcal{P}_D\mathfrak{Z}\| &= \sqrt{n}\frac{|\bar{X} - \mu|}{\sigma} \quad , \\ \mathcal{P}_{D^T}\mathfrak{Z} &= \left(\frac{X_i - \bar{X}}{\sigma}\right)_{i=1,\ldots,n} .\end{aligned}$$

Also folgt aus dem Satz von Fisher:

$$\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \quad \text{ist } \mathcal{N}(0,1)\text{-verteilt,}$$
$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \quad \text{ist } \chi^2(n-1)\text{-verteilt}$$

und die beiden Zufallsvariablen sind unabhängig.

Definition: Z sei $\mathcal{N}(0,1)$-verteilt, C sei $\chi^2(k)$-verteilt und unabhängig von Z.

Dann heißt: $T := \frac{Z}{\sqrt{C/k}}$ **Student(k)**-verteilt.

Manchmal bezeichnet man diese Verteilung auch als t-Verteilung mit k Freiheitsgraden. R kennt die Dichte-, Verteilungs- und Quantilfunktion davon als $\mathrm{dt}(x,k)$, $\mathrm{pt}(x,k)$, $\mathrm{qt}(p,k)$.

Fazit: $X_1,\ldots,X_n$ seien unabhängig und $\mathcal{N}(\mu,\sigma^2)$-verteilt. Dann ist

$$T := \frac{\sqrt{n}(\bar{X}-\mu)}{s}$$

Student$(n-1)$-verteilt.

5.4 Die Dichte der Student-Verteilung (Fishers geometrische Herleitung)

Sei $\mathfrak{Z}$ standardnormalverteilt auf $\mathbb{R}^n$,
Betrag und Richtung von $\mathfrak{Z}$ sind dann voneinander unabhängig und
gegeben $\|\mathfrak{Z}\| = r$ ist $\frac{\mathfrak{Z}}{\|\mathfrak{Z}\|}$ gleichverteilt auf der Einheitssphäre.

Es geht uns um die Verteilung von $\frac{\|\mathcal{P}_D\mathfrak{Z}\|}{\|\mathcal{P}_{D^\perp}\mathfrak{Z}\|}$.

Nach dem Strahlensatz ist dieses Längenverhältnis gleich $\frac{\overline{OM}}{\overline{PM}}$, wobei P die Projektion von $\mathfrak{Z}$ auf die Einheitssphäre S_{n-1} in $\mathbb{R}^n$, und M die Projektion von P auf die Gerade D ist.

Aufgabe: Berechne die Verteilung von $\cot\Phi$, wobei P gleichverteilt auf S_{n-1} und Φ der Winkel zwischen $\overline{OP}$ und D ist.

Berechnen wir erst mal die Verteilungsdichte von Φ:
Für $\varphi \in (0,2\pi)$ ist

$$\mathrm{Ws}\,(\Phi \in [\varphi, \varphi+d\varphi])$$

proportional zu $d\varphi\cdot$ Inhalt einer "Hypersphäre" mit Radius $\sin\varphi$.

(Beispiel: $n=3$: Hypersphäre = Kreis
Volumen = Länge ist proportional zum Radius;
$\mathrm{Ws}\,(\Phi \in [\varphi,\varphi+d\varphi])$ ist proportional zu $\sin\varphi$).

Für allgemeines $n \geq 2$:

$$\mathrm{Ws}\,(\Phi \in [\varphi,\varphi+d\varphi]) = c_n\, d\varphi\, \sin^{n-2}\varphi$$

Zur Erinnerung:

X habe Verteilungsdichte g; h sei streng monoton und differenzierbar. Dann hat $Y := h(X)$ die Verteilungsdichte

$$\frac{1}{|h'\left(h^{-1}(y)\right)|}\, g\left(h^{-1}(y)\right)$$

Anwendung auf unseren Fall: $g(\varphi) = c_n \sin^{n-2}\varphi$

$$\begin{aligned} y = h(\varphi) &= \cot\varphi \\ \cot'(\varphi) &= -(1+\cot^2\varphi) \\ \frac{1}{\cot'(\operatorname{arccot} y)} &= -\frac{1}{1+y^2} \end{aligned}$$

Also ist

$$\frac{1}{|h'\left(h^{-1}(y)\right)|}\, g\left(h^{-1}(y)\right) = \frac{1}{1+y^2} c_n \left(\frac{1}{1+\cot^2(\operatorname{arccot} y)}\right)^{\frac{n-2}{2}} = \left(\frac{1}{1+y^2}\right)^{\frac{n}{2}} c_n.$$

Berechnen wir noch die Konstante c_n. Dazu erinnern wir an

$\int_0^{2\pi} \sin^{n-2}\phi\, d\phi =$

$$\begin{aligned} &= \begin{cases} \frac{(n-3)(n-5)\ldots\cdot 1}{(n-2)(n-4)\ldots\cdot 2}\cdot\pi & \text{für } n \text{ gerade} \\ \frac{(n-3)(n-5)\ldots\cdot 1}{(n-2)(n-4)\ldots\cdot 2}\cdot 2 & \text{für } n \text{ ungerade} \end{cases} \\ &= \frac{\Gamma\left(\frac{n-1}{2}\right)\cdot\sqrt{\pi}}{\Gamma\left(\frac{n}{2}\right)} = c_n^{-1} \end{aligned}$$

Also ist die Verteilungsdichte von $\frac{\vec{OM}}{\overline{PM}}$ gleich

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \frac{1}{\sqrt{\pi}} \left(1+y^2\right)^{-\frac{n}{2}} .$$

Die Dichte der Student(n-1)-Verteilung ist gleich der Verteilungsdichte von $\frac{\vec{OM}}{\frac{1}{\sqrt{n-1}}\overline{PM}}$, also gleich

$$\frac{1}{\sqrt{(n-1)\pi}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \left(1+\frac{y^2}{n-1}\right)^{-\frac{n}{2}}$$

5.5 Die Dichte der χ^2-Verteilung

$\mathfrak{Z}$ sei standard-normalverteilt auf $\mathbb{R}^n$

Die Verteilung von $\|\mathfrak{Z}\|$ ist

$$\begin{aligned} \operatorname{Ws}\left(\|\mathfrak{Z}\| \in [r, r+dr]\right) &= c(n) e^{-\frac{r^2}{2}} \quad \operatorname{Vol}\left\{\mathfrak{z} \mid \|\mathfrak{z}\| \in [r, r+dr]\right\} \\ &= c(n) e^{-\frac{r^2}{2}} r^{n-1} dr =: g(r) dr \end{aligned}$$

Lemma
X habe Verteilungsdichte g. Die Abbildung h sei streng monoton und differenzierbar, dann hat $Y := h(X)$ die Verteilungsdichte

$$\frac{1}{|h'\left(h^{-1}(y)\right)|} g\left(h^{-1}(y)\right)$$

Betrachte speziell $h : r \to r^2$.

Die Verteilungsdichte von $Y = \|\mathfrak{Z}\|^2 = h(\|\mathfrak{Z}\|)$ ist

$$\frac{1}{h'(\sqrt{y})} g(\sqrt{y}) = c(n) \frac{1}{2\sqrt{y}} e^{-\frac{y}{2}} \sqrt{y}^{n-1} = \frac{c(n) e^{-\frac{y}{2}}}{2} y^{\frac{n}{2}-1}$$

Zur Erinnerung: Die Verteilung mit Dichte

$$f(y) = \frac{1}{\Gamma(\beta)} y^{\beta-1} e^{-y} \qquad (\beta > 0)$$

heißt Gamma(β). β heißt deren **Formparameter**.

Hat X Dichte g, so hat $Y := \frac{1}{\alpha} X$ Dichte $\alpha g(\alpha y)$.

Die Verteilung mit Dichte

$$f(y) = \frac{1}{\Gamma(\beta)} \alpha^{\beta} y^{\beta-1} e^{-\alpha y}$$

heißt Gamma(β, α). α heißt deren **Skalenparameter**.
(Merke: Gamma$(1, \alpha)$ =Exp(α).)

Beispiel:

$$\beta = \frac{n}{2} \quad , \quad \alpha = \frac{1}{2}$$
$$f(y) = \frac{1}{\Gamma(\frac{n}{2})} \frac{1}{2^{\frac{n}{2}}} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}$$

Fazit: $\chi^2(n) = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$

5.6 Student-Konfidenzintervall und Student-Test

Zur Erinnnerung:

Kennt man die Verteilung π von $\frac{L(\mathfrak{X})-\theta}{S(\mathfrak{X})}$, so kann man aus den Quantilen von π ein Konfidenzintervall für θ konstruieren.

Damit ergibt sich für $X_1, \ldots, X_n$ unabhängig und $\mathcal{N}(\mu, \sigma^2)$-verteilt:

Korollar 1:

$$I_{1-\alpha}^{\text{stud}}(\mathfrak{X}) := \left[\bar{X} - s\frac{1}{\sqrt{n}}\text{qt}\left(1-\frac{\alpha}{2}, n-1\right) \ , \ \bar{X} - s\frac{1}{\sqrt{n}}\text{qt}\left(\frac{\alpha}{2}, n-1\right)\right]$$
$$= \left[\bar{X} - s\frac{1}{\sqrt{n}}\text{qt}\left(1-\frac{\alpha}{2}, n-1\right) \ , \ \bar{X} + s\frac{1}{\sqrt{n}}\text{qt}\left(1-\frac{\alpha}{2}, n-1\right)\right]$$

ist ein $(1-\alpha)$-Konfidenzintervall für μ.

Korollar 2: $T_\theta := \frac{\sqrt{n}(\bar{X}-\theta)}{s}$ liefert eine Teststatistik für die Hypothese

$$H_\theta : \mu = \theta.$$

Fällt der Wert τ von T_θ betragsmäßig groß aus, dann können wir H_θ gegen die Alternative $\mu \neq \theta$ ablehnen zum p-Wert

$$2 \cdot \mathrm{pt}(-|\tau|, n-1)$$

Eine gängige Möglichkeit ist auch die Konstruktion einer **Entscheidungsregel** (statistischer Test im Sinn von J. Neymann und E. Pearson).

Klassischer t-Test mit Signifikanz α: Lehne H_θ ab, falls

$$|T_\theta(\mathfrak{X})| \quad \geq \quad |\mathrm{qt}(\frac{\alpha}{2}, n-1)|$$

Wir stellen fest: $\mathfrak{x}$ liegt im Annahmebereich des Tests auf H_θ

$$\iff \quad I(\mathfrak{X}) \in \left[-\mathrm{qt}(1-\frac{\alpha}{2}, n-1), \mathrm{qt}(1-\frac{\alpha}{2}, n-1)\right]$$
$$\iff \quad I^{\mathrm{stud}}_{1-\alpha}(\mathfrak{X}) \ni \theta$$

Beispiel: $\alpha = 0.05$

k	qt(0.975, k)
1	12.7
2	4.3
3	3.2
5	2.6
11	2.2
18	2.1
60	2.0
∞	1.96

5.7 Das Shift-Modell in der schönen normalen Welt

$X_1, \ldots, X_m$ $\quad \mathcal{N}(\mu_1, \sigma^2)$-verteilt
$Y_1, \ldots, Y_n$ $\quad \mathcal{N}(\mu_2, \sigma^2)$-verteilt
$X_1, X_2, \ldots, Y_n$ unabhängig

Gesucht:
a) Konfidenzintervall für $\mu_1 - \mu_2$
b) Test der Hypothese $\mu_1 = \mu_2$

Betrachte $\quad \bar{X} - \bar{Y} - (\mu_1 - \mu_2)$.

$$\mathcal{L}\left(\bar{X} - \bar{Y} - (\mu_1 - \mu_2)\right) = \mathcal{N}\left(0, \frac{1}{m}\sigma^2 + \frac{1}{n}\sigma^2\right) = \mathcal{N}\left(0, \left(\frac{1}{m} + \frac{1}{n}\right)\sigma^2\right)$$

Wie kann man diese Verteilung "pivotisieren" (so, dass sie sich nicht mehr ändert, wenn man an μ_1, μ_2 und σ^2 dreht)?
(Englisch: "pivot" = "Drehzapfen")

$$\frac{1}{\sqrt{\frac{1}{m}+\frac{1}{n}}\sigma}\left(\bar{X}-\bar{Y}-(\mu_1-\mu_2)\right) \qquad \text{ist } \mathcal{N}(0,1)\text{-verteilt.}$$

Wie bringt man σ weg? Aus dem Satz von Fisher folgt:

$$\frac{1}{\sigma^2}\left[\sum_{i=1}^{m}(X_i-\bar{X})^2+\sum_{j=1}^{n}(Y_i-\bar{Y})^2\right]$$

ist $\chi^2(m-1+n-1)$-verteilt (und unabhängig von $\bar{X}, \bar{Y}$).

Setzen wir

$$s^2_{\mathfrak{X},\mathfrak{Y}} := \frac{1}{m+n-2}\left(\sum_{i=1}^{m}(X_i-\bar{X})^2+\sum_{j=1}^{n}(Y_i-\bar{Y})^2\right).$$

Dann folgt:

$$T := T(\mathfrak{X},\mathfrak{Y}) := \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{1}{m}+\frac{1}{n}}\cdot s_{\mathfrak{X},\mathfrak{Y}}}$$

ist Student$(m+n-2)$-verteilt.

Korollar 1:

$$I^{\text{stud}}_{1-\alpha}(\mathfrak{X},\mathfrak{Y}) := \Bigg[\bar{X}-\bar{Y}-\sqrt{\frac{1}{m}+\frac{1}{n}}s_{\mathfrak{X},\mathfrak{Y}}\ \text{qt}\left(1-\frac{\alpha}{2}\,,\ m+n-2\right),$$
$$\bar{X}-\bar{Y}+\sqrt{\frac{1}{m}+\frac{1}{n}}s_{\mathfrak{X},\mathfrak{Y}}\ \text{qt}\left(1-\frac{\alpha}{2}\,,\ m+n-2\right)\Bigg]$$

ist ein $(1-\alpha)$-Konfidenzintervall für $\mu_1-\mu_2$.

Korollar 2:

$$T_\theta := \frac{\bar{X}-\bar{Y}-\theta}{\sqrt{\frac{1}{m}+\frac{1}{n}}s_{\mathfrak{X},\mathfrak{Y}}}$$

liefert eine Teststatistik für die Hypothese $H_\theta : \mu_1-\mu_2=\theta$.

Klassischer Zwei-Stichproben t-Test mit Signifikanz α:

Lehne H_θ ab, falls

$$|\,T_\theta(\mathfrak{X},\mathfrak{Y})\,| \ \geq |\ \text{qt}\left(\frac{\alpha}{2}\,,m+n-2\right)|$$

Wieder stellen wir fest:

$(\mathfrak{X},\mathfrak{Y})$ liegt im Annahmebereich des Tests auf H_θ

$$\iff T_\theta(\mathfrak{X},\mathfrak{Y}) \in \left[-\ \text{qt}\left(1-\frac{\alpha}{2}\,,n-1\right),\ \text{qt}\left(1-\frac{\alpha}{2}\,,n-1\right)\right]$$
$$\iff I^{\text{stud}}_{1-\alpha}(\mathfrak{X},\mathfrak{Y}) \ni \theta$$

Kapitel 6

Gepaarte Stichproben

6.1 Der t-Test für gepaarte Stichproben

Beispiel: (aus Box-Hunter-Hunter, Statistics for Experimenters, Wiley 1978) Zwei Materialien (A und B) für Schuhsohlen werden betrachtet. Es geht darum, einen Versuch zu planen, um heraus zu bekommen, welches Material sich weniger abnutzt. Man hat 10 Jungs als Testpersonen; sie erklären sich bereit, 2 Monate lang nur die Versuchsschuhe zu tragen.

1. Eine **schlechte** Versuchsplanung:

   Gib 5 davon Schuhe mit Material A, 5 davon Schuhe mit Material B, und miss nach den 2 Monaten die Abnutzung.

   Warum schlecht? Es gibt hampelige und ruhige Jungs. Ein etwaiger "Materialeffekt" ist mit dem Individualeffekt "Strapazierung der Sohlen", wie man sagt, "vermengt".

2. Eine **gute** Versuchsplanung:

   Gib jedem der Jungs eine Sohle A und eine Sohle B.

   So wurde es in der Tat gemacht, es ergaben sich folgende Werte für die Abnutzung:

	A	B
1	13.2	14.0
2	8.2	8.8
3	10.9	11.2
4	14.3	14.2
5	10.7	11.8
6	6.6	6.4
7	9.5	9.8
8	10.8	11.3
9	8.8	9.3
10	13.3	13.6

a) Eine **schlechte** Auswertung:

"Wirf alles wieder in einen Topf"

Der t-Test für 2 Stichproben lässt in der Tat keine signifikante Ablehnung der Hypothese $\mu_A = \mu_B$ zu.

b) Eine **gute** Auswertung:

"Nimm Rücksicht auf die Paarung"

```
plot(x,y)
```

Betrachte die Differenzen $A_i - B_i$.

Im schönen normalen Modell sind diese unabhängig und identisch normalverteilt.

6.2 Der Wilcoxon-Test für gepaarte Stichproben

Zur Erinnerung:

Die Modellannahme beim t-Test für gepaarte Stichproben (Stichwort "Bubenschuhe") war:

$$\begin{aligned} &\mathcal{L}(A_i) = \mathcal{N}\big(\mu_i, \sigma^2\big) \qquad i = 1, \ldots, n \\ &\mathcal{L}(B_i) = \mathcal{N}\big(\mu_i + \theta, \sigma^2\big) \quad i = 1, \ldots, n \\ &A_1, \ldots, B_n \quad \text{unabhängig.} \end{aligned}$$

Dies implizierte:

$$\mathcal{L}(B_i - A_i) = \mathcal{N}\big(\theta, 2\sigma^2\big)$$

und erlaubte die Anwendung des einfachen t-Tests etwa zur Prüfung der Hypothese "$\theta = 0$"

Lockerung der Modellannahme zu

$$\begin{cases} \mathcal{L}(B_i) = \mathcal{L}(A_i + \theta) \\ A_1, \ldots, B_n \quad \text{unabhängig.} \end{cases}$$

Dies impliziert: $X_i := B_i - A_i$, $i = 1, \ldots, n$, sind unabhängig und identisch verteilt, und $\nu := \mathcal{L}(X_i - \theta)$ ist symmetrisch um den Ursprung.

Dies passt somit in den folgenden Exkurs.

6.2.1 Das Shiftmodell für symmetrische Verteilungen

ν sei eine **um den Ursprung symmetrische** Verteilung auf $\mathbb{R}$.

Der Einfachheit halber nehmen wir an:

ν ist **stetig** (d.h. $\nu(\{x\}) = 0$ für alle $z \in \mathbb{R}$).

Modell: $X_1, X_2, \ldots, X_n$ sei u.i.v., für irgendein θ gelte:

$$\mathcal{L}(X_1 - \theta) = \nu.$$

Idee des Wilcoxon-"signed rank"-Tests auf

$$H_\mu : \theta = \mu :$$

Wenn die deutliche Mehrzahl der betragsmäßig größeren Werte $X_i - \mu$ ein positives Vorzeichen hat, wird $\mathcal{L}(X_1)$ eher rechts von μ liegen.

Betrachte dazu die mit dem Vorzeichen von $X_i - \mu$ gewichteten Ränge von $|X_i - \mu|$.

$$W_\mu(\mathfrak{X}) := \sum_{i=n}^{n} 1_{\{X_i > \mu\}} \text{ rank } |X_i - \mu|.$$

Unter $H_\mu : \theta = \mu$ ist W_μ so verteilt wie

$$S := \sum_{i=n}^{n} iV_i$$

wobei V_i ein fairer {0,1}-Münzwurf ist. R kennt die Verteilung von S als "signrank".

Diese ist symmetrisch um $\frac{1}{4}n(n+1)$ (denn $\sum_{i=n}^{n} i(V_i - \frac{1}{2})$ ist symmetrisch verteilt um 0).

Test auf H_μ (gegen die Alternative $\theta \neq \mu$)

Ist $W_\mu(x) < \frac{1}{4}n(n+1)$, dann ergibt sich der p-Wert $2\text{psignrank}(W_\mu(x), n)$.

Ist $W_\mu(x) > \frac{1}{4}n(n+1)$, dann ergibt sich der p-Wert $2\text{Ws}(\{S \geq W_\mu(x)\})$.

Dafür gibt es die R-Routine

```
wilcox.test (x,mu = \mu)
```

Mit vorgegebenem Signifikanzniveau α:

Lehne H_μ ab, falls

$$W_\mu(x) \notin \left[\text{qsignrank}\left(\frac{\alpha}{2}, n\right), \text{qsignrank}\left(1 - \frac{\alpha}{2}, n\right)\right]$$

Wie gehabt, führt dies auf ein $(1-\alpha)$-Konfidenzintervall für θ:

$$I(\mathfrak{x}) := \{\theta : W_\theta(\mathfrak{x}) \in \left[\text{q signrank}\left(\frac{\alpha}{2}, n\right), \text{qsignrank}\left(1 - \frac{\alpha}{2}, n\right)\right]\}$$

R-Routine:

```
wilcox.test (x, conf.int = T , conf.level = 1-\alpha)
```

(Dabei wird auch der Stichprobenpseudomedian

$$\text{median}\left(\frac{1}{2}(x_i + x_j)_{i,j}\right)$$

als Schätzer für den Median m von $\mathcal{L}\left(\frac{1}{2}(X_1 + X_2)\right)$ mitgeliefert. Man beachte, dass $\mathcal{L}\left(\frac{1}{2}(X_1 + X_2) - m\right)$ auch dann symmetrisch ist, wenn ν nicht symmetrisch ist).

Kehren wir zurück zu den gepaarten Stichproben und zu unserem Beispiel der Bubenschuhe. Der Befehl

```
wilcox.test (a,b,paired=T)
```

liefert den p-Wert 0.014.

Der "Vorzeichentest" ist weniger mächtig: er liefert den p-Wert
$2 * \text{pbinom}\ (2, 10, 0.5) = 0.11$.

Kapitel 7

Zur Dualität zwischen Tests und Konfidenzintervallen

7.1 Die Dualitätsbeziehung

Unsere zufälligen Daten sind eine B-wertige Zufallsgröße $\mathfrak{X}$. (B ist der "Beobachtungsraum", "Datenraum", z.B.: $B = \mathbb{R}^n$). Die Verteilung von $\mathfrak{X}$ sei parametrisiert durch θ, $m(\theta)$ sei ein (skalares) Parametermerkmal.

Für jedes $\mu \in \mathbb{R}$ sei A_μ eine Teilmenge von B.

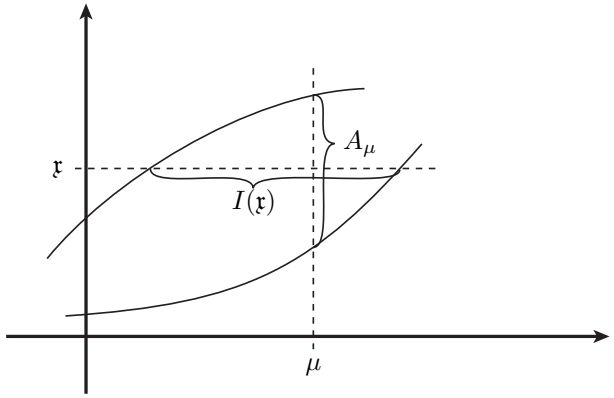
Für jedes $\mathfrak{x} \in B$ sei $I(\mathfrak{x})$ eine Teilmenge von $\mathbb{R}$.

Es gelte die Beziehung:

$$\mathfrak{x} \in A_\mu \iff \mu \in I(\mathfrak{x}).$$

Also stimmen die beiden Ereignisse $\{\mathfrak{X} \in A_{m(\theta)}\}$ und $\{I(\mathfrak{X}) \ni m(\theta)\}$ überein.

Daraus folgt sofort: $I(\mathfrak{X})$ ist (1-α)-Konfidenzbereich für μ
$\iff$ für jedes μ ist A_μ der Annahmebereich eines Tests der Hypothese "$m(\theta) = \mu$" zur Signfikanz α.



7.2 Wilcoxon-Test und Hodges-Lehmann-Schätzer

Der Wilcoxon-Test passt in das folgende **Shiftmodell**:

$$\begin{array}{lll} H_\theta & X_1, \ldots, X_m & \text{u.i.v.} \quad \text{gemäß } \nu(. - \theta) \\ & Y_1, \ldots, Y_n & \text{u.i.v.} \quad \text{gemäß } \nu \\ & (\iff \mathcal{L}(X_1 - \theta) = \mathcal{L}(Y_1)). & \end{array}$$

Unter H_θ ist

$$W_\theta(\mathfrak{X}, \mathfrak{Y}) := \sum_{i=1}^{m} \sum_{j=1}^{n} I_{\{X_i - \theta - Y_j > 0\}} - \frac{mn}{2}$$

Wilxocon (m, n)-verteilt.

Annahmebereich eines Tests auf H_θ mit Signifikanz α:

$$A_\theta := \{(\mathfrak{x}, \mathfrak{y}) | W_\theta(\mathfrak{x}, \mathfrak{y})\} \in \left[\text{qwilcox}\left(\frac{\alpha}{2}, m, n\right),\ \text{qwilcox}\left(1 - \frac{\alpha}{2}, m, n\right)\right]$$

1. Betrachten wir (zu den beobachteten Daten $\mathfrak{x}, \mathfrak{y}$) alle diejenigen Verschiebungen θ, für die $(\mathfrak{x}, \mathfrak{y})$ im Annahmebereich für H_θ liegt:

   $$I(\mathfrak{x}, \mathfrak{y}) := \{\theta | (\mathfrak{x}, \mathfrak{y}) \in A_\theta\}$$

   Beh: $I(\mathfrak{X}, \mathfrak{Y})$ ist ein (1-α)-Konfidenzintervall für θ.

   In der Tat: Für alle θ ist

   $$\text{Ws}_\theta\left(I(\mathfrak{X}, \mathfrak{Y}) \ni \theta\right) = \text{Ws}_\theta\left(I(\mathfrak{X}, \mathfrak{Y}) \in A_\theta\right) = 1 - \alpha.$$

2. Betrachten wir noch (zu den beobachteten Daten $\mathfrak{x}, \mathfrak{y}$) diejenige(n) Verschiebung(en) θ , bei denen $(\mathfrak{x}, \mathfrak{y})$ am deutlichsten für die Annahme von H_θ spricht. D.h. wir suchen dasjenige $\hat{\theta} = \hat{\theta}(\mathfrak{x}, \mathfrak{y})$, für welches $W_{\hat{\theta}}(\mathfrak{x}, \mathfrak{y})$ betragsmäßig am kleinsten (nämlich 0) wird.

   Setzen wir der Einfachheit halber voraus: ν ist **stetig**, d.h. $\nu(\{z\}) = 0$ für alle $z \in \mathbb{R}$. Dann gilt fast sicher:

   $$\sum_{i=1}^{m} \sum_{j=1}^{n} I_{\{X_i - \theta - Y_j > 0\}} = \frac{1}{2} mn \iff$$

   $\iff$ genau die Hälfte der $m \cdot n$ Werte $X_i - Y_j - \theta$ ist > 0.
   $\hat{\theta} := \text{median}\left((X_i - Y_j)_{i=1,..m, j=1,..n}\right)$ heißt **Hodges-Lehmann-Schätzer** für θ.

Kapitel 8

Passen die Daten zur Hypothese oder zur Alternative?

8.1 Tests à la Neyman-Pearson

Zufällige Daten $\mathfrak{X}$ (erzeugt durch einen Zufallsmechanismus mit Parameter ϑ) fallen in den Beobachtungsraum B.

Θ ... Menge der Parameter.

Seien Θ_0 und Θ_1 disjunkte Teilmengen von Θ.

Ein **Test** von Θ_0 gegen Θ_1 (im Sinn von Neyman und Pearson) besteht in einer Partition von B in zwei Teilmengen, A_0 und $A_1 := B \setminus A_0$ (die sogenannten Annahmebereiche für Θ_0 und Θ_1.) Man will die Wahrscheinlichkeit von Fehlentscheidungen unter Kontrolle haben.

Definition: Ein Test (A_0, A_1) von Θ_0 gegen Θ_1 hat Signifikanz $\alpha :\Longleftrightarrow$

$$\forall \vartheta \in \Theta_0 : \mathrm{Ws}_\vartheta(\mathfrak{X} \notin A_0) \leq \alpha$$

$\sup_{\vartheta \in \Theta_0} \mathrm{Ws}_\vartheta(\mathfrak{X} \notin A_0)$ heißt Risiko 1. Art (Produzentenrisiko)

$\sup_{\vartheta \in \Theta_1} \mathrm{Ws}_\vartheta(\mathfrak{X} \notin A_1)$ heißt Risiko 2. Art (Konsumentenrisiko)

Klassisches Beispiel: Annahmekontrolle:

Betrachte eine Lieferung, bestehend aus vielen Einzelstücken. Ein Anteil π davon ist schlecht, der Rest ist gut. Konsument und Produzent einigen sich (z.B.) auf

$$\Theta_0 := \{\pi | \pi \leq 0.03\} \qquad \Theta_1 := \{\pi | \pi \geq 0.15\}$$

Θ_0 entspricht "guter Qualität", Θ_1 entspricht "schlechter Qualität". 20 Stück werden rein zufällig gezogen, die Anzahl der schlechten Stücke wird notiert.

$B := \{0, 1, \ldots, 20\}$

$X := \sharp$ schlechte Stücke in der Stichprobe.

Man einigt sich auf

$A_0 := \{0, 1, 2\}$

$A_1 := \{3, 4, \ldots, 20\}$.

Angenommen, Θ_0 trifft zu. Dann hat $\mathrm{Ws}_\pi(X > 2)$ höchstens (nämlich für $\pi = 0.03$) den Wert

$$1 - \mathrm{Ws}_{0.03}(X \leq 2) = 1 - \text{pbinom } (2, 20, 0.03) = 0.021.$$

Angenommen, Θ_1 trifft zu. Dann hat $\mathrm{Ws}_\pi(X \leq 2)$ höchstens (nämlich für $\pi = 0.15$) den Wert

$$\mathrm{Ws}_{0.15}(X \leq 2) = \text{pbinom } (2, 20, 0.15) = 0.4.$$

Die Abbildung

$$\vartheta \mapsto \mathrm{Ws}_\vartheta(X \in A_0)$$

heißt **Operationscharakteristik** des Tests.

8.2 Das Lemma von Jerzy Neyman und Egon Pearson

Definition: Ein Test von Θ_0 gegen Θ_1 heißt **mächtiger** als ein anderer, wenn er ein kleineres Risiko 2. Art hat.

Ziel: Finde den mächtigsten Test, der ein vorgegebenes Signifikanzniveau α einhält.

Die Situation ist übersichtlich im Fall von einelementigen Θ_0 und Θ_1 (man spricht auch von **einfacher** Hypothese und **einfacher** Alternative).

Lemma von Neyman und Pearson: Seien

$$\Theta_0 = \{\vartheta_0\}, \qquad \Theta_1 = \{\vartheta_1\}\ .$$

f_0 und f_1 seien die Verteilungsdichten (bzw. -gewichte) von $\mathfrak{X}$ unter $\mathrm{Ws}_{\vartheta_0}$ bzw. $\mathrm{Ws}_{\vartheta_1}$. Sei $c > 0$ fest. Wir setzen

$$A(c) := \{\mathfrak{x} \in B | f_0(\mathfrak{x}) \geq c f_1(\mathfrak{x})\}$$

Dann ist unter allen Tests von $\{\vartheta_0\}$ gegen $\{\vartheta_1\}$ mit Signifikanz α

$$\text{der Test mit } A_0 = A(c)$$

der mächtigtste.

Beweis: Sei A ein weiterer zulässiger Annahmebereich für $\{\vartheta_0\}$,
d.h. $\mathrm{Ws}_{\vartheta_0}(X \in A) \geq 1 - \alpha$.

Zu zeigen ist: $\mathrm{Ws}_{\vartheta_1}(X \in A) \geq \mathrm{Ws}_{\vartheta_1}(X \in A(c))$.

In der Tat:

$$\begin{aligned}
\int_A f_1(\mathfrak{x})\, d\mathfrak{x} &= \int_{A(c)} f_1(\mathfrak{x})\, d\mathfrak{x} + \int_{A\setminus A(c)} f_1(\mathfrak{x})\, d\mathfrak{x} - \int_{A(c)\setminus A} f_1(\mathfrak{x})\, d\mathfrak{x} \\
&\geq \int_{A(c)} f_1(\mathfrak{x})\, d\mathfrak{x} + \frac{1}{c}\left(\int_{A\setminus A(c)} f_0(\mathfrak{x})\, d\mathfrak{x} - \int_{A(c)\setminus A} f_0(\mathfrak{x})\, d\mathfrak{x}\right) \\
&= \int_{A(c)} f_1(\mathfrak{x})\, d\mathfrak{x} + \frac{1}{c}\Bigg(\underbrace{\int_A f_0(\mathfrak{x})\, d\mathfrak{x}}_{\geq 1-\alpha} - \underbrace{\int_{A(c)} f_0(\mathfrak{x})\, d\mathfrak{x}}_{=1-\alpha}\Bigg) \\
&\geq \int_{A(c)} f_1(\mathfrak{x})\, d\mathfrak{x}\,. \quad \square
\end{aligned}$$

Zum Merken Die Aufgabenstellung beim Neyman-Pearson-Lemma war: Bestimme A so, dass

$$\begin{aligned}
&(i) \int_A f_0(\mathfrak{x}) d\mathfrak{x} \geq 1-\alpha \\
&(ii) \int_A f_1(\mathfrak{x}) d\mathfrak{x} \quad \text{möglichst klein.}
\end{aligned}$$

Anschauliche Erklärung:

Aufgabenstellung: Kaufe eine Menge von mindestens $1-\alpha$ Einheiten (vielleicht Tonnen) Zucker mit möglichst wenig Geld.

Lösung: Gehe erst in die billigen Läden.

$f_0(\mathfrak{x})\ldots$ im Laden $\mathfrak{x}$ angebotene Zuckermenge

$\frac{f_1(\mathfrak{x})}{f_0(\mathfrak{x})}\ \ldots$ Zuckerpreis im Laden $\mathfrak{x}$

$f_1(\mathfrak{x}) = \frac{f_1(\mathfrak{x})}{f_0(\mathfrak{x})} \cdot f_0(\mathfrak{x})$ ...im Laden gelassenes Geld, wenn man dessen gesamten Zucker kauft

$$A(c) := \left\{\mathfrak{x} \mid \frac{f_0(\mathfrak{x})}{f_1(\mathfrak{x})} \geq c\right\}$$

("Leistungs-Preis-Verhältnis ist mindestens c").

Bestimme das maximale c so, dass

$$\int_{A(c)} f_0(\mathfrak{x})\, d\mathfrak{x} \geq 1-\alpha\,.$$

Kapitel 9

Wie gut passen die Daten zu einer Hypothese?

9.1 Das Konzept der Likelihood

Wir erinnern uns an die Idee des statistischen Tests: Ziehe die Hypothese dann in Zweifel, wenn (verglichen mit einer Alternative) die Daten unter der Hypothese hinreichend unwahrscheinlich werden.

Das "Passen der Daten $\mathfrak{x}$ zur Hypothese H_ϑ" kann man messen durch die sogenannte

$$\textbf{Likelihood} \ : f_\vartheta(\mathfrak{x})$$

wobei wie gehabt $f_\vartheta(\mathfrak{x})$ die Wahrscheinlichkeitsdichte bzw. das Wahrscheinlichkeitsgewicht von $\mathfrak{x}$ unter Ws_ϑ ist.

Ein Parameter $\hat{\vartheta} = \hat{\vartheta}(\mathfrak{X})$, zu dem die zufälligen Daten $\mathfrak{X}$ am besten passen in dem Sinn, dass $\hat{\vartheta}(\mathfrak{X})$ Maximalstelle von

$$\vartheta \mapsto f_\vartheta(\mathfrak{X})$$

ist, heißt **Maximum-Likelihood-Schätzer** von ϑ.

9.2 Zwei Beispiele für Maximum-Likelihood-Schätzer

Beispiel 1:

$\mathfrak{X} = (X_1, \ldots, X_n)$ mit $X_1, \ldots, X_n$ unabhängig, $\mathcal{N}(\mu, \sigma^2)$ verteilt

$$f_\vartheta(\mathfrak{X}) = \frac{1}{\sqrt{2\pi}^n} \, \frac{1}{\sigma^n} \, e^{-\frac{1}{2\sigma^2} \sum (X_i - \mu)^2}.$$

Für jedes σ wird dies maximal, wenn die Summe im Exponenten minimal wird, also bei $\hat{\mu} := \bar{X}$.

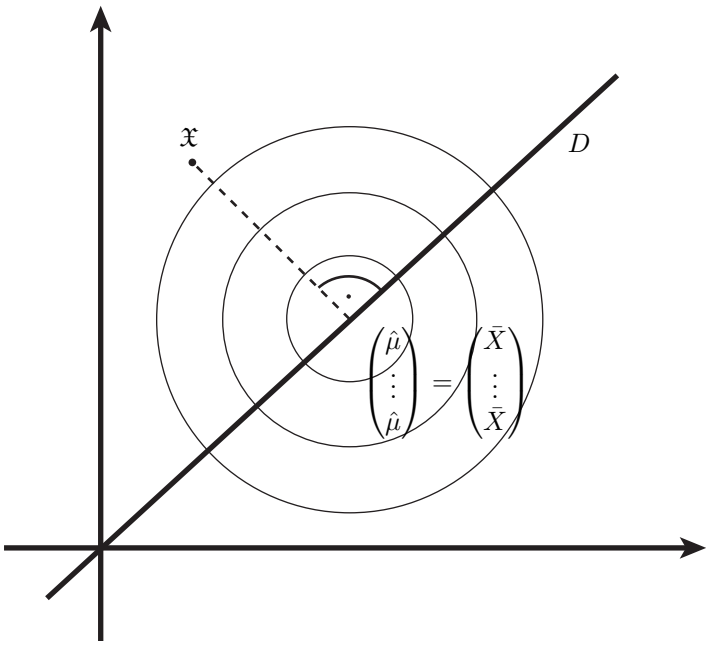


Abbildung 9.1: Wähle denjenigen Punkt auf D, zu dem $\mathfrak{X}$ am nächsten ist.

Bestimme noch σ so, dass

$$\log f_{(\bar{X},\sigma)}(\mathfrak{X}) = \text{const} - n\log\sigma - \frac{\sum(X_i - \bar{X})^2}{2\sigma^3}$$

maximal wird. Durch Nullsetzen der Ableitung erhält man:

$$\frac{1}{\sigma} = \frac{\frac{1}{n}\sum(X_i - \bar{X})^2}{\sigma^2},$$

dies führt auf

$$\widehat{\sigma^2} := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Also ist $(\hat{\mu}, \widehat{\sigma^2})$ der ML-Schätzer für (μ, σ^2).

Bemerkung: Hätten wir mit (μ, σ) anstelle mit (μ, σ^2) parametrisiert, so hätte sich als ML-Schätzer $(\hat{\mu}, \hat{\sigma})$ ergeben mit

$$\hat{\sigma} := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad .$$

Wir stellen fest: $\hat{\sigma^2} = \widehat{\sigma^2}$.

Dies ist keine Koinzidenz, denn man überzeugt sicht leicht: Das Bilden des ML-Schätzers ist verträglich mit Umparametrisieren.

Wir werden später den Schätzer $\widehat{\sigma^2}$ mit dem guten alten

$$s^2 := \mathrm{sd}(\mathfrak{X}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

vergleichen.

Beispiel 2: (Schätzen von Anteilen)

$$s \in \mathbb{N}, \quad p_1 + \ldots + p_s = 1\ ,$$

$(K_1, \ldots, K_s)$ multinomial $(n; p_1, \ldots, p_s)$ verteilt

$$\vartheta = (p_1, \ldots, p_s) =: \vec{p}$$

$$f_{\vec{p}}(\mathcal{K}) = \binom{n}{K_1 \ldots K_s} p_1^{K_1} \ldots p_s^{K_s}\ .$$

Maximiere $\vec{p} \mapsto f_{\vec{p}}(\mathcal{K})$ unter der Nebenbedingung $\sum p_j = 1$ (Lagrange-Methode).

$$\frac{\partial}{\partial p_\ell} \log f_{\vec{p}}(\mathcal{K}) = \frac{\partial}{\partial p_\ell}\left(\sum_{j=1}^{s} K_j \log p_j\right) = \frac{K_\ell}{p_\ell}$$

$$\frac{\partial}{\partial p_\ell}\gamma \sum p_j = \gamma$$

$$\frac{K_\ell}{p_\ell} - \gamma = 0 \quad \Longrightarrow p_\ell = \frac{K_\ell}{\gamma}$$

Aus der Nebenbedingung folgt: $\quad 1 = \frac{1}{\gamma}\sum K_\ell = \frac{n}{\gamma} \quad .$

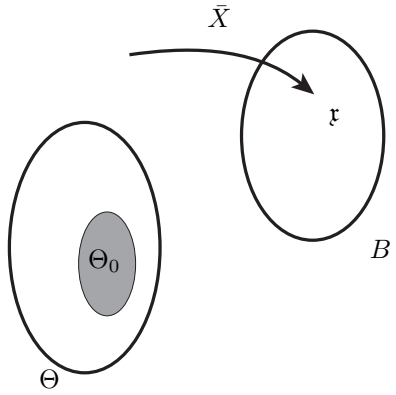
Damit ergibt sich als ML-Schätzer

$$\hat{p_\ell} = \frac{K_\ell}{n}\ .$$

9.3 Die Idee des Maximum-Likelihood-Quotiententests

Sei $\Theta_0 \subseteq \Theta$ gegeben.

Wie gut passt ein $\mathfrak{x} \in B$ zur Hypothese Θ_0?



Eine Maßzahl dafür ist der "ML-Quotient":

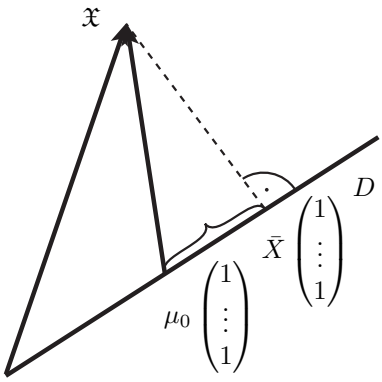
$$Q(\mathfrak{x}) := \frac{\sup\limits_{\vartheta \in \Theta_0} f_\vartheta(\mathfrak{x})}{\sup\limits_{\vartheta \in \Theta} f_\vartheta(\mathfrak{x})} \quad (\leq 1)$$

Beispiel 1: (Normalverteilungen)

$\Theta = \mathbb{R} \times \mathbb{R}_+$, $\Theta_0 := \{\mu_0\} \times \mathbb{R}_+$

$$\arg\max_{\sigma^2} f_{(\mu_0,\sigma^2)}(\mathfrak{X}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2 =: \hat{\sigma}_0^2$$

$$\arg\max_{(\mu,\sigma^2)} f_{(\mu,\sigma^2)}(\mathfrak{X}) = \left(\bar{X}, \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \left(\bar{X}, \hat{\sigma}^2\right)$$



Mit Pythagoras ergibt sich:

$$n(\bar{X}-\mu_0)^2+\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}(X_i-\mu_0)^2$$

$$\Longrightarrow \hat{\sigma}^2=\hat{\sigma}_0^2-(\bar{X}-\mu_0)^2$$
$$\Longrightarrow \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}=1-\frac{(\bar{X}-\mu_0)^2}{\sigma_0^2}$$

$$\begin{aligned}Q(\mathfrak{X}) &= (2\pi\hat{\sigma}_0^2)^{-\frac{n}{2}}\exp\left(-\frac{\sum(X_i-\mu_0)^2}{2\hat{\sigma}_0^2}\right)/(2\pi\hat{\sigma}^2)^{-\frac{n}{2}}\exp\left(-\frac{\sum(X_i-\bar{X})^2}{2\hat{\sigma}^2}\right)\\ &= \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{\frac{n}{2}}\exp\left(-\frac{n}{2}\right)/\exp\left(-\frac{n}{2}\right)\\ &= \left(1-\frac{1}{n}\frac{(\bar{X}-\mu_0)^2}{\hat{\sigma}_0^2/n}\right)^{n/2}\approx\exp\left(-\frac{1}{2}\frac{n(\bar{X}-\mu_0)^2}{\hat{\sigma}_0^2/n}\right)=\exp\left(-\frac{1}{2}\frac{n(\bar{X}-\mu_0)^2}{\sigma^2}\frac{\sigma^2}{\hat{\sigma}_0^2}\right)\end{aligned}$$

Unter der Hypothese $H_0:\mu=\mu_0$ ist $n(\bar{X}-\mu_0)^2/\sigma^2$ $\chi^2(1)$-verteilt. Nach dem Gesetz der Großen Zahlen konvergiert $\hat{\sigma}_0^2$ für $n\to\infty$ gegen σ^2. Also ist $-2\log Q(\mathfrak{X})$ unter H_0 asymptotisch $\chi^2(1)$ verteilt.

Beispiel 2: (Multinomialverteilungen)

$$\Theta:=\{(p_1,\ldots,p_s) \quad | \quad p_j\geq 0, \quad \sum p_j=1\}$$

Wir betrachten eine einfache Hypothese

$$\Theta_0:=\{(\pi_1,\ldots,\pi_s)\} \quad \text{mit festem } \pi_1,\ldots,\pi_s$$

Der ML-Quotient ist

$$Q(\mathcal{K})=\prod_{j=1}^{s}\left(\frac{\pi_j}{\hat{p}_j}\right)^{K_j} \quad ,$$

mit $\hat{p}_j=\frac{K_j}{n}$.

Für $\hat{p}_j$ nahe bei π_j gilt:

$$\begin{aligned}\log Q &= \sum_{j=1}^{s}K_j\log\left(1+\frac{\pi_j-\hat{p}_j}{\hat{p}_j}\right)\\ &\approx \sum_{j=1}^{s}K_j\left(\frac{\pi_j-\hat{p}_j}{\hat{p}_j}-\frac{1}{2}\left(\frac{\pi_j-\hat{p}_j}{\hat{p}_j}\right)^2\right)\\ &= \underbrace{n\sum_{j=1}^{s}(\pi_j-\hat{p}_j)}_{=0\ (\text{wegen}\ \sum\pi_j=\sum\hat{p}_j=1)} - \frac{n}{2}\sum_{j=1}^{s}\frac{(\pi_j-\hat{p}_j)^2}{\hat{p}_j}\end{aligned}$$

Es ergibt sich

$$-2\log Q \approx n\sum_{j=1}^{s}\frac{(\hat{p}_j-\pi_j)^2}{\hat{p}_j}=\sum_{j=1}^{s}\frac{(K_j-n\pi_j)^2}{n\hat{p}_j}=\sum_{j=1}^{s}\frac{(K_j-n\pi_j)^2}{n\pi_j}$$

Definition: $\sum_{j=1}^{s}\frac{(K_j-n\pi_j)^2}{n\pi_j}$

heißt Pearson'sche Teststatistik für die Hypothese $H_\pi : \pi = (\pi_1, \ldots, \pi_s)$.

Wir werden beweisen, dass diese unter H_π für $n \to \infty$ asymptotisch $\chi^2(s-1)$-verteilt ist.

Fazit: Sowohl in Beispiel 1 wie in Beispiel 2 ist $-2\log Q$ für $n \to \infty$ asymptotisch $\chi^2(\dim\Theta - \dim\Theta_0)$-verteilt.

Ein Resultat von Wilks (siehe z.B. P. Bickel, K. Doksum, Mathematical Statistics, Vol. 1 (2nd ed.), Prentice Hall 2001, Abschnitt 6.3.1) besagt, dass dies unter viel allgemeineren Voraussetzungen zutrifft.

Kapitel 10

Wie gut passen beobachtete Häufigkeiten?

10.1 Feste hypothetische Gewichte: Der Satz von Karl Pearson

Satz (Karl Pearson, 1900) Sei $s \in \mathbb{N}$, und seien $\pi_1, \ldots, \pi_s$ Wahrscheinlichkeitsgewichte.

Ist für jedes n der zufällige Vektor $\mathcal{K}^{(n)} := (\mathcal{K}_1^{(n)}, \ldots, K_s^{(n)})$ multinomial $(n; \pi_1, \ldots, \pi_s)$-verteilt, dann gilt

$$\sum_{j=1}^{s} \frac{(K_j - n\pi_j)^2}{n\pi_j} \Longrightarrow \chi^2(s-1) \quad \text{in Verteilung.}$$

Mnemo: Die linke Seite ist

$$\sum \frac{(\text{beobachtete Häufigkeit - erwartete Häufigkeit})^2}{\text{erwartete Häufigkeit}}$$

Erste Bemerkung: Für jedes feste j ist $K_j^{(n)}$ Binom(n, π_j)-verteilt, also folgt mit dem Satz von de Moivre-Laplace:

$$\frac{(K_j^{(n)} - n\pi_j)^2}{n\pi_j(1-\pi_j)} \overset{n\to\infty}{\longrightarrow} \chi^2(1)$$

Fishers heuristischer Beweis des Satzes von Pearson:

Wir brechen die Abhängigkeit auf durch Poissonisieren:

$N_1^{(n)}, \ldots, N_s^{(n)}$ seien unabhängig, mit $N_j^{(n)}$Poisson$(n\pi_j)$-verteilt.

Die Gesamtanzahl $N^{(n)} := N_1^{(n)} + \ldots + N_s^{(n)}$ ist dann Poisson(n)-verteilt.

Bemerkung 1: Bedingt unter $N^{(n)} = n$ ist $N_1^{(n)}, \ldots, N_s^{(n)}$ multinomial $(n; \pi_1, \ldots, \pi_s)$-verteilt. In der Tat:

$$\begin{aligned}\frac{\mathrm{Ws}(N_1^{(n)} = k_1, \ldots, N_s^{(n)} = k_s)}{\mathrm{Ws}(N^{(n)} = n)} &= \frac{e^{-n\pi_1}\frac{(n\pi_1)^{k_1}}{k_1!} \ldots e^{-n\pi_s}\frac{(n\pi_s)^{k_s}}{k_s!}}{e^{-n}\frac{n^n}{n!}} \\ &= \frac{n!}{k_1! \ldots k_s!}\, \pi_1^{k_1} \ldots \pi_s^{k_s} \,. \qquad \square\end{aligned}$$

$N^{(n)}$ läßt sich darstellen als Summe von n unabhängigen Poisson(π_j)-verteilten Zufallsvariablen. Die Varianz von Poisson(α) ist α.

Nach dem klassischen Zentralen Grenzwertsatz gilt für die Folge $N_j^{(n)}, n = 1, 2, \ldots$ (mit festem j):

$$\frac{N_j^{(n)} - n\pi_j}{\sqrt{n\pi_j}} \longrightarrow \mathcal{N}(0,1) \quad \text{in Verteilung.}$$

Diese Komponenten sind unabhängig. Also folgt:

$$\mathcal{Y}_n := \left(\frac{N_j^{(n)} - n\pi_j}{\sqrt{n\pi_j}}\right)_{j=1,\ldots,s} \overset{n\to\infty}{\longrightarrow} \mathfrak{Z} \quad \text{in Verteilung.}$$

mit $\mathfrak{Z}$ standard normalverteilt auf $\mathbb{R}^s$.

Wir versuchen nun, durch eine geeignete Projektion

$$\text{von} \quad \frac{\left(N_j^{(n)} - n\pi_j\right)}{\sqrt{n\pi_j}} \quad \text{zu} \quad \frac{\left(N_j^{(n)} - N^{(n)}\pi_j\right)}{\sqrt{n\pi_j}}$$

überzugehen, mit $N^{(n)} := \sum_{j=1}^s N_j^{(n)}$.

(Mit anderen Worten: wir versuchen, den einen Freiheitsgrad "zufällige Gesamtanzahl" wegzuprojezieren.)

Die Differenz der beiden Vektoren ist

$$\left(\sqrt{\pi_j}\frac{1}{\sqrt{n}}\,(N^{(n)} - n)\right).$$

Dieser Vektor ist nichts anderes als die Projektion von $\mathcal{Y}_n$ auf die durch den Einheitsvektor

$$e := \begin{pmatrix}\sqrt{\pi_1} \\ \ldots \\ \sqrt{\pi_s}\end{pmatrix}$$

gegebene Richtung. In der Tat ist

$$e^T \cdot \mathcal{Y}_n = \sum_{j=1}^s \sqrt{\pi_j}\frac{N_j^{(n)} - n\pi_j}{\sqrt{n\pi_j}} = \frac{1}{\sqrt{n}}\sum_{j=1}^s \left(N_j^{(n)} - n\pi_j\right) = \frac{1}{\sqrt{n}}(N^{(n)} - n) \,.$$

Wir fassen zusammen:

$$\left(\sqrt{\pi_j}\frac{1}{\sqrt{n}}(N^{(n)}-n)\right)_j = ee^T\mathcal{Y}_n \overset{n\to\infty}{\longrightarrow} ee^T\mathfrak{Z} \quad \text{in Verteilung}$$

$$\left(\frac{N_j^{(n)}-N^{(n)}\pi_j}{\sqrt{n\pi_j}}\right)_j = (I-ee^T)\mathcal{Y}_n \overset{n\to\infty}{\longrightarrow} (I-ee^T)\mathfrak{Z} \quad \text{in Verteilung}$$

(i) Nach Bemerkung 1 ist, bedingt unter $N^{(n)}=n$,

$$\left(\frac{N_j^{(n)}-N^{(n)}\pi_j}{\sqrt{n\pi_j}}\right) \quad \text{so verteilt wie} \quad \left(\frac{K_j^{(n)}-n\pi_j}{\sqrt{n\pi_j}}\right)$$

(ii) Nach Fishers Lemma und der obigen Konvergenzaussage ist $(I-ee^T)\mathcal{Y}_n$ für $n\to\infty$ asymptotisch unabhängig von $ee^T\mathcal{Y}_n$ und damit von $N^{(n)}$.

Fazit aus (i) und (ii):

$\left(\frac{K_j^{(n)}-n\pi_j}{\sqrt{n\pi_j}}\right)$ ist für große n asymptotisch so verteilt wie $(I-ee^T)\mathfrak{Z}$. Nach dem Satz von Fisher ist aber dessen Längenquadrat $\chi^2(s-1)$-verteilt. □

Bemerkung

Stellen wir $K_j^{(n)}$ als Summe von Indikatoren dar, so ergibt sich aus den vorigen Beweisgedanken eine Aussage über die Verteilungskonvergenz der Summe gewisser zentrierter Zufallsvektoren.

Sei dazu $X_{1,}, X_2, \ldots$ u.i.v., mit

$$\mathrm{Ws}(X_i=j)=\pi_j, \quad j=1,\ldots,s.$$

Wir stellen $K_j^{(n)}$ dar als

$$K_j^{(n)} = \sum_{i=1}^n I_{\{X_i=j\}}$$

Damit ergibt sich:

$$\begin{aligned}\frac{K_j^{(n)}-n\pi_j}{\sqrt{n\pi_j}} &= \frac{1}{\sqrt{n}}\left(\frac{K_j^{(n)}}{\sqrt{\pi_j}}-n\sqrt{\pi_j}\right)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{1}{\sqrt{\pi_j}}I_{\{X_i=j\}}-\sqrt{\pi_j}\right) =: \frac{1}{\sqrt{n}}\sum_{i=1}^n V_j^{(i)}\end{aligned}$$

Für jedes i ist $\mathfrak{V}^{(i)} := (V_1^{(i)},\ldots,V_s^{(i)})^T$ ein zufälliger Vektor in $\mathbb{R}^s$. Die $\mathfrak{V}^{(i)}$ sind unabhängig und identisch verteilt. Die "Kovarianzmatrix" von $\mathfrak{V}^{(1)} =: \mathfrak{V}$ ist gegeben durch

$$\mathrm{Cov}(V_j,V_k) = \begin{cases} 1-\pi_j & \text{für} \quad j=k \\ -\sqrt{\pi_j\pi_k} & \text{für} \quad j\neq k \end{cases}$$

Dies schlägt sich auf die Kovarianzmatrix der Grenzverteilung durch. In der Tat ist

$$\begin{aligned}&\mathrm{Cov}\left(Z_j - \sqrt{\pi_j}\sum_{l=1}^{s}\sqrt{\pi_l}Z_l\ ,\ Z_k - \sqrt{\pi_k}\sum_{m=1}^{s}\sqrt{\pi_m}Z_m\right)\\ =&\begin{cases} 1-2\pi_j+\pi_j = 1-\pi_j & \text{für} \quad j=k\\ -2\sqrt{\pi_i\pi_k}+\sqrt{\pi_j\pi_k}\sum_{l=1}^{s}\pi_e = -\sqrt{\pi_j\pi_k} & \text{für} \quad j\neq k\end{cases}\end{aligned}$$

In beiden Fällen ist die Kovarianzmatrix also nichts anderes als $I - ee^T$, mit

$$e = (\sqrt{\pi_1}, \ldots, \sqrt{\pi_s})^T\ .$$

Im nächsten Abschnitt werden wir einen Zentralen Grenzwertsatz für unabhängige Zufallsvektoren kennenlernen, der ein klares Licht auf diese Angelegenheit wirft.

10.2 Der multivariate zentrale Grenzwertsatz

10.2.1 Kovarianzmatrizen

$\mathfrak{X} = (X_1, \ldots, X_k)^T$ sei eine $\mathbb{R}^k$-wertige Zufallsvariable mit $\mathrm{Var}\, X_i < \infty$.

Die Matrix

$$C^{\mathfrak{X}} := (\mathrm{Cov}(X_i, X_j))_{1\leq i,j\leq k}$$

heißt Kovarianzmatrix von $\mathfrak{X}$.

Bemerkung 1: (Kovarianzmatrix nach einer linearen Transformation)

Sei $A := (a_{ij})$ eine $d \times k$-Matrix, und

$$\mathcal{Y} := A \cdot \mathfrak{X}.$$

Dann gilt

$$\begin{aligned}C^{\mathcal{Y}} &= \left(\mathrm{Cov}\left(\sum_{r=1}^{k} a_{ir}X_r, \sum_{s=1}^{k} a_{js}X_s\right)\right)_{i,j}\\ &= \left(\sum_r\sum_s a_{ir}\mathrm{Cov}(X_r, X_s)a_{js}\right)_{i,j}\\ &= AC^{\mathfrak{X}}A^T\ .\end{aligned}$$

Satz: Jede Kovarianzmatrix ist symmetrisch und nicht negativ definit.

Beweis: a) Die Symmetrie ist klar. Zur positiven Definitheit:

Für $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ ist

$$\sum_{i,j}\alpha_i\,(\mathrm{Cov}X_i, X_j)\,\alpha_j = \left(\mathrm{Cov}\sum\alpha_i, X_i, \sum\alpha_j, X_j\right) = \mathrm{Var}\,\left(\sum\alpha_i X_i\right) \geq 0\,.\square$$

Tritt umgekehrt jede symmetrische, nicht negativ definite Matrix als Kovarianzmarix auf? Jawohl! Um dies einzusehen, erinnern wir uns, dass man aus symmetrischen, nicht negativ definiten Matrizen die Wurzel ziehen kann: Dies folgt aus dem fundamentalen

Lemma (über die Diagonalisierbarbeit symmetrischer Matrizen)

C sei eine symmetrische, nicht negativ definite $k \times k$-Matrix. Dann existieren Eigenwerte $\lambda_1^2 \geq \lambda_2^2 \geq \ldots \geq \lambda_k^2 \geq 0$ und eine zugehörige ON-Basis von Eigenvektoren $e_1, \ldots, e_k$.

Beweis: Lineare Algebra . □

Korollar 1: $C = \sum_{i=1}^k \lambda_i^2 e_i e_i^T$

denn: $C \cdot e_j = \lambda_j^2 e_j$

und: $\sum_{i=1}^k \lambda_i^2 e_i e_i^T e_j = \lambda_j^2 e_i$.

Korollar 2: Für $A := \sum_{i=1}^k \lambda_i e_i e_i^T$ gilt:

$$AA^T = C.$$

Korollar 3: $\mathfrak{Z}$ habe Kovarianzmatrix I, (also z.B.: $\mathfrak{Z}$ standard-normalverteilt auf $\mathbb{R}^k$), A sei wie in Korollar 2. Dann hat wegen Bemerkung 1 $\mathfrak{X} := A\mathfrak{Z}$ die Kovarianzmatrix C.

10.2.2 Normalverteilungen auf $\mathbb{R}^k$

Wir schließen an das eben formulierte Korollar an. Sei

$$C = \sum_{i=1}^k \lambda_i^2 e_i e_i^T \quad \text{mit } (e_i) \quad \text{ON-Basis im } \mathbb{R}^k$$

und

$$A := \sum_{i=1}^k \lambda_i e_i e_i^T .$$

Machen wir uns ein Bild der Verteilung von $\mathfrak{X} = A \cdot \mathfrak{Z}$, mit $\mathfrak{Z}$ standard-normalverteilt auf $\mathbb{R}^k$.

Sei $W_j = e_j^T \mathfrak{X} = \lambda_j e_j^T \mathfrak{Z}$ die Koordinate von $\mathfrak{X}$ in e_j-Richtung. Nach dem Lemma von Fisher sind die $e_j^T \mathfrak{Z}$, $j = 1, \ldots, k$, unabhängig und normalverteilt. Diese werden mit den Faktoren λ_j gestreckt. Also sind die W_j $\mathcal{N}(0, \lambda_j^2)$-verteilt und unabhängig.

Satz 1 C sei symmetrisch und positiv semidefinit, mit Orthogonalzerlegung $C = \sum_{i=1}^k \lambda_i^2 e_i e_i^T$.

B sei eine $(k \times m)$-Matrix mit

$$BB^T = C.$$

$\tilde{\mathfrak{Z}}$ sei standard-normalverteilt auf $\mathbb{R}^m$. Dann sind die Koordinaten von $B\tilde{\mathfrak{Z}}$ in der Basis (e_j) unabhängig und $\mathcal{N}\big(0, \lambda_j^2\big)$-verteilt.

Beweis: Sei

$$J := \{i | i \in \{1, \ldots, k\}, \lambda_i > 0\}.$$

Die Vektoren $\frac{1}{\lambda_i} B^T e_i \ , \ i \in I$ bilden ein Orthogonalsystem in $\mathbb{R}^m$.
(In der Tat:

$$\frac{1}{\lambda_i \lambda_j} e_i^T \underbrace{BB^T e_j}_{\lambda_j^2 e_j} = \frac{\lambda_j}{\lambda_i} e_i^T e_j = \delta_{ij}).$$

Also sind nach Fishers Lemma die Zufallsvariablen

$$\frac{1}{\lambda_i} e_i^T B \cdot \tilde{\mathfrak{Z}} \ , \ i \in I,$$

$\mathcal{N}(0,1)$-verteilt und unabhängig.

Andererseits gilt für $i \notin I$:

$$\mathrm{Var}\,(e_i^T B\tilde{\mathfrak{Z}}) = e_i^T BB^T e_i = e_i^T C e_i = \lambda_i^2 = 0,$$

also $\quad e_i^T B\tilde{\mathfrak{Z}} = 0 \quad$ f.s.

Definition: $\mathfrak{X}$ heißt **normalverteilt** auf $\mathbb{R}^k : \Longleftrightarrow \ \mathfrak{X}$ entsteht durch affin lineare Transformation eines standard-normalverteilten $\mathfrak{Z}$.

Satz 2 Eine mehrdimensionale Normalverteilung ist eindeutig bestimmt durch ihren Mittelwertvektor und ihre Kovarianzmatrix.

Beweis: Angenommen

$$\mathfrak{X} = \mu + A\mathfrak{Z} \quad \text{ und } \mathcal{Y} = \tilde{\mu} + B\tilde{\mathfrak{Z}}$$

habe denselben Erwartungswert und dieselbe Konvarianzmatrix. Dann folgt: $\mu = \tilde{\mu}$, und $AA^T = BB^T$. Nach Satz 1 ergibt sich daraus: $\quad \mathcal{L}(A\mathfrak{Z}) = \mathcal{L}(B\tilde{\mathfrak{Z}})$.

Bemerkung 1: Affin lineare Transformationen von normalverteilten zufälligen Vektoren sind wieder normalverteilt!

Bemerkung 2: Sind $\mathfrak{V}_1, \mathfrak{V}_2, \ldots$ unabhängig und identisch verteilt auf $\mathbb{R}^k$ mit Mittelwert 0 und Kovarianzmatrix C, und ist $\quad a \in \mathbb{R}^k$, so gilt nach dem klassischen zentralen Grenzwertsatz:

$$a^T \cdot \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathfrak{V}_i \right) \longrightarrow \mathcal{N}\big(0, a^T C a\big) \qquad \text{in Verteilung.}$$

10.2.3 Verteilungskonvergenz von $\mathbb{R}^k$-wertigen Zufallsvariablen

Definition: $\mathfrak{X}_n, \mathfrak{X}$ seien $\mathbb{R}^k$-wertige Zufallsvariable. Wir definieren:

$$\mathfrak{X}_n \to \mathfrak{X} \text{ in Verteilung } :\Longleftrightarrow \ \mathbb{E}\, h(\mathfrak{X}_n) \longrightarrow \mathbb{E}\, h(\mathfrak{X})$$

für alle $h : \mathbb{R}^k \to \mathbb{R}$ stetig und beschränkt.

Bemerkung 1: Dieser Konvergenzbegriff hängt nur von den Verteilungen der einzelnen $\mathfrak{X}_i$ und der Verteilung von $\mathfrak{X}$ ab. Anstelle von "$\mathfrak{X}_n \to \mathfrak{X}$ in Verteilung" schreiben wir daher auch "$\mathfrak{X}_n \to \mathcal{L}(\mathfrak{X})$".

Bemerkung 2: Gilt $\mathfrak{X}_n \to \mathfrak{X}$ in Verteilung, und ist $g : \mathbb{R}^k \to \mathbb{R}^m$ stetig, so folgt: $g(\mathfrak{X}_n) \to g(\mathfrak{X})$ in Verteilung.

Der folgende Satz, den wir hier nicht beweisen werden, erlaubt es, die Verteilungskonvergenz von $\mathbb{R}^k$-wertigen Zufallsvariablen auf die Verteilungskonvergenz ihrer eindimensionalen Projektionen zurück zu spielen.

Satz: $\mathfrak{X}_n, \mathfrak{X}$ seien $\mathbb{R}^k$-wertige Zufallsvariable. Dann gilt:

$$\mathfrak{X}_n \to \mathfrak{X} \quad \text{in Verteilung} \iff$$

$$\forall a \in \mathbb{R}^k : a^T \mathfrak{X}_n \to a^T \mathfrak{X} \quad \text{in Verteilung}$$

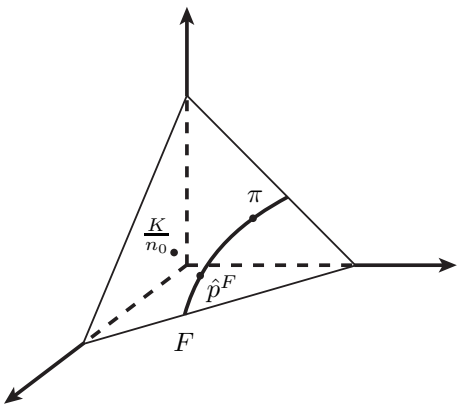
Beweis Siehe z.B. O. Kallenberg, Foundations of modern probability, 2. Aufl., Springer 2002.

Als Korollar zu diesem Satz und zur Bemerkung am Ende des vorigen Abschnitts ergibt sich der **Multivariate Zentrale Grenzwertsatz**: Sind $\mathfrak{V}_1, \mathfrak{V}_2, \ldots$ unabhängig und identisch verteilt auf $\mathbb{R}^k$ mit Mittelwert 0 und Kovarianzmatrix C, so gilt:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathfrak{V}_i \longrightarrow \mathcal{N}(0, C)\,.$$

10.3 Der χ^2-Test

Wie gut passen beobachtete Häufigkeiten zu einer hypothetischen Klasse von Häufigkeitsverteilungen? Die beobachteten Häufigkeiten seien $K_1, \ldots, K_s$. Der Vektor $K = (K_1, \ldots, K_s)$ sei multinomial $(n; \pi)$-verteilt. Hypothese: $\pi \in F$, wobei F eine r-dimensionale Hyperfläche des Simplex $\{(p_j) | \sum_{j=1}^{s} p_j = 1\}$ ist.



$\hat{p}^F$ sei der ML-Schätzer für π unter der Hypothese $\pi \in F$.

Teststatistik:

$$T := \sum_{j=1}^{s} \frac{\left(K_j - n\hat{p}_j^F\right)^2}{n\hat{p}_j^F} \ .$$

Satz (K. Pearson, R.A. Fisher) Unter der Hypothese $\pi \in F$ ist T für große n annähernd $\chi^2(s-1-r)$-verteilt.

Beweisskizze für den Fall $r = 1$ und

$$F = \{p(\gamma) | \gamma \in I \subseteq \mathbb{R}\}$$

(d.h. F ist durch γ parametrisiert).

$\hat{p}^F = p(\hat{\gamma})$, wobei $\hat{\gamma}$ bestimmt wird durch die ML-Gleichung

$$0 = \frac{\partial}{\partial \gamma} \log \prod_j p_j(\hat{\gamma})^{K_j} = \sum_{j=1}^{s} K_j \frac{1}{p_j(\hat{\gamma})} \frac{\partial p_j(\hat{\gamma})}{\partial \gamma} \qquad (\star)$$

Sei $\pi = p(\gamma_0)$.

Wir zerlegen mit einer Taylorentwicklung um γ_0:

$$\begin{aligned} \frac{K_j - n\pi_j}{\sqrt{np_j(\hat{\gamma})}} &= \frac{K_j - np_j(\hat{\gamma})}{\sqrt{np_j(\hat{\gamma})}} + \frac{n(p_j(\hat{\gamma}) - p_j(\gamma_0))}{\sqrt{np_j(\hat{\gamma})}} \\ &\approx \frac{K_j - np_j(\hat{\gamma})}{\sqrt{np_j(\hat{\gamma})}} + \sqrt{n} \frac{1}{\sqrt{p_j(\hat{\gamma})}} \frac{\partial p_j}{\partial \gamma}(\hat{\gamma})(\hat{\gamma} - \gamma_0) \end{aligned}$$

Wir setzen

$$\mathfrak{G} := \left(\frac{K_j - n\pi_j}{\sqrt{np_j(\hat{\gamma})}} \right)_{j=1,\ldots,s} , \ \mathfrak{H} := \left(\frac{K_j - np_j(\hat{\gamma})}{\sqrt{np_j(\hat{\gamma})}} \right)_{j=1,\ldots,s}$$

$$\mathcal{M} := \frac{1}{\sqrt{p_j(\hat{\gamma})}} \frac{\partial p_j}{\partial \gamma}(\hat{\gamma}) \ .$$

Wir haben gesehen:

$$\mathfrak{G} \approx \mathfrak{H} + \sqrt{n}(\hat{\gamma} - \gamma_0)\mathcal{M} \qquad (\star\star)$$

Es gilt: $\mathfrak{H} \perp \mathcal{M}$,

denn:

$$\begin{aligned} \mathfrak{H}^T \cdot \mathcal{M} &= \frac{1}{\sqrt{n}} \sum_{j=1}^{s} \frac{K_j - np_j(\hat{\gamma})}{\sqrt{p_j(\hat{\gamma})}} \cdot \frac{1}{\sqrt{p_j(\hat{\gamma})}} \frac{\partial p_j(\hat{\gamma})}{\partial \gamma} \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^{s} K_j \frac{1}{\sqrt{p_j(\hat{\gamma})}} \frac{\partial p_j(\hat{\gamma})}{\partial \gamma} - \sqrt{n} \frac{\partial}{\partial \gamma} \sum_{j=1}^{s} p_j(\hat{\gamma}) \ . \end{aligned}$$

Der erste Term ist Null wegen der ML-Gleichung $(\star)$, der zweite verschwindet wegen $\sum_{j=1}^{s} p_j(\gamma) \equiv 1$.

Projektion von $(\star\star)$ auf den zu $\mathcal{M}$ orthogonalen Raum (d.h. Multiplikation von $(\star\star)$ mit $I - \frac{\mathcal{M}\mathcal{M}^T}{\|\mathcal{M}\|^2}$ liefert

$$\mathfrak{H} \approx \left(I - \frac{\mathcal{M}\mathcal{M}^T}{\| \mathcal{M} \|^2} \right) \mathfrak{G}$$

Nun ist aber (vgl. Beweis des Satzes von Pearson) $\mathfrak{G}$ approximativ so verteilt wie $(I - ee^T)\mathfrak{Z}$, mit $\mathfrak{Z}$ standard-normalverteilt auf $\mathbb{R}^s$ und $e = (\sqrt{\pi_1}, \ldots, \sqrt{\pi_s})^T$.

Und - oh Wunder - $\mathcal{M}$ und e stehen approximativ senkrecht aufeinander, denn

$$\sum_j \underbrace{\frac{\sqrt{\pi_j}}{\sqrt{p_j(\hat{\gamma})}}}_{\approx 1} \frac{\partial p_j(\hat{\gamma})}{\partial \gamma} \approx \frac{\partial}{\partial \gamma} \sum_j p_j(\hat{\gamma}) = 0 \, .$$

Somit hat $\mathfrak{H}$ approximativ dieselbe Verteilung wie

$$\left(I - \frac{\mathcal{M}\mathcal{M}^T}{\| \mathcal{M} \|^2} \right) (I - ee^T)\mathfrak{Z} \approx \left(I - \frac{\mathcal{M}\mathcal{M}^T}{\| \mathcal{M} \|^2} - ee^T \right) \mathfrak{Z}$$

Wegen $T = \| \mathfrak{H} \|^2$ folgt die Behauptung nun aus dem Satz von Fisher. □

10.3.1 χ^2-Test auf Unabhängigkeit zweier Merkmale

Wir betrachten Gewichte $p_{ij}, i = 1, \ldots, I, j = 1, \ldots, J$.

Die Hypothese ist: p_{ij} ist von Produktform, d.h.

$$p_{ij} = \alpha_i \beta_j \quad , \quad \sum \alpha_i = \sum \beta_j = 1.$$

$$\dim (\text{Hypothese}) = (I - 1) + (J - 1).$$

Es ist $s = I \cdot J$.

Ist $\vec{k} = (k_{ij})$ der beobachtete Vektor der Häufigkeiten, so ist die Likelihood

$$\begin{aligned} f_{p(\alpha,\beta)}(\vec{k}) &= \binom{n}{k_{11} \ldots k_{IJ}} \prod_{i,j} p_{ij}(\alpha, \beta)^{k_{ij}} \\ &= \text{const} \prod_{i,j} (\alpha_i \beta_j)^{k_{ij}} = \text{const} \prod_i \alpha_i^{k_{i+}} \prod_i \beta_j^{k_{+j}} \, , \end{aligned}$$

wobei

$$k_{i+} := \sum_j k_{ij} \quad , \quad k_{+j} := \sum_i k_{ij} \, .$$

Maximierung von $f_{p(\alpha,\beta)}(\vec{k})$ unter der Nebenbedingung $\sum \alpha_i = \sum \beta_j = 1$ führt auf (vgl. frühere Rechnung)

$$\hat{\alpha_i} = \frac{k_{i+}}{n} \quad , \quad \hat{\beta_j} = \frac{k_{+j}}{n}.$$

Somit folgt aus dem Satz von Pearson/Fisher:

Ist (K_{ij}) multinomial $(n, (\alpha_i\beta_j))$-verteilt, dann ist

$$\sum_{i,j} \frac{(K_{ij} - n\hat{\alpha}_i\hat{\beta}_j)^2}{n \cdot \hat{\alpha}_i\hat{\beta}_j} = \sum_{i,j} \frac{\left(K_{ij} - \frac{K_{i+}K_{+j}}{n}\right)^2}{\frac{K_{i+}K_{+j}}{n}}$$

für großes n approximativ

$$\chi^2(IJ - 1 - (I - 1 + J - 1)) = \chi^2((I - 1)(J - 1))\text{-verteilt.}$$

Kapitel 11

Zur Asymptotik der ML-Schätzer

$X_1, X_2, \ldots$ seien unabhängige, identsich verteilte Zufallsvariablen mit Verteilungsdichten (bzw. -gewichten) $g_\vartheta(x)$. Für die Verteilungsdichte (bzw. das -gewicht) f_ϑ von $\mathfrak{X} := (X_1, \ldots, X_n)$ gilt dann:

$$f_\vartheta(x_1, \ldots, x_n) = g_\vartheta(x_1) \ldots g_\vartheta(x_n).$$

Wir setzen

$$L(\vartheta, x) := \log g_\vartheta(x).$$

Der ML-Schätzer $\hat{\vartheta}_n(\mathfrak{X})$ ist Maximalstelle von

$$\vartheta \to \sum_{i=1}^{n} L(\vartheta, X_i)$$

und somit (in gutartigen Fällen) Nullstelle von

$$\vartheta \to \sum_{i=1}^{n} L'(\vartheta, X_i)$$

mit $L'(\vartheta, x) := \frac{\partial}{\partial \vartheta} L(\vartheta, x)$.

Betrachten wir erst den Fall, dass Θ ein Intervall ist. Die Zufallsvariable $L'(\vartheta, X_1)$ besitzt die bemerkenswerte Eigenschaft

$$\mathbb{E}_\vartheta L'(\vartheta, X_i) = 0 \tag{11.1}$$

Denn:

$$\begin{aligned} \mathbb{E}_\vartheta L'(\vartheta, X_i) &= \int \frac{\partial}{\partial \vartheta} \log g_\vartheta(x) \cdot g_\vartheta(x) dx \\ &= \int \frac{\partial}{\partial \vartheta} g_\vartheta(x) dx = \frac{\partial}{\partial \vartheta} 1 = 0 \,. \end{aligned}$$

Dabei setzen wir voraus

$$\vartheta \to g_\vartheta(x) \text{ ist stetig differenzierbar mit}$$
$$\int \frac{\partial}{\partial\vartheta} g_\vartheta(x)dx = \frac{\partial}{\partial\vartheta}\int g_\vartheta(x)dx \tag{11.2}$$

Satz 1 (über die Konsistenz des ML-Schätzers) Außer (11.2) gelte:

$$\vartheta \to L'(\vartheta, x) \text{ sei strikt monoton fallend.}$$

Dann gilt: $\mathrm{Ws}_\vartheta\{|\hat{\vartheta}_n - \vartheta| > \varepsilon\} \overset{n\to\infty}{\longrightarrow} 0$.

Beweis: $\hat{\vartheta}_n$ ist eindeutig bestimmt durch die ML-Gleichung

$$\sum_{i=1}^n L'(\hat{\vartheta}_n, X_i) = 0 \tag{11.3}$$

Weil $\vartheta \to L'(\vartheta, x)$ nach Voraussetzung streng monoton fällt, folgt:

$$\vartheta'_n < \vartheta - \varepsilon \iff \sum_{i=1}^n L'(\vartheta - \varepsilon, X_i) < 0$$

sowie

$$\mathbb{E}_\vartheta L'(\vartheta - \varepsilon, X_i) > \mathbb{E}_\vartheta L'(\vartheta, X_i) \quad (= 0 \text{ wegen}(11.1)).$$

Also:

$$\begin{aligned}
&\mathrm{Ws}_\vartheta\{\vartheta'_n < \vartheta - \varepsilon\} = \mathrm{Ws}_\vartheta\left\{\frac{1}{n}\sum_{i=1}^n L'(\vartheta - \varepsilon, X_i) < 0\right\} \\
&= \mathrm{Ws}_\vartheta\left\{\frac{1}{n}\sum_{i=1}^n L'(\vartheta - \varepsilon, X_i) - \mathbb{E}_\vartheta L'(\vartheta - \varepsilon, X_1) < -\mathbb{E}_\vartheta L'(\vartheta - \varepsilon, X_1)\right\} \\
&\le \mathrm{Ws}_\vartheta\left\{|\frac{1}{n}\sum_{i=1}^n L'(\vartheta - \varepsilon, X_i) - \mathbb{E}_\vartheta L'(\vartheta - \varepsilon, X_1)| > \mathbb{E}_\vartheta L'(\vartheta - \varepsilon, X_1)\right\} \\
&\overset{n\to\infty}{\longrightarrow} 0
\end{aligned}$$

nach dem Schwachen Gesetz der Großen Zahlen. □

Unter geeigneten Regularitätsvoraussetzungen an $g_\vartheta(x)$ ist $\hat{\vartheta}_n$ für $n \to \infty$ nicht nur "konsistent" (im Sinn von Satz 1), sondern sogar "asymptotisch normal", mit asymptotischer Varianz $\frac{1}{n}I(\vartheta)^{-1}$, wobei

$$I(\vartheta) := \mathbb{E}_\vartheta\left[L'(\vartheta, X_1)^2\right].$$

Dahinter steckt der Zentrale Grenzwertsatz und eine Taylorentwicklung der ML-Gleichung um den wahren Parameter ϑ bis zur 2. Ordnung. Hier ist die
Beweisidee: Wegen (11.3) gilt für eine geeignete Zwischenstelle $\tilde{\vartheta}_n = \tilde{\vartheta}_n(\mathfrak{X})$ (mit $|\tilde{\vartheta}_n - \vartheta| \le |\hat{\vartheta}_n - \vartheta|$):

$$0 = \sum_{i=1}^n L'(\hat{\vartheta}_n, X_i) + (\hat{\vartheta}_n - \vartheta)\sum_{i=1}^n L''(\tilde{\vartheta}_n, X_i)$$

oder äquivalent dazu

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n L'(\vartheta, X_i)}{\frac{1}{n} \sum_{i=1}^n L''(\tilde{\vartheta}_n, X_i)}.$$

Nach dem Zentralen Grenzwertsatz (und wegen der Zentriertheit von $L'(\vartheta, X_1)$: konvergiert der Zähler in Verteilung gegen $\mathcal{N}(0, I(\vartheta))$.

Nach dem Gesetz der großen Zahlen konvergiert der Nenner gegen $\mathbb{E}_\vartheta L''(\vartheta, X_1)$.

Nun ist aber (sieh an!):

$$\begin{aligned}
\mathbb{E}_\vartheta L''(\vartheta, X_1) &= \int g_\vartheta(x) \frac{\partial^2}{\partial \vartheta^2} \log g_\vartheta(x) dx \\
&= \int g_\vartheta(x) \left(\frac{\frac{\partial^2}{\partial \vartheta^2} g_\vartheta(x)}{g_\vartheta(x)} - \left(\frac{\frac{\partial}{\partial \vartheta} g_\vartheta(x)}{g_\vartheta(x)} \right)^2 \right) dx \\
&= \underbrace{\int \frac{\partial^2}{\partial \vartheta^2} g_\vartheta(x) dx}_{=0} - \mathbb{E}_\vartheta \left[\left(\frac{\partial}{\partial \vartheta} \log g_\vartheta(X_1) \right)^2 \right] \\
&= -I(\vartheta) \, .
\end{aligned}$$

Also: $\sqrt{n}(\hat{\vartheta}_n - \vartheta) \overset{n \to \infty}{\longrightarrow} \mathcal{N}\big(0, I(\vartheta)^{-1}\big) \quad \forall \vartheta \, .$

Definition: $I(\vartheta) := \mathbb{E}_\vartheta \left[L'(\vartheta, X_1)^2 \right] = -\mathbb{E}_\vartheta \left[L''(\vartheta, X_1) \right]$
heißt "Fisher-Information".

Satz 2 (über die Asymptotik der ML-Schätzer)

(Harald Cramér 1946, siehe C.R. Rao, Linear Statistical Inference and its Applications, Wiley 1973, p. 364 f. und p. 349).

Θ sei ein offenes Intervall. Unter Ws_ϑ seien $X_1, X_2, \ldots$ unabhängig und identisch verteilt mit Dichte (oder Gewichten) g_ϑ. Es gelte :

(i) $\vartheta \mapsto L(\vartheta, x) := \log g_\vartheta(x)$ ist 3 mal differenzierbar.

(ii) $\int \frac{\partial}{\partial \vartheta} g_\vartheta(x) dx = \int \frac{\partial^2}{\partial \vartheta^2} g_\vartheta(x) dx = 0$

$I(\vartheta) := \int g_\vartheta(x) (L'(\vartheta, x)^2) dx \in (0, \infty)$

(iii) $\exists H(x)$, so dass

$|L'''(\vartheta, x)| \leq H(x)$ mit $\sup_\vartheta \int g_\vartheta(x) H(x) dx < \infty.$

Dann existiert ein $\hat{\vartheta}_n = \hat{\vartheta}_n(X_1, \ldots, X_n)$ so, dass für alle $\vartheta \in \Theta$ unter Ws_ϑ gilt:

a) $\sum_{i=1}^n L'(\hat{\vartheta}_n, X_i) = 0$ f.s.

b) $\hat{\vartheta}_n \overset{n\to\infty}{\longrightarrow} \vartheta$ f.s.

c) $\sqrt{n}(\hat{\vartheta}_n - \vartheta) - I(\vartheta)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n L'(\vartheta, X_i) \overset{n\to\infty}{\longrightarrow} 0$ stochastisch

d) $\sqrt{n}(\hat{\vartheta}_n - \vartheta) \overset{n\to\infty}{\longrightarrow} \mathcal{N}\big(0, I(\vartheta)^{-1}\big)$ in Verteilung

Definition: Eine Folge S_n von Schätzern für ϑ heißt:

i) stark konsistent $:\Longleftrightarrow$
$\forall \vartheta : \mathrm{Ws}_\vartheta\{S_n \to \vartheta\} = 1$

ii) asymptotisch normal mit asymptotischer Varianz $\frac{1}{n}\sigma^2(\vartheta) :\Longleftrightarrow$
$\forall \vartheta : \sqrt{n}(S_n - \vartheta) \overset{n\to\infty}{\longrightarrow} \mathcal{N}\big(0, \sigma^2(\vartheta)\big)$ in Verteilung (unter Ws_ϑ).

Damit ergibt sich als
Korollar 1 Ist unter den Voraussetzungen des Satzes von Cramèr/Rao die ML-Gleichung (a) eindeutig lösbar mit Lösung =: $\hat{\vartheta}_n$, dann ist die Folge $\hat{\vartheta}_n$ stark konsistent und asymptotisch normal mit asymptotischer Varianz $\frac{1}{n} I(\vartheta)^{-1}$.

Korollar 2 Insbesondere gilt dann: mit

$$K_n := \left[\hat{\vartheta}_n - \frac{1}{\sqrt{n}} \frac{1}{\sqrt{I(\hat{\vartheta}_n)}} \text{ qnorm } \left(1 - \frac{\alpha}{2}\right),\ \hat{\vartheta}_n + \frac{1}{\sqrt{n}} \frac{1}{\sqrt{I(\hat{\vartheta}_n)}} \text{ qnorm } \left(1 - \frac{\alpha}{2}\right) \right] :$$

$$\mathrm{Ws}_\vartheta\{K_n \ni \vartheta\} \overset{n\to\infty}{\longrightarrow} 1 - \alpha \qquad \forall \vartheta.$$

In diesem Sinn ist also K_n ein asymptotisches $(1-\alpha)$-Konfidenzintervall für ϑ.

Wenden wir uns schließlich noch der multivariaten Variante des Satzes von Cramèr-Rao zu. Sei $\Theta \subseteq \mathbb{R}^k$.

Für $\vartheta \in \Theta$ ist die Fisher-Informationsmatrix definiert als die $k \times k$-Matrix

$$I(\vartheta) := \mathbb{E}_\vartheta \left[\frac{\partial}{\partial \vartheta_i} L(\vartheta, X_1) \frac{\partial}{\partial \vartheta_j} L(\vartheta, X_1) \right]_{i,j} .$$

Unter ähnlichen Regularitätsvoraussetzungen wie im 1−dim. Fall gilt für die ML-Schätzer $\hat{\vartheta}_n$ wieder a) - d), wobei jetzt

$$L'(\vartheta, x) := \left(\frac{\partial}{\partial \vartheta_1} L(\vartheta, x), \ldots, \frac{\partial}{\partial \vartheta_k} L(\vartheta, x) \right)$$

der Gradient von $L(\cdot, x)$ ist.

Korollar (Asymptotische Verteilung für ein Parametermerkmal)

Sei $m : \Theta \to \mathbb{R}^k$ differenzierbar. Für große n gilt:

$$\sqrt{n} \left(m(\hat{\vartheta}_n) - m(\vartheta) \right) \approx \text{ grad } m(\vartheta) \cdot \sqrt{n} \left((\hat{\vartheta}_n - \vartheta) \right) ,$$

also

$$\sqrt{n} \left(m(\hat{\vartheta}_n) - m(\vartheta) \right) \overset{n\to\infty}{\longrightarrow} \mathcal{N}\big(0,\ \text{grad } m(\vartheta) I(\vartheta)^{-1} \text{ grad } m(\vartheta)^T\big)$$

Kapitel 12

Suffizienz und Vollständigkeit

12.1 Bedingte Verteilung und bedingte Erwartung

Einer der wichtigsten in der Elementaren Stochastik besprochenen Zusammenhänge ist der zwischen gemeinsamer Verteilung und bedingter Verteilung.

Ist ρ die Verteilung von Y, $\pi(y,\cdot)$ die bedingte Verteilung von X gegeben $Y=y$, und ν die gemeinsame Verteilung von X und Y, so gilt

$$\nu(dy,dx) = \rho(dy)\pi(y,dx) \ .$$

Für reellwertiges $\varphi = \varphi(x)$ setzen wir

$$\mathbb{E}\left[\varphi(X) \mid Y=y\right] := \int \varphi(x)\pi(y,dx)$$

Für jedes y ist dies eine Zahl, wir nennen sie den **bedingten Erwartungswert** von $\varphi(X)$, **gegeben** $Y=y$.

Die Zufallsvariable

$$\mathbb{E}\left[\varphi(X) \mid Y\right] := \int \varphi(x)\pi(Y,dx)$$

heißt **bedingte Erwartung** von $\varphi(X)$, **gegeben** Y.

Der Satz von der totalen Wahrscheinlichkeit übersetzt sich in die wichtige Beziehung

(a) $\mathbb{E}\left[\mathbb{E}\left[\varphi(X) \mid Y\right]\right] := \mathbb{E}\left[\varphi(X)\right]$.

Denn in der Tat ist

$$\int\int \varphi(x)\pi(y,dx)\rho(dy) = \int \varphi(x)\nu(dx,dy) = \mathbb{E}\left[\varphi(X)\right] \ .$$

Eine weitere wichtige Beziehung ist

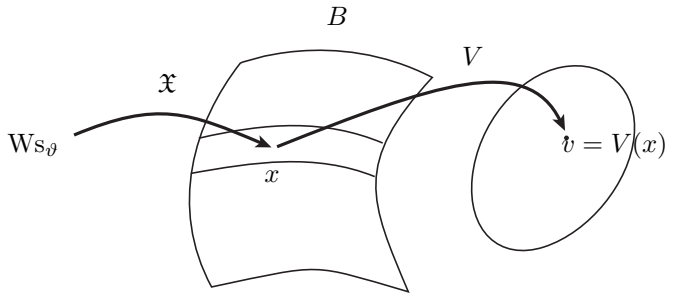
(b) $\mathbb{E}\left[\mathbb{E}\left[\varphi(X) \mid Y\right]^2\right] \leq \mathbb{E}\left[\varphi(X)^2\right]$.

In der Tat gilt (weil das Quadrat eines Erwartungswerts stets $\leq$ dem Erwartungswert des Quadats ist)

$$\mathbb{E}\left[\varphi(X) \mid Y\right]^2 = \left(\int \varphi(x)\pi(Y, dz)\right)^2 \leq \int \varphi(x)^2 \pi(Y, dz) = \mathbb{E}\left[\varphi(X)^2 \mid Y\right] .$$

Bildet man über dieser Ungleichung den Erwartungswert, so wird die rechte Seite wegen (a) zu $\mathbb{E}\left[\varphi(X)^2\right]$.

12.2 Suffizienz



Angenommen, die Verteilung von $\mathfrak{X}$, gegeben $V(\mathfrak{X}) = v$, hängt nicht von ϑ ab. Für die Schätzung von ϑ bringt dann die Beobachtung von x keine zusätzliche Information gegenüber der vergröberten Beobachtung $v = V(x)$.

Definition: Eine Statistik $V(\mathfrak{X})$ heißt **suffizient** für ϑ , falls die bedingte Verteilung von $\mathfrak{X}$ gegeben $V(\mathfrak{X})$ nicht von ϑ abhängt.

Beispiel 1:

a) $\mathfrak{X} = (X_1, \ldots, X_n)$, mit $X_1, X_2, \ldots$ Münzwurffolge mit Parameter $p \in (0, 1)$.

Dann ist $K := \sum_{i=1}^n X_i$ suffizient für p.

Denn: gegeben $K = k$ ist $(X_1, \ldots, X_n)$ Laplaceverteilt auf $\{(x_1, \ldots, x_n) \in \{0, 1\}^n \mid \sum_{i=1}^n x_i = k\}$.

In der Tat: Für $\sum x_i = k$ ist

$$\mathrm{Ws}_p\left(\left\{(X_1, \ldots, X_n) = (x_1, \ldots, x_n) \;\middle|\; \sum_{i=1}^n X_i = k\right\}\right) =$$

$$= \frac{p^k(1-p)^{n-k}}{\binom{n}{k} p^k(1-p)^{n-k}} = \frac{1}{\binom{n}{k}} .$$

b) Analog gilt:

$\mathfrak{X} = (X_1, \ldots, X_n)$, mit $X_1, X_2, \ldots$ "Würfelfolge" zu den Verteilungsgewichten $(p_1, \ldots, p_s) = \vec{p}$.

Dann ist der Vektor der Besetzungszahlen $\mathcal{K} = (K_1, \ldots, K_s)$ suffizient für $\vec{p}$.

Satz (Faktorisierungssatz von Neyman-Pearson)

Für alle ϑ besitze die Verteilung von $\mathfrak{X}$ unter Ws_ϑ die Dichte (bzw. die Gewichte) $f_\vartheta(x)$.

Dann sind äquivalent:

(i) $V(\mathfrak{X})$ ist suffizient für ϑ

(ii) $\exists \psi_\vartheta(v), \rho(x) : f_\vartheta(x) = \psi_\vartheta(V(x)) \cdot \rho(x)$.

Beweis: hier nur für den diskreten Fall

$$\begin{aligned}(i) \Rightarrow (ii): \quad f_\vartheta(x) &= \mathrm{Ws}_\vartheta(\{\mathfrak{X} = x\}) \\ &= \mathrm{Ws}_\vartheta(\{\mathfrak{X} = x\} \cap \{V(\mathfrak{X}) = V(x)\}) \\ &= \mathrm{Ws}_\vartheta(\{V(\mathfrak{X}) = V(x)\}) \cdot \mathrm{Ws}_\vartheta(\{\mathfrak{X} = \mathfrak{x}\} | (\{V(\mathfrak{X}) = V(x)\}) \; .\end{aligned}$$

Wegen der Suffizienz von $V(\mathfrak{X})$ hängt der 2. Faktor nicht von ϑ (und somit nur von von x) ab. Wir nennen ihn $\rho(x)$, und den ersten Faktor nennen wir $\psi_\vartheta(V(x))$.

$$\begin{aligned}(ii) \Rightarrow (i): \quad \mathrm{Ws}_\vartheta(\{\mathfrak{X} = x\} | \{V(\mathfrak{X}) = V(x)\}) &= \frac{\mathrm{Ws}_\vartheta(\{\mathfrak{X} = x\})}{\mathrm{Ws}_\vartheta(\{V(\mathfrak{X}) = V(x)\})} \\ = \frac{\psi_\vartheta(V(x))\rho(x)}{\sum_{y:V(y)=V(x)} \psi_\vartheta(V(y))\rho(y)} &= \frac{\rho(x)}{\sum_{y:V(y)=V(x)} \rho(y)} \quad ,\end{aligned}$$

und dies hängt nicht von ϑ ab. Also ist $V(\mathfrak{X})$ suffizient. □

Beispiel 2: $\mathfrak{X} = (X_1, \ldots, X_n)$ mit X_i unabhängig, $\mathcal{N}(\mu, \sigma^2)$-verteilt.

$$\begin{aligned}f_\vartheta(x_1, \ldots, x_n) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \, e^{-\frac{\sum (x_i - \mu)^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \, e^{-\frac{\sum x_i^2 - 2\mu \sum x_i + n\mu^2}{2\sigma^2}}\end{aligned}$$

Also ist $V(\mathfrak{X}) := \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ suffizient für (μ, σ^2).

Beispiel 3: $\mathfrak{X} = (X_1, \ldots, X_n)$ mit X_i unabhängig, Gamma(k, α)-verteilt.

$$f_{k,\alpha}(x_1, \ldots, x_n) = \mathrm{const}(\alpha, k)(x_1 \ldots x_n)^{k-1} e^{-\alpha(x_1 + \ldots + x_k)}$$

Also ist $V(\mathfrak{X}) := (\prod_{i=1}^n X_i, \sum_{i=1}^n X_i)$ suffizient für (k, α).

12.3 Der Satz von Rao-Blackwell

Ist $V(\mathfrak{X})$ suffizient für ϑ, dann hängt für jede Verarbeitung φ von $\mathfrak{X}$ die bedingte Erwartung $\mathbb{E}_\vartheta\left[\varphi(\mathfrak{X})|V(\mathfrak{X})\right]$ nicht von ϑ ab! Wie wir sehen werden, kann's etwas bringen, von einem Schätzer $S = h(\mathfrak{X})$ zu dessen "Glättung" $\mathbb{E}_\vartheta\left[h(\mathfrak{X})|V(\mathfrak{X})\right]$ über zu gehen, man wird uninformatives Rauschen los.

Satz (Rao-Blackwell)

Sei $S = h(\mathfrak{X})$ ein Schätzer für das reelle Parametermerkmal $m(\vartheta)$. Sei $V(\mathfrak{X})$ eine suffiziente Statistik für ϑ. Dann gilt für den (auf $V(\mathfrak{X})$ basierenden) Schätzer

$$S^\star := \mathbb{E}\left[S|V(\mathfrak{X})\right] := \mathbb{E}_\vartheta\left[S|V(\mathfrak{X})\right].$$

1. Die Erwartungswerte bleiben gleich:

$$\forall\vartheta : \mathbb{E}_\vartheta[S^\star] = \mathbb{E}_\vartheta[S]$$

.

2. Der mittlere quadratische Fehler wird nicht größer:

$$\forall\vartheta : \mathbb{E}_\vartheta\left[(S^\star - m(\vartheta))^2\right] \leq \mathbb{E}_\vartheta\left[(S - m(\vartheta))^2\right].$$

Beweis

1. $\mathbb{E}_\vartheta[S^\star] = \mathbb{E}_\vartheta\left[\mathbb{E}_\vartheta[S|V(\mathfrak{X})]\right] = \mathbb{E}_\vartheta[S]$
   nach dem Satz von der totalen Wahrscheinlichkeit.

2. $S^\star - m(\vartheta) = \mathbb{E}_\vartheta[S|V(\mathfrak{X})] - m(\vartheta) = \mathbb{E}_\vartheta[S - m(\vartheta)|V(\mathfrak{X})]$

   Quadrieren ergibt

$$\left(S^\star - m(\vartheta)\right)^2 = \left(\mathbb{E}_\vartheta[S - m(\vartheta)|V(\mathfrak{X})]\right)^2$$

   Also mit Bemerkung 2 aus dem vorletzten Abschnitt:

$$\mathbb{E}\left[\left(S^\star - m(\vartheta)\right)^2\right] \leq \mathbb{E}_\vartheta\left[(S - m(\vartheta))^2\right] \qquad \square$$

Bemerkung $S^\star$ (wie im vorigen Satz) heißt auch die Rao-Blackwellisierung von S (auf der Basis von V).

Definition $\vartheta \mapsto \mathbb{E}_\vartheta[S] - m(\vartheta)$ heißt **Bias** (oder **Verzerrung**) des Schätzers S (für das Parametermerkmal $m(\vartheta)$).

S heißt erwartungstreu (oder unverzerrt) :$\Leftrightarrow$

$$:\Leftrightarrow \forall\vartheta : \mathbb{E}_\vartheta S = m(\vartheta).$$

Bemerkung (vgl. Übungen): Der mittlere quadratische Fehler eines Schätzers ist die Summe aus Varianz und quadriertem Bias.

Wie bekommt man einen erwartungstreuen Schätzer mit (möglichst) kleiner Varianz?

Rezept: Man nehme einen erwartungstreuen Schätzer und Rao-Blackwellisiere ihn.

Beispiel: $X_1, X_2, \ldots, X_n$ Münzwurffolge zum Parameter $p, m(p) := p^2$

$S := X_1 \cdot X_2 \quad ; \quad \mathbb{E}_p S = p^2$

$K := \sum_{i=1}^{n} X_i$ ist suffizient für p.

Wie sieht die Rao-Blackwellisierung aus?

Gegeben $K = k$ ist $(X_1, , \ldots, X_n)$ Laplace-verteilt auf

$$\left(\left\{ (x_1, \ldots, x_n) \in \{0,1\}^n \middle| \sum x_i = k \right\} \right).$$

Also:

$$\mathbb{E}_p \left[X_1 X_2 \middle| K = k \right] = \mathrm{Ws}_p \left(\{X_1 = 1, X_2 = 1\} \middle| K = k \right) = \frac{k(k-1)}{n(n-1)},$$

$$S^\star = \frac{K(K-1)}{n(n-1)}.$$

12.4 Vollständigkeit

Für ein reelles Parametermerkmal $m(\vartheta)$ gibt es im allgemeinen viele erwartungstreue Schätzer. Gibt es eine übersichtliche Bedingung, dass deren Rao-Blackwellisierungen (unter einer für ϑ) suffizienten Statistik $V(\mathfrak{X})) f.s.$ übereinstimmen?

Für die Differenz zweier solcher Rao-Blackwellsisierungen $r(V(\mathfrak{X})), \rho(V(\mathfrak{X}))$

$$\triangle(V(\mathfrak{X})) := r(V(\mathfrak{X})) - \rho(V(\mathfrak{X}))$$

gilt

$$\forall \vartheta : \mathbb{E}_\vartheta \left[\triangle(V(\mathfrak{X})) \right] = m(\vartheta) - m(\vartheta) = 0.$$

Dies sollte erzwingen, dass

$$\forall \vartheta : \triangle(V(\mathfrak{X})) = 0 \quad \mathrm{Ws}_\vartheta - f.s.$$

Die gefragte Bedingung lautet also so:

Definition Eine Statistik $V(\mathfrak{X})$ heißt vollständig für ϑ, wenn für alle Zufallsvariablen der Form $\varphi(V(\mathfrak{X}))$ mit $\mathbb{E}_\vartheta \varphi(V(\mathfrak{X})) = 0 \quad \forall \vartheta$ gilt:

$$\forall \vartheta : \varphi(V(\mathfrak{X})) = 0 \quad \mathrm{Ws}_\vartheta - f.s.$$

Bemerkung: Sei $V(\mathfrak{X})$ nicht nur suffizient, sondern auch vollständig für ϑ, und sei $m(\vartheta)$ ein reelles Parametermerkmal. Dann stimmen die Rao-Blackwellisierungen von je zwei für $m(\vartheta)$ erwartungetreuen Schätzern überein.

Beispiel: Erfolgswahrscheinlichkeit p beim Münzwurf: $n \in \mathbb{N}$ sei fest

$$K_n := X_1 + \ldots + X_n \quad \text{ist vollständig für } p$$

In der Tat: Für ein $\varphi : \{0, \ldots, n\} \to \mathbb{R}$ gelte

$$0 \equiv \mathbb{E}_p \varphi(K_n) = \sum_{k=1}^{n} \varphi(k) \binom{n}{k} p^k (1-p)^{n-k}$$

Dann ist $\varphi \equiv 0$ (denn das einzige Polynom, das auf einem Intervall verschwindet, ist das Nullpolynom).

Anders formuliert:

Ist $b(n,p)$ der Vektor der Binomialgewichte, dann impliziert

$$\langle b(n,p),\, \varphi \rangle = 0 \quad \forall p$$

dass $\varphi \equiv 0$ ist.

In diesem Sinn bilden die Vektoren $b(n,p), 0 < p < 1$, ein vollständiges System im $\mathbb{R}^{n+1}$.

Allgemein nennt man eine Familie von linearen Funktionalen vollständig, wenn der Nullvektor der einzige Vektor ist, den sie gemeinsam annullieren.

Anstelle von "$V(\mathfrak{X})$ ist vollständig für ϑ" wäre es in diesem Sinn suggestiver zu sagen: "Θ ist vollständig für $V(\mathfrak{X})$" (oder noch genauer: Die Erwartungswerte $\mathbb{E}_\vartheta, \vartheta \in \Theta$ sind vollständig auf den Raum der $\varphi(V(\mathfrak{X}))$).

Definition S_1 und S_2 seien zwei erwartungstreue Schätzer für $m(\vartheta)$.

S_1 heißt wirksamer (eigentlich: nicht weniger wirksam) als S_2
$:\iff \mathrm{Var}_\vartheta(S_1) \leq \mathrm{Var}_\vartheta(S_2) \quad \forall \vartheta$.

Lemma: Für je zwei wirksamste erwarungstreue Schätzer S_1, S_2 für $m(\vartheta)$ gilt:

$$\mathrm{Ws}_\vartheta(S_1 = S_2) = 1 \quad \forall \vartheta.$$

Beweis:

$$\begin{aligned}
\mathbb{E}_\vartheta\left[(S_1 - S_2)^2\right] &= -\mathbb{E}_\vartheta\left[(S_1 + S_2)^2\right] + 2\mathbb{E}_\vartheta[S_1^2] + 2\mathbb{E}_\vartheta[S_2^2] \\
&= -4\mathbb{E}_\vartheta\left[\left(\frac{S_1+S_2}{2}\right)^2\right] + 4m(\vartheta)^2 + \\
&\quad + 2\left(\mathbb{E}_\vartheta[S_1^2] - m(\vartheta)^2\right) + 2\left(\mathbb{E}_\vartheta[S_2^2] - m(\vartheta)^2\right) \\
&= -4\mathrm{Var}_\vartheta \frac{S_1+S_2}{2} + 2\mathrm{Var}_\vartheta S_1 + 2\mathrm{Var}_\vartheta(S_2) \\
&\leq 0\,. \qquad \square
\end{aligned}$$

Satz (Lehmann-Scheffé)

Ist $V(\mathfrak{X})$ eine suffiziente und vollständige Statistik für ϑ, und ist $S^\star = \varphi(V(\mathfrak{X}))$ erwartungstreu für $m(\vartheta)$, so ist $S^\star$ ein wirksamster (und nach vorigem Lemma der wirksamste) erwartungstreue Schätzer für $m(\vartheta)$.

Beweis:

a) $S^\star$ ist Rao-Blackwellisierung eines erwartungstreuen Schätzers (nämlich von sich selbst).

b) Sei S irgendein erwartungstreuer Schätzer für $m(\vartheta)$. Nach obiger Bemerkung stimmt $S^\star$ mit der Rao-Blackwellisierung von S überein. Nach dem Satz von Rao-Blackwell ist diese aber wirksamer als S. $\square$

Gibt es übersichtliche Kriterien dafür, wann eine Statistik V suffizient und vollständig ist?

In gewissem Sinn sind die beiden Eigenschaften "gegenläufig":

Je "feiner" eine Statistik $V(\mathfrak{X})$ ist, um so leichter tut sie sich suffizient zu sein.

Je "gröber" sie ist um so leichter tut sie sich, vollständig zu sein.

Beispiel:

a) $V = id$ ist stets suffizient, aber i.a. nicht vollständig:

Θ sei einelementig, $\mathfrak{X}$ gleichverteilt auf $[-1, 1]$ unter $\mathrm{Ws}_\vartheta, g := id$

$$\mathbb{E}_\vartheta g(V(\mathfrak{X})) = 0 \text{ , aber } \quad g(V(\mathfrak{X})) \not\equiv 0 \quad \mathrm{Ws}_\vartheta f.s.$$

b) $V = \text{const}$ ist stets vollständig, aber i.a. nicht suffizient
denn $\quad \mathcal{L}_\vartheta(\mathfrak{X}\big|V) = \mathcal{L}_\vartheta(\mathfrak{X})$ hängt von ϑ ab?

Satz

$X_1, \ldots, X_n$ u.i.v., mit Dichten bzw. Gewichten folgender Form:

$$g_\vartheta(x) = K(\vartheta) e^{a(x) + \sum_{j=1}^{\ell} cj^{(\vartheta) V j(x)}} \ .$$

Umfasst die Menge

$$\{c_1(\vartheta), \ldots, c_\ell(\vartheta) \big| \vartheta \in \Theta\}$$

einen Quader mit positivem Volumen im $\mathbb{R}^\ell$, dann ist

$$V(\mathfrak{X}) := \left(\sum_{i=1}^{n} V_1(X_i), \ldots, \sum_{i=1}^{n} V_\ell(X_i) \right)$$

eine suffiziente und vollständige Statistik.

Beweis: Die Suffizienz folgt aus dem Faktorisierungssatz von Neyman. Zur Vollständigkeit siehe z.B. E.B. Lehmann, Testing Statistical Hypotheses, 2nd ed., Wadsworth 1991, p.142.

Beispiel: $\quad X_i$ unabhängig $\mathcal{N}\big(\mu, \sigma^2\big)$-verteilt

$$g_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{X^2}{2\sigma^2} + \frac{X\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}} \ .$$

Also ist $(X_1 + \ldots + X_n, X_1^2 + \ldots + X_n^2)$ (und damit auch $(\bar{X}, s_X^2)$) suffizient und vollständig für (μ, σ^2).

Kapitel 13

Die Cramér-Rao-Ungleichung

$\mathfrak{X}$ habe Verteilungsdichte (oder -gewichte) $f_\vartheta(x), \vartheta \in \Theta \subseteq \mathbb{R}$.

$\vartheta \mapsto f_\vartheta(x)$ sei hinreichend regulär

$S = h(\mathfrak{X})$ sei eine reellwertige Zufallsvariable mit $\operatorname{Var}_\vartheta S < \infty \quad \forall \vartheta$.

Sei $\psi(\vartheta) := \mathbb{E}_\vartheta[S], \quad I_f(\vartheta) := \mathbb{E}_\vartheta\left[\left(\frac{\partial}{\partial\vartheta}\log f_\vartheta(\mathfrak{X})\right)^2\right]$.

Dann gilt:

$$\operatorname{Var}_\vartheta S \geq \frac{\left(\psi'(\vartheta)\right)^2}{I_f(\vartheta)} \, .$$

Beweiskern (vgl. Bickel-Doksum (2001) p.181)

$$\begin{aligned}
\psi'(\vartheta) &= \int h(x)\frac{\partial}{\partial\vartheta} f_\vartheta(x)dx = \int h(x)\left(\frac{\partial}{\partial\vartheta}\log f_\vartheta(x)\right) f_\vartheta(x)dx \\
&= \mathbb{E}_\vartheta\left[h(\mathfrak{X})\frac{\partial}{\partial\vartheta}\log f_\vartheta(\mathfrak{X})\right] = \mathbb{E}_\vartheta\left[\left(h(\mathfrak{X}) - \psi(\vartheta)\right)\frac{\partial}{\partial\vartheta}\log f_\vartheta(\mathfrak{X})\right] \\
&\leq \sqrt{\operatorname{Var}_\vartheta[S]\,\mathbb{E}_\vartheta\left[\left(\frac{\partial}{\partial\vartheta}\log f_\vartheta(\mathfrak{X})\right)^2\right]} \qquad \square
\end{aligned}$$

Dabei wurde verwendet, dass $\mathbb{E}_\vartheta\left[\frac{\partial}{\partial\vartheta}\log f_\vartheta(\mathfrak{X})\right] = 0$. Die letzte Abschätzung folgt aus der Cauchy-Schwarz-Ungleichung: $\mathbb{E}[YZ] \leq \sqrt{\mathbb{E}Y^2}\sqrt{\mathbb{E}Z^2}$.

Bemerkung 1. Sei $\mathfrak{X} = (X_1, \ldots, X_n)$ mit $X_1, \ldots, X_n$ u.i.v., X_i habe Verteilungsdichte $g_\vartheta(x)$.

Dann gilt:

$$\begin{aligned}
I_f(\vartheta) &= -\mathbb{E}_\vartheta\left[\frac{\partial^2}{\partial\vartheta^2}\log f_\vartheta(\mathfrak{X})\right] \\
&= -n\mathbb{E}_\vartheta\left[\frac{\partial^2}{\partial\vartheta^2} g_\vartheta(X_1)\right] = nI_g(\vartheta)
\end{aligned}$$

Ist überdies S noch erwartungstreu für ϑ, so ist

$$\operatorname{Var}_\vartheta S \geq \frac{1}{nI_g(\vartheta)} \, .$$

In diesem Sinn ist also (in schönen Fällen) der ML-Schätzer "asymptotisch optimal".

Kapitel 14

Das Lageproblem für mehrere Stichproben

14.1 Die einfache Varianzanalyse

Angenommen, k unabhängige Stichproben kommen aus k Normalverteilungen mit ein- und derselben Varianz. Gibt es Grund, an der Hypothese zu zweifeln, dass alle Mittelwerte $\mu_1, \ldots, \mu_k$ übereinstimmen?

Fishers Idee war gleichermaßen einfach und scharfsinnig: Vergleiche die Variabilität innerhalb der Gruppen mit der zwischen den Gruppen. Wenn letzte übermäßig groß ist., dann gibt dies Grund, an der Gleichheit der μ_j zu zweifeln. Wie lässt sich diese Idee quantifizieren?

Die Stichworte sind "Quadratzerlegung der Variabilität", "Satz von Pythagoras" und (wieder einmal) Fishers Satz über die Projektionen eines standardnormalverteilten $\mathfrak{Z}$.

Zum Aufwärmen betrachten wir den Fall $k = 2$; hier finden wir Altbekanntes wieder.

Unser Modell ist:

$$X_{1i} = \mu_1 + \sigma Z_{1i} \ , \ i = 1, \ldots, n_1$$
$$X_{2i} = \mu_2 + \sigma Z_{2i} \ , \ i = 1, \ldots, n_2$$

mit Z_{ji} unabhängig, standard-normalverteilt.

In Vektorschreibweise:

$$\mathfrak{X} = \vec{\mu} + \sigma\mathfrak{Z}$$

mit $\mathfrak{Z}$ standard-normalverteilt auf $\mathbb{R}^n, n := n_1 + n_2$.

$\vec{\mu}$ wird als *(wahrer) Mittelwertsvektor* oder auch als *systematische Komponente* bezeichnet, $\sigma\mathfrak{Z}$ nennt man auch *Rauschen*.

$\vec{\mu}$ ist Element des *(2-dimensionalen linearen) Modellraums*

$$M := \left\{(\mu_1, \ldots, \mu_1, \mu_2, \ldots, \mu_2)^T \middle| \mu_1, \mu_2 \in \mathbb{R}\right\} \subseteq \mathbb{R}^n \ .$$

Den *(empirischen) Gesamtmittelwert*

$$\bar{X}_{..} := \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} X_{1i} + \sum_{i=1}^{n_2} X_{2i}\right)$$

und die *(empirischen) Gruppenmittelwerte*

$$\bar{X}_{j.} := \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ji} \quad , \quad j = 1,2$$

finden wir in den Projektionen von $\mathfrak{X}$ auf die *Diagonale*

$$D = \left\{(\mu,\ldots,\mu)\middle|\mu \in \mathbb{R}\right\} \subseteq \mathbb{R}^n$$

und auf M wieder:

$$\begin{aligned}\mathcal{P}_D\mathfrak{X} &= (\bar{X}_{..},\ldots,\bar{X}_{..})^T \\ \mathcal{P}_M\mathfrak{X} &= (\bar{X}_{1.},\ldots,\bar{X}_{1.},\bar{X}_{2.},\ldots,\bar{X}_{2.})^T \ .\end{aligned}$$

Die Abweichungen der Gruppenmittel vom Gesamtmittel finden sich in der Differenz $\mathcal{P}_M\mathfrak{X} - \mathcal{P}_D\mathfrak{X}$.

Was liegt nun näher als den Raum M orthogonal zu zerlegen in die zwei eindimensionalen Teilräume D und

$$E := \left\{(\beta_1,\ldots,\beta_1,\beta_2,\ldots,\beta_2)\middle|\beta_1,\beta_2 \in \mathbb{R}, n_1\beta_1 + n_2\beta_2 = 0\right\}.$$

Damit ergibt sich:

$$\mathcal{P}_E\mathfrak{X} = (\bar{X}_{1.} - \bar{X}_{..},\ldots,\bar{X}_{1.} - \bar{X}_{..},\bar{X}_{2.} - \bar{X}_{..},\ldots,\bar{X}_{2.} - \bar{X}_{..}) \ .$$

Nicht nur die Variabilität zwischen den Gruppen, sondern auch die in den Gruppen lässt sich aus einer Projektion ablesen, und zwar auf das orthogonale Komponent zu M in $\mathbb{R}^n$:

$$\begin{aligned}\mathcal{P}_{M^\perp}\mathfrak{X} &= \mathfrak{X} - \mathcal{P}_M\mathfrak{X} = \\ &= \left(X_{11} - \bar{X}_{1.},\ldots,X_{1n_1} - \bar{X}_{1.},X_{21} - \bar{X}_{2.},\ldots,X_{2n_2} - \bar{X}_{2.}\right).\end{aligned}$$

Also:

$$\begin{aligned}\|\mathcal{P}_{M^\perp}\mathfrak{X}\| &= \sqrt{\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_{1.})^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{X}_{2.})^2} \\ \|\mathcal{P}_E\mathfrak{X}\| &= \sqrt{n_1(\bar{X}_{1.} - \bar{X}_{..})^2 + n_2(\bar{X}_{2.} - \bar{X}_{..})^2} \ .\end{aligned}$$

Die Länge $\|\mathcal{P}_E\mathfrak{X}\|$ lässt sich noch eleganter berechnen:

Ein Einheitsbektor von E ist

$$e := \left(-\frac{1}{n_1},\ldots,-\frac{1}{n_1}\ ,\frac{1}{n_2},\ldots,\frac{1}{n_2}\right)^T \frac{1}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \ .$$

Damit:

$$\|\mathcal{P}_E\mathfrak{X}\| = \big|\,\langle e, \mathfrak{X}\rangle\,\big| = \frac{1}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\left(\bar{X}_{2\cdot} - \bar{X}_{1\cdot}\right) .$$

Wir sehen:

$$\frac{\langle e, \mathfrak{X}\rangle}{\frac{1}{\sqrt{n-2}}\|\mathcal{P}_{M^\perp}\mathfrak{X}\|}$$

ist nichts anderes als die gute alte Student-Statistik für zwei Stichproben.

Wenden wir uns nun der zu Beginn des Abschnitt beschriebenen Frage für k zu.

Unser Modell ist jetzt:

$$\begin{aligned} X_{1i} &= \mu_1 + \sigma Z_{1i}\ ,\ i = 1, \ldots, n_1 \\ X_{ki} &= \mu_k + \sigma Z_{ki}\ ,\ i = 1, \ldots, n_k \end{aligned}$$

mit Z_{ji} unabhängig, standard-normalverteilt.

Wir setzen $n := n_1 + \ldots + n_k$.

Das im Fall $k = 2$ beschriebene Vorgehen lässt sich geradewegs übertragen. Der Modellraum M ist jetzt k-dimensional.

$$M := \left\{(\mu_1, \ldots, \mu_1, \mu_2, \ldots, \mu_2, \ldots, \mu_k, \ldots, \mu_k)^T \big| \mu_1, \ldots \mu_k \in \mathbb{R}\right\} \subseteq \mathbb{R}^n .$$

D bleibt wie gehabt die Diagonale; der Raum E (das orthogonale Komplement von D zu M) ist $k-1$-dimensional:

$$E := \left\{(\beta_1, \ldots, \beta_1, \beta_2, \ldots, \beta_2, \ldots, \beta_k, \ldots, \beta_k)\big| \beta_1, \ldots \beta_k \in \mathbb{R}, n_1\beta_1 + \ldots + n_k\beta_k = 0\right\} .$$

Schreiben wir

$$\vec{\mu} = \mathcal{P}_D\vec{\mu} + \mathcal{P}_E\vec{\mu} = \vec{\alpha} + \vec{\beta}$$

so ergibt sich aus unserer Modellgleichung

$$\mathfrak{X} = \vec{\mu} + \sigma\mathfrak{Z}$$

die Beziehung

$$\begin{aligned} \mathfrak{X} &= \mathcal{P}_D\mathfrak{X} && +\mathcal{P}_E\mathfrak{X} && +\mathcal{P}_{M^\perp}\mathfrak{X} \\ &= \vec{\alpha} + \sigma\mathcal{P}_D\mathfrak{Z} && +\vec{\beta} + \sigma\mathcal{P}_E\mathfrak{Z} && +\sigma\mathcal{P}_{M^\perp}\mathfrak{Z} \end{aligned}$$

Unsere Frage war: Ist $\mu_1 = \mu_2 = \ldots = \mu_k$?

Oder mit anderen Worten: Ist $\vec{\beta} = 0$?

Dazu vergleichen wir (wie oben) $\|\mathcal{P}_E\mathfrak{X}\|$ mit $\|\mathcal{P}_{M^\perp}\mathfrak{X}\|$.

In unserem Modell gilt:

$$\begin{aligned} \|\mathcal{P}_{M^\perp}\mathfrak{X}\|^2 &= \sigma^2\|\mathcal{P}_{M^\perp}\mathfrak{Z}\|^2 \\ \|\mathcal{P}_E\mathfrak{X}\|^2 = \|\vec{\beta} + \sigma\mathcal{P}_E\mathfrak{Z}\|^2 &= \|\vec{\beta}\|^2 + 2\sigma \underbrace{\left\langle \vec{\beta}, \mathcal{P}_E\mathfrak{Z}\right\rangle}_{\mathcal{N}\left(0, \|\vec{\beta}\|^2\sigma^2\right) \text{ verteilt}} + \sigma^2\|\mathcal{P}_E\mathfrak{Z}\|^2 . \end{aligned}$$

Aus dem Satz von Fisher folgt:

Unter der Hypothese $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ ist

$$F := \frac{\frac{\|\mathcal{P}_E \mathfrak{X}\|^2}{k-1}}{\frac{\|\mathcal{P}_{M^\perp} \mathfrak{X}\|^2}{n-k}}$$

so verteilt wie der Quotient von zwei unabhängigen, durch die Anzahl ihrer Freiheitsgrade normierten $\chi^2(k-1)$ und $\chi^2(n-k)$ verteilten Zufallsvariablen.

Ist H_0 verletzt, so fällt F tendenziell größer aus
(wegen $\mathbb{E}(\|\mathcal{P}_E \mathfrak{X}\|^2) = \|\vec{\beta}\|^2 + \sigma^2(k-1)$).

Definition: Seien Y_1 $\chi^2(l_1)$-verteilt, Y_2 $\chi^2(l_2)$-verteilt und unabhängig. Dann heißt $\mathcal{L}\left(\frac{Y_1/l_1}{Y_2/l_2}\right)$ *Fisher-Verteilung* mit Parametern l_1 und l_2.

R kennt diese Verteilung als $f(., l_1, l_2)$.

Fassen wir zusammen:

Die zentrale Frage der Varianzanalyse (die eigentlich besser Variabilitätsanalyse heißen sollte) ist: "Wo steckt die Variabilität in den Daten? Nur in der zufälligen Komponente oder auch in der systematischen Komponente?"

Geometrisch steckt dahinter

- eine Quadratzerlegung von $\|\mathfrak{Y}\|^2$
- Verteilungseigenschaften von $\mathfrak{Z}$

Einen Überblick über die Ergebnisse der Varianzanalyse gibt die *ANOVA-Tafel* oder *Varianztafel*:

Quelle	Quadratsummen (SS)	Freiheitsgrade (DF)	Mittlere Quadratsumme $\left(\frac{\text{SS}}{\text{DF}}\right)$
(D) Mittelwert	$\|\mathcal{P}_D \mathfrak{Y}\|^2 = n \cdot \bar{Y}^2$	1	$\|\mathcal{P}_D \mathfrak{Y}\|^2$
(E) Gruppe	$\|\mathcal{P}_E \mathfrak{Y}\|^2$ $= \sum n_i \left(Y_{i.} - \bar{Y}_{..}\right)^2$	$k-1$	$\frac{\|\mathcal{P}_E \mathfrak{Y}\|^2}{k-1}$
$\left(M^\perp\right)$ Residuum	$\|\mathcal{R}\|^2 = \|\mathcal{P}_{M^\perp} \mathfrak{Y}\|^2$ $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i.}\right)^2$	$n-k$	$s^2 := \frac{\|\mathcal{R}\|^2}{n-k}$

s^2 ist wegen $\mathbb{E}\mathcal{R}^2 = \sigma^2 (n-k)$ ein erwartungstreuer Schätzer für σ^2.

Beispiel: Blutgerinnungszeit bei Ratten. Es gab vier unterschiedliche Behandlungsarten der Wunde (A,B,C,D).

Unterscheiden sich die vier Gruppen signifikant in Bezug auf ihre Lage?

A	62	60	63	59				
B	63	67	71	64	65	66		
C	68	66	71	67	68	68		
D	56	62	60	61	63	64	63	59

Y_{ij} = Gerinnungszeit der j-ten Ratte in der i-ten Gruppe

Gruppenmittel $\bar{Y}_{i\cdot}$:

A	61	61	61	61				
B	66	66	66	66	66	66		
C	68	68	68	68	68	68		
D	61	61	61	61	61	61	61	61

"Großes Mittel" $\bar{Y}_{\cdot\cdot}$

A	64	64	64	64				
B	64	64	64	64	64	64		
C	64	64	64	64	64	64		
D	64	64	64	64	64	64	64	64

$$
\begin{aligned}
(Y_{ij}) &= \mathcal{P}_D\mathfrak{Y} + \mathcal{P}_E\mathfrak{Y} + \overbrace{\mathcal{P}_{M^\perp}\mathfrak{Y}}^{=:\mathcal{R}} \\
&= \left(\bar{Y}_{\cdot\cdot}\right) + \left[\left(\bar{Y}_{i\cdot}\right) - \left(\bar{Y}_{\cdot\cdot}\right)\right] + \left[(Y_{ij}) - \left(\bar{Y}_{i\cdot}\right)\right] \\
&= \begin{pmatrix} 64 & 64 & 64 & 64 & & & & \\ 64 & 64 & 64 & 64 & 64 & 64 & & \\ 64 & 64 & 64 & 64 & 64 & 64 & & \\ 64 & 64 & 64 & 64 & 64 & 64 & 64 & 64 \end{pmatrix} \\
&\quad + \begin{pmatrix} -3 & -3 & -3 & -3 & & & & \\ 2 & 2 & 2 & 2 & 2 & 2 & & \\ 4 & 4 & 4 & 4 & 4 & 4 & & \\ -3 & -3 & -3 & -3 & -3 & -3 & -3 & -3 \end{pmatrix} \\
&\quad + \begin{pmatrix} 1 & -1 & 2 & 2 & & & & \\ -3 & 1 & 5 & -2 & -1 & 0 & & \\ 0 & -2 & 3 & -1 & 0 & 0 & & \\ -5 & 1 & -1 & 0 & 2 & 3 & 2 & -2 \end{pmatrix}
\end{aligned}
$$

Quelle	SS	DF	$\frac{\text{SS}}{\text{DF}}$
MW	$24 \cdot 64^2$	1	$24 \cdot 64^2$
Gruppe	228	$k - 1 = 3$	76
Residuum	112	$n - k = 20$	5,6

$$F = 13,6$$

$$
\begin{aligned}
Y_{ij} &= \mu_i + \sigma Z_{ij} \\
\mathfrak{Y} &= \vec{\mu} + \sigma\mathfrak{Z}
\end{aligned}
$$

Unter der Hypothese $\mu_1 = \mu_2 = \cdots = \mu_k$ ist

$$F := \frac{\frac{\|\mathcal{P}_E \mathfrak{Y}\|^2}{k-1}}{\frac{\left\|\mathcal{P}_{M^\perp} \mathfrak{Y}\right\|^2}{n-k}} = 13,6$$

Fisher$(3, 20)$-verteilt.

Mit einem p-Wert von $0.000\ldots$ ist damit die Hypothese zu verwerfen.

Wir illustrieren an unserem Beispiel der Blutgerinnungsdaten, wie man die einfache Varianzanalyse mit R durchführt.

Die Daten sind dabei in einer $n \times 2$ Matrix (im Textfile namens Blutgerinnung.dat) gespeichert, die Kopfzeile beinhaltet die Spaltennamen "Zeit" und "Gruppe".

```
> gerinnung<-read.table("Blutgerinnung.dat", header=T)
> gerinnung.aov<-aov(Zeit~Gruppe, data=gerinnung)
> summary(gerinnung.aov)
            Df Sum Sq Mean Sq F value    Pr(>F)
Gruppe       3  228.0    76.0  13.571 4.658e-05 ***
Residuals   20  112.0     5.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> model.tables(gerinnung.aov,type="means")
Tables of means
Grand mean

64

 Gruppe
     A  B  C  D
    61 66 68 61
rep  4  6  6  8
> model.tables(gerinnung.aov,type="effects")
Tables of effects

 Gruppe
     A B C  D
    -3 2 4 -3
rep  4 6 6  8
>
```

14.2 Paarweise Vergleiche

Wir gehen aus vom Modell des vorigen Abschnitts:

$$\mathfrak{X} = \vec{\mu} + \sigma \mathfrak{Z} \,,$$
$$\vec{\mu} \in M := \left\{ (\mu_1, \ldots, \mu_1, \ldots, \mu_k, \ldots, \mu_k)^T \middle| \mu_j \in \mathbb{R} \right\} \subseteq \mathbb{R}^n$$

Weil die Projektionen $\mathcal{P}_M\mathfrak{X}$ und $\mathcal{P}_{M^\perp}\mathfrak{X}$ unabhängig sind, ist für $j \neq l$

$$T_{jl} := \frac{1}{\sqrt{\frac{1}{n_j} + \frac{1}{n_l}}} \left(\bar{X}_{j.} - \bar{X}_{l.} - (\mu_j - \mu_l)\right) \Big/ s$$

Student$(n-k)$-verteilt (mit $s^2 := \|\mathcal{P}_{M^\perp}\mathfrak{X}\|^2/n-k$).

Also ist

$$I_{jl} := \left[\bar{X}_{j.} - \bar{X}_{l.} - \sqrt{\frac{1}{n_j} + \frac{1}{n_l}} s \cdot \mathrm{qt}\left(1 - \frac{a}{2}\right), \bar{X}_{j.} - \bar{X}_{l.} + \ldots\right]$$

ein $(1-\alpha)$-Konfidenzintervall für $\mu_j - \mu_l$.

Für ein fest gewähltes Paar (j,l) ist also

$$\mathrm{Ws}_{\vec{\mu}}\left(I_{jl} \ni \mu_j - \mu_l\right) = 1 - \alpha$$

Aber Achtung! Die Wahrscheinlichkeit, dass *alle* I_{jl} simultan die jeweiligen Differenzen $\mu_j - \mu_l$ überdecken, ist i.a. kleiner als $(1-\alpha)$.

Eine pessimistische Schranke (die sogenannte Bonferroni-Schranke) ist

$$\mathrm{Ws}_{\mu}\left(I_{jl} \ni \mu_j - \mu_l \text{ für alle Paare } (j,l)\right) \geq 1 - \binom{k}{2}\alpha$$

(es gibt $\binom{k}{2}$ Paare, verwende die Subadditivität der Wahrscheinlichkeit für die Nichtüberdeckungs-Ereignisse).

Um die Schranke $1-\alpha$ einzuhalten, müssen die Intervalle also wohl oder übel verlängert werden.

Betrachten wir den *balancierten Fall*

$$n_1 = \ldots = n_k =: m$$

Das pessimistische Rezept wäre dann, $\mathrm{qt}(1-\frac{\alpha}{2})$ durch $\mathrm{qt}\Big(1 - \frac{\alpha}{k(k-1)}\Big)$ zu ersetzen (um dem Faktor $\binom{k}{2}$ gegenzusteuern).

Aber geht's nicht kürzer? Sei

$$I_{(j,l),c} := \left[\bar{X}_{j.} - \bar{X}_{l.} - c\frac{s}{\sqrt{m}}\ , \bar{X}_{j.} - \bar{X}_{l.} + c\frac{s}{\sqrt{m}}\right]\ .$$

Wie groß ist c zu wählen, damit die Intervalle $I_{(j,l),c}$ simultane Konfidenzintervalle zum Niveau $(1-\alpha)$ für alle (j,l) werden?

$$\begin{aligned}
&\mathrm{Ws}_{\vec{\mu}}\left(\bar{X}_{j.} - \bar{X}_{l.} - c\frac{s}{\sqrt{m}} \leq \mu_j - \mu_l \leq \bar{X}_{j.} - \bar{X}_{l.} + c\frac{s}{\sqrt{m}} \quad \forall (j,l)\right) \\
&\quad = \mathrm{Ws}_{\vec{\mu}}\left(\max_{(j,l)} \frac{\sqrt{m}}{s}\Big|\bar{X}_{j.} - \mu_j - \left(\bar{X}_{l.} - \mu_l\right)\Big| \leq c\right) \\
&\quad = \mathrm{Ws}_{\vec{\mu}}\Bigg(\underbrace{\frac{\sigma}{s}\left(\max_j \frac{\sqrt{m}\left(\bar{X}_{j.} - \mu_j\right)}{\sigma} - \min_l \frac{\sqrt{m}\left(\bar{X}_{l.} - \mu_l\right)}{\sigma}\right)}_{\text{dessen Verteilung hängt nicht von } \vec{\mu} \text{ und } \sigma \text{ ab!}} \leq c\Bigg)
\end{aligned}$$

Definition (Tukey) $W_1, \ldots, W_k$ seien $\mathcal{N}(0,1)$-verteilt, Y sei $\chi^2(g)$-verteilt und die $W_1, \ldots, W_k, Y$ seien unabhängig.

Dann heißt

$$\mathcal{L}\left(\frac{\max_j W_j - \min_l W_l}{\sqrt{\frac{Y}{g}}}\right)$$

die *studentisierte Spannweitenverteilung (studentized range)* für k Mittelwerte und g Freiheitsgrade.

R kennt diese Verteilung als tukey$(., k, g)$.

Fazit: Für $c =$ qtukey $(1-\alpha, k, n-k)$ sind die $I_{(j,l),c}$ simultane Konfidenzintervalle für $\mu_j - \mu_l, 1 \leq j \neq l \leq k$, zum Niveau $1-\alpha$.

Bemerkung: Dem Auffinden von simultanen Konfidenzintervallen entspricht über die bekannte Dualität ein *multipler* Test:

Betrachte die folgende Familie von Tests für die Hypothese

$$H_{j,l} : \mu_j - \mu_l \quad :$$

Lehne $H_{j,l}$ ab, falls $\left|\bar{X}_{j\cdot} - \bar{X}_{l\cdot}\right| \geq$ qtukey $(1-\frac{\alpha}{2}, k, n-k)s/\sqrt{m}$.

Dann ist die Wahrscheinlichkeit, auch nur eine der zutreffenden Hypothesen irrtümlich abzulehnen, nicht größer als α.

Die Differenzen $\mu_j - \mu_l$ sind speziell lineare Funktionale von $\vec{\mu}$ der Form $\sum_{r=1}^k w_r \mu_r$ mit $\sum_{r=1}^k w_r = 0$. Derartige lineare Funktionale heißen auch **Kontraste**. Für

$$\vec{\mu} = \mathcal{P}_D \vec{\mu} + \mathcal{P}_E \vec{\mu} = \vec{\alpha} + \vec{\beta}$$

gilt dann:

$$\varphi := \sum_{r=1}^k w_r \mu_r = \sum_{r=1}^k w_r \beta_r.$$

Suchen wir nun noch (auf den Spuren von H. Scheffé) nach simultanen Konfidenzintervallen für alle Kontraste von $\vec{\mu}$.

Der Schätzer $\hat{\varphi} := \sum_{r=1}^k w_r \bar{X}_r$ ist normalverteilt mit MW $\sum_{r=1}^k w_r \mu_r$ und Varianz $\frac{\sigma^2}{m} \sum_{r=1}^k w_r^2 =: \frac{\sigma^2}{m} \tau(w)^2$.

Also ist

$$\left[\hat{\varphi} - \frac{s}{\sqrt{m}} \tau(w) \text{ qt } (1-\frac{\alpha}{2}, n-k), \hat{\varphi} + \ldots\right]$$

ein $(1-\alpha)$-Konfidenzintervall für φ.

Kann man ein C so wählen, dass

$$I_{w,K} := \left[\hat{\varphi} - \frac{s}{\sqrt{m}} \tau(w) C, \hat{\varphi} + \frac{s}{\sqrt{m}} \tau(w) C\right]$$

simultane $(1-\alpha)$-Konfidenzintervall für alle Kontraste φ sind?

Die Forderung ist gleichbedeutend mit

$$(\star)\ \mathrm{Ws}_{\vec{\mu}}\left[\hat{\varphi}-\frac{s}{\sqrt{m}}\tau(w)C\le\varphi\le\hat{\varphi}+\frac{s}{\sqrt{m}}\tau(w)C\quad\forall w\right]$$
$$=\mathrm{Ws}_{\vec{\mu}}\left[\max_w\frac{\sqrt{m}}{s}\frac{|\hat{\varphi}-\varphi|}{\tau(w)}\le C\right]=1-\alpha.$$

Nun gilt aber (mit $\hat{\beta}_r:=\bar{X}_{r\cdot}-\bar{X}_{\cdot\cdot}$) :

$$\left|\hat{\varphi}-\varphi\right|=\left|\sum_{r=1}^k w_r(\hat{\beta}_r-\beta_r)\right|$$
$$\le\underbrace{\sqrt{\sum_{r=1}^k w_r^2}}_{=\tau(w)}\sqrt{\sum_{r=1}^k(\hat{\beta}_r-\beta_r)^2}\,.$$

Die Gleichheit gilt genau dann, wenn (w_r) proportional zu $(\hat{\beta}_r-\beta_r)$ ist. Diese Wahl von (w_r) ist möglich, weil auch $(\hat{\beta}_r-\beta_r)$ im Raum

$$E_0:=\left\{(\beta_1,\ldots,\beta_k)^T\middle|\beta_1+\ldots+\beta_k=0\right\}\subseteq\mathbb{R}^k$$

liegt. Wir schreiben also $(\star)$ um zu

$$\mathrm{Ws}_{\vec{\mu}}\left(\frac{1}{s^2}m\sum_{r=1}^k(\hat{\beta}_r-\beta_r)^2\le C^2\right)$$
$$=\mathrm{Ws}_{\vec{\mu}}\left(\frac{\sigma^2\|\mathcal{P}_E\mathfrak{Z}\|^2}{\sigma^2\|\mathcal{P}_{M^\perp}\mathfrak{Z}\|^2/(n-k)}\le C^2\right)$$
$$=\mathrm{Ws}\left(Y\le\frac{C^2}{k-1}\right)$$

mit Y Fisher $(k-1,mk-k)$ verteilt.

Fazit: Für $C:=\sqrt{(k-1)\ \mathrm{qf}\ (1-\alpha,k-1,k(m-1))}$ sind die $I_{w,K}$, $w\in E_0$, simultane $(1-\alpha)$-Konfidenzintervalle für alle Kontraste.

Bemerkung: Jede einzelne der Bedingungen

$$\varphi\in I_{w,C}$$

legt einen Konfidenzbereich für $\beta=(\beta_1,\beta_2,\ldots,\beta_k)^T$ fest.

Es gilt nämlich:

$$\varphi\in I_{w,C}\iff\frac{\sqrt{m}}{s}\frac{|\hat{\varphi}-\varphi|}{\tau(w)}\le C$$
$$\iff\left|\left\langle\frac{w}{\tau(w)},\hat{\beta}-\beta\right\rangle\right|\le\frac{Cs}{\sqrt{m}}$$
$$\iff\hat{\beta}-\beta\in B_w$$

wobei $B_w \subseteq E_0$ durch die beiden auf w senkrecht stehenden Hyperebenen (jeweils mit Abstand $\frac{Cs}{\sqrt{m}}$ zum Ursprung) berandet wird.

Offenbar gilt:

$$\bigcap_{w \in E_0} B_w$$

ist der Schnitt von E_0 mit der Kugel (um 0) mit Radius $\frac{Cs}{\sqrt{m}}$ im $\mathbb{R}^k$.

14.3 Der Kruskal-Wallis-Test

Der Kruskal-Wallis-Test ist das Rangsummen-Analogon zur einfachen Varianzanalyse bzw. das Mehrstichproben-Analogon zum Wilcoxon-Test.

	Normalverteilungsmodell	Nichtparametrisch, "Rangsummen"
2 Stichproben	Student	Wilcoxon
k Stichproben	einfache Varianzanalyse	Kruskal-Wallis

Modell:

$$\left.\begin{array}{ll} X_{11}, \ldots, X_{1n_1} & \text{mit Verteilung } \pi_1 \\ \vdots & \\ X_{k1}, \ldots, X_{kn_k} & \text{mit Verteilung } \pi_k \end{array}\right\} \text{ unabhängig}$$

$n := n_1 + \ldots + n_k$.

Hypothese $H_0 : \pi_1 = \ldots = \pi_k$.

Wie lässt sich diese Hypothese testen?

Idee: Man ersetzt die Daten X_{ji} durch ihre Ränge R_{ji} in der Gesamtstichprobe und untersucht die Schwankungen der Gruppenmittelwerte $\bar{R}_{j\cdot}$ um das Gesamtmittel

$$\bar{R}_{\cdot\cdot} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{nj} R_{ji} = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2} \ .$$

Teststatistik:

$$K := \sum_{j=1}^{k} n_j \left(\bar{R}_{j\cdot} - \bar{R}_{\cdot\cdot} \right)^2 \ .$$

Unter H_0 ist (zumindest wenn die π_j als kontinuierlich vorausgesetzt werden) $(R_{11}, \ldots, R_{kn_k})$ eine zufällige Permutation von $(1, \ldots, n)$.

Tabellen der Verteilung von K (für $k = 3$ und $n_j \leq 5$) finden sich in E. Lehmann, Nonparametrics: Statistical methods based on ranks, Holden-Day, 1975.

Für größere Stichprobenumfänge (Lehmann, p.207, empfiehlt $n_j \geq 5$ für $k = 3$ oder $n_j \geq 4$ für $k \geq 3$) ist die folgende Normalapproximation brauchbar:

Satz: Für $n \to \infty$ und $n_j/n \to p_j > 0$ gilt unter H_0:

$$\frac{12}{n(n+1)} K \longrightarrow \chi^2(k-1) \quad \text{in Verteilung.}$$

Wir können diesen Satz verstehen über einen multivariaten Grenzwertsatz für Stichprobenmittel beim Ziehen ohne Zurücklegen:

Für $n \in \mathbf{N}$ sei

$$L_n = \left\{ y_1^{(n)}, \ldots, y_n^{(n)} \right\}$$

eine Liste der Länge n mit reellwertigen Einträgen $y_i^{(n)}$.

Sei $n = n_1 + \ldots + n_k$.

$(Y_{11}^{(n)}, \ldots, Y_{1n_1}^{(n)}, \ldots, Y_{kn_k}^{(n)})$ entstehe durch rein zufälliges Ziehen ohne Zurücklegen aus L_n. Wir setzen

$$S_j^{(n)} := \sum_{i=1}^{n_j} Y_{ji}^{(n)}.$$

Es gilt:

Für $j \neq l$ ist der Korrelationskoeffizient von $S_j^{(n)}$ und $S_l^{(n)}$ gleich

$$-\sqrt{\frac{n_j n_l}{(n-n_j)(n-n_j)}}\,.$$

Beweis: O.B.d.A. sei L_n zentriert, d.h. $\bar{y} = 0$.

Sei $Z_1, \ldots, Z_n$ eine rein zufällige Permutation von $\{1, \ldots, n\}$. Für $i \neq j$ ist

$$\mathrm{Cov}\left(y_{Z_i}, y_{Z_j}\right) = -\frac{\sigma_y^2}{n-1}$$

denn

$$0 = \mathrm{Var} \sum_{i=1}^{n} y_{Z_i} = n\sigma_y^2 + n(n-1)\mathrm{Cov}\left(y_{Z_i}, y_{Z_j}\right)$$

Also ist

$$\mathrm{Var}\left(\sum_{i=1}^{n_1} y_{Z_i}\right) = n_1\sigma_y^2 - n_1(n_1-1)\frac{\sigma_y^2}{n-1} = \sigma_y^2 \frac{n_1(n_1-1)}{n-1}$$

$$\mathrm{Cov}\left(\sum_{i=1}^{n_1} y_{Z_i}, \sum_{j=n_1+1}^{n_1+n_2} y_{Z_j}\right) = n_1 n_2 \mathrm{Cov}\left(y_{Z_1} y_{Z_2}\right) = -\sigma_y^2 \frac{n_1 n_2}{n-1}\,.$$

Für j, l gilt daher

$$\mathrm{Corr}\,\left(S_j^{(n)}, S_l^{(n)}\right) = \frac{1}{\sqrt{(1-n_j/n)(1-n_l/n)}}\left(\delta_{jl} - \sqrt{\frac{n_j}{n}\frac{n_l}{n}}\right)\,.$$

Hat man eine Folge von Listen L_n, bei denen nicht jeweils einige wenige Elemente als Ausreißer dominieren, dann sind die $(S_j^{(n)})_{j=1,\ldots,k}$ für $n \to \infty$ gemeinsam asymptotisch normal:

Satz (siehe E. Lehmann, loco citato, Thm 19 auf p.393)

Es gelte $\frac{n_j}{n} \overset{n\to\infty}{\longrightarrow} p_j$,

$$\frac{\max_i \left(y_i^{(n)} - \bar{y}^{(n)}\right)^2}{\sum_{i=1}^n \left(y_i^{(n)} - \bar{y}^{(n)}\right)^2} \cdot \max_j \left(\frac{n-n_j}{n_j}\right) \overset{n\to\infty}{\longrightarrow} 0\ .$$

Dann gilt:

$$\left(\frac{S_j^{(n)} - \mathbf{E}S_j^{(n)}}{\sqrt{\operatorname{Var} S_j^{(n)}}}\right)_{j=1,\ldots,k} \overset{n\to\infty}{\longrightarrow} \mathcal{N}\left(0, \left(\frac{1}{\sqrt{(1-p_j)(1-p_l)}}\left(\delta_{jl} - \sqrt{p_j p_l}\right)\right)_{j,l}\right)$$

in Verteilung.

Wir wenden diesen Satz an auf

$$L_n = (1, 2, \ldots, n).$$

Damit ist $\bar{R}_{j\cdot} = \frac{1}{n_j} S_j^{(n)}$.

Im Abschnitt über den Rangsummentext von Wilcoxon hatten wir hergeleitet:

$$\operatorname{Var} S_j^{(n)} = \frac{n_j(n-n_j)(n+1)}{12}\ .$$

Also ist

$$\operatorname{Var} \bar{R}_{j\cdot} = \frac{n(n+1)}{12}\frac{1}{n_j}\left(1 - \frac{n_j}{n}\right)\ .$$

Kombiniert mit obigem Zentralem Grenzwertsatz liefert das:

$$\left(\sqrt{\frac{12}{n(n+1)}}\sqrt{n_j}\left(\bar{R}_{j\cdot} - \bar{R}_{\cdot\cdot}\right)\right)_{j=1,\ldots,k} \to \mathcal{N}\left(0, \left(\delta_{jl} - \sqrt{p_j}\sqrt{p_l}\right)\right)\ .$$

Der Satz über die asymptotische Verteilung der Kruskal-Wallis-Statistik folgt nun unmittelbar aus dem Satz von Fisher! □

R gibt auf den Befehl kruskal.test als Wert der Teststatistik den Wert 12/(n(n+1)),K aus und berechnet den p-Wert grundsätzlich über die χ^2-Approximation. Hier ist unser kleines Beispiel:

```
> gerinnung<-read.table("Blutgerinnung.dat", header=T)
> attach(gerinnung)
> kruskal.test(Zeit, Gruppe)

    Kruskal-Wallis rank sum test

data:  Zeit and Gruppe
Kruskal-Wallis chi-squared = 17.0154, df = 3, p-value = 0.0007016
```

Zur Kontrolle:

```
> pchisq(17.0154,3,lower.tail=F)
[1] 0.000701607
```

Kapitel 15

Das normale lineare Modell

15.1 Schätzen von $\vec{\mu}$ und σ^2

M sei ein k-dimensionaler linearer Teilraum von $\mathbb{R}^n$.

Unser Modell ist:

$$\mathfrak{Y} = \vec{\mu} + \sigma\mathfrak{Z}$$

wobei $\vec{\mu} \in M, \sigma \geq 0, \mathfrak{Z}$ standard-normalverteilt auf $\mathbb{R}^n$.

Wie schätzt man $\vec{\mu}$ und σ^2 am besten?

Definition und Bemerkung

$$\begin{aligned}\mathfrak{M} &:= \mathcal{P}_M\mathfrak{Y} = \vec{\mu} + \sigma\mathcal{P}_M\mathfrak{Z} \\ \mathcal{R} &:= \mathcal{P}_{M^\perp}\mathfrak{Y} = \sigma\mathcal{P}_{M^\perp}\mathfrak{Z} \\ s^2 &:= \frac{\|\mathcal{R}\|^2}{n-k} = \sigma^2\frac{\|\mathcal{P}_{M^\perp}\mathfrak{Z}\|^2}{n-k}\ .\end{aligned}$$

$\mathfrak{M}$ und s^2 sind erwartungstreue Schätzer für $\vec{\mu}$ und σ^2.

Satz

(i) Für jedes $w \in \mathbb{R}^n$ ist $w^T \cdot \mathfrak{M}$ der wirksamste erwartungstreue Schätzer von $w^T\vec{\mu}$.

(ii) s^2 ist der wirksamste erwartungstreue Schätzer von σ^2.

Beweis: Nach dem Satz von Lehmann und Scheffé reicht es zu zeigen, dass $\mathfrak{M}$ und s^2 Funktionen einer suffizienten und vollständigen Statistik sind. Dazu gehen wir zu einer Darstellung in einer geeigneten ON-Basis über.

Sei $e_1, \ldots, e_k$ eine ON-Basis von M, ergänzt zu einer ON-Basis $e_1, \ldots, e_n$ von $\mathbb{R}^n$.

Wir schreiben

$$\begin{aligned}\mathfrak{Y} &= \sum_{i=1}^{n} U_i e_i \\ \vec{\mu} &= \sum_{i=1}^{n} \eta_i e_i \quad (\text{mit } \eta_{k+1} = \ldots = \eta_n = 0) \\ \vartheta &:= \left(\eta_1, \ldots, \eta_k, \sigma^2\right) .\end{aligned}$$

Die Verteilungsdichte von $(U_1, \ldots, U_n)$ ist

$$\begin{aligned}f_\vartheta(u_1, \ldots, u_n) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(u_i - \eta_i)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} u_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^{k}\eta_i u_i - \frac{1}{2\sigma^2}\sum_{i=1}^{k}\eta_i^2\right) \\ &=: \psi_\vartheta\left(\sum_{i=1}^{n} u_i, u_1, \ldots, u_k\right) .\end{aligned}$$

Mit dem Faktorisierungssatz von Neyman folgt:

$$\mathcal{S} := V(U_1, \ldots, U_n) := \left(U_1, \ldots, U_k, \sum_{i=1}^{n} U_i^2\right)$$

ist suffizient für ϑ.

Wegen eines in Abschnitt 12.4 zitierten Satzes über Exponentialverteilungen gilt: $\mathcal{S}$ ist auch vollständig für ϑ.

Es bleibt nur mehr festzustellen, dass sich sowohl $s^2 = \frac{1}{n-k}\sum_{i=1}^{k} U_i^2$ als auch $\mathfrak{M} = \sum_{i=1}^{k} U_i e_i$ als Funktionen von $\mathcal{S}$ schreiben lassen. □

Bemerkung 1 Umgekehrt lässt sich $\mathcal{S}$ aus $(\mathfrak{M}, s^2)$ berechnen. Damit ist $(\mathfrak{M}, s^2)$ eine suffiziente und vollständige Statistik für $(\vec{\mu}, \sigma^2)$.

Bemerkung 2

a) $\mathfrak{M}$ und $\mathcal{R}$ sind unabhängig (nach dem Satz von Fisher).

b) Wir wollen nun ein Konfidenzintervall für $w^T\vec{\mu}$ herleiten. Mit $w_M := \mathcal{P}_M w$ ist

$$w^T\vec{\mu} = w_M^T\vec{\mu} .$$

Wegen

$$w_M^T\mathfrak{Y} = w_M^T\vec{\mu} + \sigma w_M^T\mathfrak{Z}$$

hat $w_M^T\mathfrak{Y} = w^T\mathfrak{M}$ die Verteilung $\mathcal{N}\left(w^T\vec{\mu}, \sigma^2\|w_M^T\|^2\right)$.

Also ist $\frac{\left(w^T\mathfrak{M} - w^T\vec{\mu}\right)/\|w_M\|}{s}$ Student$(n-k)$-verteilt, und

$$\left[w^T\mathfrak{M} - s\|w_M\| \text{ qt}\left(1 - \frac{\alpha}{2}, n-k\right), w^T\mathfrak{M} + \ldots\right]$$

ist ein $(1-\alpha)$-Konfidentintervall für $w^T\vec{\mu}$.

Bemerkung 3 Weil $\mathfrak{M} = \mathcal{P}_M\mathfrak{Y}$ unter allen $\vec{m} \in M$ den Abstand zu $\mathfrak{Y}$, d.h. die Quadratsumme $\sum_{i=1}^n (Y_i - m_i)^2$, minimiert, heißt $\mathfrak{M}$ auch der **Kleinst-Quadrate-Schätzer** für $\vec{\mu}$.

15.2 Testen von linearen Hypothesen

Wir betrachten das Modell des vorigen Abschnitts, zusammen mit einem linearen Teilraum L unseres Modellraums M.

Wie testet man die Hypothese

$$(H_0) \qquad \vec{\mu} \in L \quad ?$$

M.a.W: wie bringt man zutage, ob in den Daten eine systematische Komponente "aus L heraus" steckt?

Sei E das orthogonale Komplement von L bezüglich M:

$$M = L \oplus E.$$

Wir vergleichen die Projektion von $\mathfrak{Y}$ auf E mit der auf $M^\perp$:

$$\begin{aligned} \mathcal{P}_E\mathfrak{Y} &= \mathcal{P}_E\vec{\mu} + \sigma\mathcal{P}_E\mathfrak{Z} \\ \mathcal{P}_{M^\perp}\mathfrak{Y} &= \sigma\mathcal{P}_E\mathfrak{Z} \ . \end{aligned}$$

Bemerkung: Unter H_0 ist (mit $\dim M = k, \dim L = r$)

$$F := \frac{\|\mathcal{P}_E\mathfrak{Y}\|^2/(k-r)}{\|\mathcal{P}_{M^\perp}\mathfrak{Y}\|^2/(n-k)}$$

Fisher $(k-r, n-k)$-verteilt. Unter der Alternative $\vec{\mu} \in M \setminus L$ trägt die Komponente $\mathcal{P}_E\vec{\mu}$ zu einer systematischen Vergrößerung des Zählers bei.

Die Frage "wie groß ist der F-Wert" kann man übrigens umdeuten zu der Frage "Ein wie großer Anteil von $\|\mathcal{P}_{L^\perp}\mathfrak{Y}\|^2$ wird durch die Projektion auf M erklärt?

In der Tat: Schreiben wir $\mathfrak{V} := \mathcal{P}_{L^\perp}\mathfrak{Y}$.

Dann ist wegen $\mathcal{P}_E = \mathcal{P}_M\mathcal{P}_{L^\perp}$ und $\mathcal{P}_{M^\perp} = \mathcal{P}_{M^\perp}\mathcal{P}_{L^\perp}$

$$\begin{aligned} \frac{\|\mathcal{P}_E\mathfrak{Y}\|^2}{\|\mathcal{P}_{M^\perp}\mathfrak{Y}\|^2} = \frac{\|\mathcal{P}_M\mathfrak{V}\|^2}{\|\mathcal{P}_{M^\perp}\mathfrak{V}\|^2} &= \frac{\|\mathcal{P}_M\mathfrak{V}\|^2}{\|\mathfrak{V}\|^2 - \|\mathcal{P}_M\mathfrak{V}\|^2} \\ &= \frac{R^2}{1-R^2} \ , \end{aligned}$$

mit $R := \frac{\|\mathcal{P}_M\mathfrak{V}\|}{\|\mathfrak{V}\|}$.

Zwischen F und R^2 ergibt sich somit die Beziehung

$$F = \frac{n-k}{k-r}\frac{R^2}{1-R^2} \ .$$

R lässt sich umschreiben zu

$$R = \frac{\|\mathcal{P}_M \mathcal{U}\|^2}{\|\mathcal{U}\| \, \|\mathcal{P}_M \mathcal{U}\|} = \frac{\langle \mathcal{U}, \mathcal{P}_M \mathcal{U} \rangle}{\|\mathcal{U}\| \, \|\mathcal{P}_M \mathcal{U}\|} \; .$$

Ausgedrückt durch $\mathfrak{Y}$ und $\mathfrak{M} := \mathcal{P}_M \mathfrak{Y}$ ist dies (wegen $\mathcal{P}_M \mathcal{P}_{L^\perp} = \mathcal{P}_{L^\perp} \mathcal{P}_M$)

$$R = \frac{\langle \mathcal{P}_{L^\perp} \mathfrak{Y}, \mathcal{P}_{L^\perp} \mathfrak{M} \rangle}{\|\mathcal{P}_{L^\perp} \mathfrak{Y}\| \, \|\mathcal{P}_{L^\perp} \mathfrak{M}\|} \; .$$

Besonders schön lässt sich R deuten, wenn unser linearer Teilraum gleich der Hauptdiagonalen D ist.

Dann gilt nämlich (mit $\mathfrak{M} = (M_1, \ldots, M_n)^T$)

$$R = \frac{\sum_{i=1}^n \left(Y_i - \bar{Y}\right) \left(M_i - \bar{M}\right)}{\sqrt{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2 \sum_{i=1}^n \left(M_i - \bar{M}\right)^2}} \; .$$

Also ist in diesem Fall R nichts anderes als der Stichprobenkorrelationskoeffizient von $\mathfrak{Y}$ und $\mathfrak{M} = \mathcal{P}_M \mathfrak{Y}$.

Dieser heißt auch *multipler Korrelationskoeffizient* des durch M beschriebenen Modells.

Kapitel 16

Das lineare Modell mit weißem Rauschen

Unser Modell ist jetzt:

$$Y_i = \mu_i + W_i \quad , \quad i = 1, \ldots, n \, ,$$

wobei wie gehabt $\vec{\mu} = (\mu_1, \ldots, \mu_n)^T$ in einem k-dimensionalen linearen Teilraum $M(\subseteq \mathbb{R}^n)$ liegt, und $W_1, \ldots, W_n$ zentrierte, unkorrelierte Zufallsvariablen mit ein- und derselben Varianz σ^2 seien.

Definition $\mathfrak{W} = (W_1, \ldots, W_n)^T$ heißt *weißes Rauschen* auf $\mathbb{R}^n$ mit Varianz $\sigma^2 : \Longleftrightarrow$

$$\mathbb{E}\mathfrak{W} = 0 \quad , \quad \mathbf{Var}\mathfrak{W} = \sigma^2 I \, .$$

Dabei bezeichnen wir hier und im Folgenden die Kovarianzmatrix eines Zufallsvektors $\mathfrak{V}$ mit $\mathbf{Var}\, \mathfrak{V}$.

Die schönsten Eigenschaften in normalen linearen Modell hatten wir dem Satz von Fisher zu verdanken. Was kann man davon retten?

Lemma Sei $\mathfrak{W}$ ein weißes Rauschen auf $\mathbb{R}^n$ mit Varianz σ^2.

a) Ist $\mathfrak{O}$ eine orthogonale Matrix, so ist auch $\mathfrak{O} \cdot W$ ein weißes Rauschen.

b) Die Koordinaten von $\mathfrak{W}$ in jeder ON-Basis von $\mathbb{R}^n$ bilden wieder ein weißes Rauschen.

c) Für zwei orthogonale Unterräume L, N von $\mathbb{R}^n$ ist jede Koordinate von $\mathcal{P}_L\mathfrak{W}$ zu jeder Koordinate von $\mathcal{P}_N\mathfrak{W}$ unkorreliert.

d) Für jeden Unterraum $L \subseteq \mathbb{R}^n$ gilt

$$\mathbb{E}\|\mathcal{P}_L\mathfrak{W}\|^2 = \dim L \cdot \sigma^2 \, .$$

Beweis

a) $\mathbb{E}\mathfrak{O}\mathfrak{W} = 0$, $\mathbf{Var}\, \mathfrak{O}\mathfrak{W} = \sigma^2 \mathfrak{O} I \mathfrak{O}^T = \sigma^2 I$.

b) Sei $e_1, \ldots, e_n$ eine ON-Basis von $\mathbb{R}^n$, $U_i := e_i^T \mathfrak{W}$. Dann gilt:

$$\begin{aligned} \mathbb{E}U_i = 0 \ , \ \mathrm{Cov}(U_i, U_j) &= \mathbb{E}\left[e_i^T \mathfrak{W}\mathfrak{W}^T e_j\right] \\ &= e_i^T \mathbb{E}\left[\mathfrak{W}\mathfrak{W}^T\right] e_j = \sigma^2 \delta_{ij} \ . \end{aligned}$$

c) Es reicht, die ON-Basis in b) verträglich mit L und N zu wählen.

d) Sei $e_1, \ldots, e_n$, $U_1, \ldots, U_n$ wie in b) und c) mit $e_1, \ldots, e_l$ ON-Basis von L. Dann ist

$$\|\mathcal{P}_L \mathfrak{W}\|^2 = U_1^2 + \ldots + U_l^2 \ ,$$

dies hat Erwartungswert $\ell \cdot \sigma^2$. □

Kehren wir zurück zu unserem Modell

$$\mathfrak{Y} = \vec{\mu} + \mathfrak{W} \ .$$

Wie früher schreiben wir

$$\mathfrak{M} := \mathcal{P}_M \mathfrak{Y} \ , \ \mathcal{R} := \mathcal{P}_{M^\perp} \mathfrak{Y} \ , \ s^2 := \frac{\|\mathcal{R}\|^2}{n-k} \ .$$

Aus Teil d) des Lemmas folgt sofort die

Bemerkung: s^2 ist ein erwartungstreuer Schätzer für σ^2. Darüber hinaus gilt der

Satz (von Gauß-Markov):

Sei $g \in \mathbb{R}^n$ fest. Dann ist $g^T \mathfrak{M}$ der wirksamste unter allen in $\mathfrak{Y}$ linearen erwartungstreuen Schätzern für das lineare Parametermerkmal $g^T \vec{\mu}$.

Beweis: Sei (mit $h \in \mathbb{R}^n$) $h^T \mathfrak{Y}$ ein weiterer linearer erwartungstreuer Schätzer für $g^T \vec{\mu}$. Die Erwartungstreue besagt:

$$\mathbb{E}_{\vec{\mu}} h^T \mathfrak{Y} = g^T \vec{\mu} \quad \forall \vec{\mu} \in M \ .$$

Andererseits ist nach Modellannahme

$$\mathbb{E}_{\vec{\mu}} h^T \mathfrak{Y} = \mathbb{E}_{\vec{\mu}} h^T (\vec{\mu} + \mathfrak{W}) = h^T \vec{\mu} \quad \forall \vec{\mu} \in M \ .$$

Aus den beiden letzten Beziehungen zusammen folgt

$$g^T m = h^T m \quad \forall m \in M \ . \tag{16.1}$$

Wir zeigen nun, dass $h^T \mathfrak{Y}$ nicht wirksamer ist als $g^T \mathfrak{M}$:

$$\begin{aligned} \mathrm{Var}\, h^T \mathfrak{Y} &= \mathrm{Var}\,(h^T \mathfrak{M} + h^T \mathcal{R}) \\ &= \mathrm{Var}\,(h^T \mathfrak{M}) + 2\mathrm{Cov}(h^T \mathfrak{M}, h^T \mathcal{R}) + \mathrm{Var}\,(h^T \mathcal{R}) \\ &= \mathrm{Var}\,(g^T \mathfrak{M}) + 2\mathrm{Cov}\left(h^T \mathcal{P}_M \mathfrak{W}, h^T \mathcal{P}_{M^\perp} \mathfrak{W}\right) + \mathrm{Var}\,(h^T \mathcal{R}) \ . \end{aligned}$$

Dabei haben wir (16.1) sowie unsere Modellgleichung verwendet. Der mittlere Term der rechten Seite verschwindet aber wegen Teil c) des Lemmas, und der dritte Term ist nichtnegativ. □

Kapitel 17

Lineare Regression

17.1 Vom linearen Modell zum linearen Regressionsmodell

Sei M wie gehabt ein k-dimensionaler Teilraum des $\mathbb{R}^n$.

Wir betrachten eine lineare Parametrisierung von M. Dazu seien $c_1, \ldots, c_l$ Vektoren im $\mathbb{R}^n$ mit

$$\begin{aligned} M &= \left\{ \beta_1 c_1 + \ldots + \beta_l c_l \middle| \beta := (\beta_1, \ldots, \beta_l)^T \in \mathbb{R}^l \right\} \\ &= \left\{ C \cdot \beta \middle| \beta \in \mathbb{R}^l \right\} \end{aligned}$$

mit der $n \times l$-Matrix $C := (c_1, \ldots, c_l)$.

Es gilt jedenfalls: $\text{Rang}(C) = k \leq l$.

C heißt Systemmatrix, die β_j heißen Regressionskoeffizienten.

Sei $\mathfrak{y}$ ein Datenvektor im $\mathbb{R}^n$.

Die Kleinst-Quadrate-Schätzer für β ist dann gegeben durch die Gleichung

$$\mathcal{P}_M \vec{\mathfrak{y}} = C\hat{\beta} \,. \tag{17.1}$$

(17.1) ist gleichbedeutend zur Forderung, dass die Differenz von $\vec{\mathfrak{y}}$ und $C\hat{\beta}$ orthogonal steht auf ganz M, m.a.W. dass gilt:

$$(\vec{\mathfrak{y}} - C\hat{\beta})^T C\beta = 0 \quad \text{für alle } \beta \in \mathbb{R}^l \,.$$

Daraus ergibt sich folgende Kette von Äquivalenzen:

$$\begin{aligned} (17.1) &\iff (\vec{\mathfrak{y}} - C\hat{\beta})^T C = 0 \\ &\iff \vec{\mathfrak{y}}^T C = \hat{\beta}^T C^T C \\ &\iff C^T \vec{\mathfrak{y}} = C^T C\hat{\beta} \end{aligned} \tag{17.2}$$

Die Gleichungen (17.2) heißen die *Normalgleichungen* im Regressionsmodell. Offenbar gilt:

(17.1) ist eindeutig lösbar $\iff$ $c_1, \ldots, c_l$ linear unabhängig

(17.2) ist eindeutig lösbar $\iff$ $C^T C$ ist regulär.

Dies ist genau dann der Fall, wenn $k = l$ ist, d.h. wenn man nicht überparametrisiert hat.

Beispiel 1. Die einfache Varianzanalyse (schon mit $k = 2$) ist ein Beispiel eines "überparametrisierten" Regressionsmodells:

$$M = \left\{ \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_2 \end{pmatrix} \right\} = \left\{ \beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \beta_2 \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right\}.$$

Hier ist C eine $n \times 3$-Matrix mit Rang 2.

Beispiel 2. (Die klassische lineare Regression)

$x_1, \ldots, x_n$ seien *feste* reelle Zahlen

$$C := \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

$$\beta = (\beta_0, \beta_1)^T .$$

Modell: $Y_i = \beta_0 + \beta_1 x_i +$ Rauschen.

Die rechte Seite von (17.1), d.h. der Vektor $C\hat{\beta}$, nimmt jetzt folgende Gestalt an:

$$C\hat{\beta} = \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right)_{i=1,\ldots,n}$$

$\hat{\mathfrak{y}} - C\hat{\beta}$ muss orthogonal stehen auf M; dieses aber wird aufgespannt von den beiden Vektoren $(1, \ldots, 1)^T$ und $(x_1, \ldots, x_n)^T$.

Also ergeben sich die beiden Bedingungen:

(1) $\hat{\mathfrak{y}} - C\hat{\beta} \perp (1, \ldots, 1)^T$:

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \Rightarrow \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$$

d.h. $(\bar{x}, \bar{y})$ liegt auf der Geraden $y = \hat{\beta}_0 - \hat{\beta}_1 x$.

(2) $\hat{\mathfrak{y}} - C\hat{\beta} \perp (x_1, \ldots, x_n)^T$

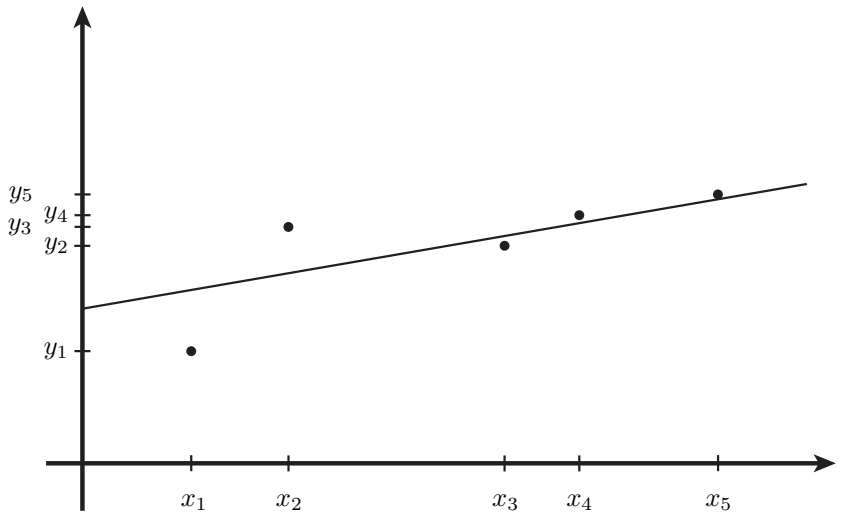
$$\begin{aligned} 0 &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i \\ &\overset{(1)}{=} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) (x_i - \bar{x}) \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 x_i \right) (x_i - \bar{x}) \\ &\overset{(1)}{=} \sum_{i=1}^{n} \left((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right) (x_i - \bar{x}) \end{aligned}$$

also

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$\hat{\beta}_1$ wird auch als Regressionskoeffizient von (y_i) nach (x_i) bezeichnet.

Die Gerade $y = \hat{\beta}_0 + \hat{\beta}_1 x$ heißt Regressionsgerade.



$\hat{\beta}_0$ und $\hat{\beta}_1$ lösen *auch* das Problem

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 = \min !$$

Es besteht ein bemerkenswerter Zusammenhang zwischen dem Regressions- und dem Korrelationskoeffizienten

$$\hat{\beta}_1 = \kappa_{\vec{\mathfrak{x}}\vec{\mathfrak{y}}} \frac{\sigma_{\mathfrak{y}}}{\sigma_{\mathfrak{x}}}$$

(dabei ist $\sigma_{\mathfrak{x}}^2 := \frac{1}{n} (x_i - \bar{x})^2$ die Stichprobenvarianz von $\vec{\mathfrak{x}}$.

Gehen wir zu den *standardisierten Werten* über

$$u_i := \frac{x_i - \bar{x}}{\sigma_{\mathfrak{x}}} \quad , \quad \upsilon_i := \frac{y_i - \bar{y}}{\sigma_{\mathfrak{y}}} \ ,$$

so ändert sich der Korrelationskoeffizient nicht:

$$\kappa_{\mathfrak{u},\mathfrak{v}} = \kappa_{\mathfrak{x}\mathfrak{y}} \ .$$

Der Regressionskoeffizient von (υ_i) nach (u_i) stimmt mit $\kappa_{u\upsilon}$ überein (und ist somit verschieden vom Regressionskoeffizienten von (y_i) nach (x_i)).

Hier ist ein Sachverhalt zu erwähnen, der auf den ersten Blick paradox klingt und der von Sir Francis Galton im 19. Jhdt. entdeckt wurde.

Wir betrachten n Paaren (x_i, y_i) mit positiver Korrelation $\kappa < 1$; die Streuung der (x_i) sei gleich der von (y_i).

Dann hat die Regressionsgerade die Steigung κ. Die "beste Vorhersage" von $x - \bar{x}$ im Sinn der kleinsten Quadrate ist also $\kappa(x - \bar{x})$; wegen $\kappa < 1$ bedeuted dies eine "regression towards the mean", ein "Zurückgehen Richtung Mittelwert" (diese Sicht ist übrigens die historische Wurzel des Names "Regression"). So weit, so gut.

Nun ist aber $\kappa_{\mathfrak{x}\mathfrak{y}} = \kappa_{\mathfrak{y}\mathfrak{x}}$, und deshalb hat auch die Regressionsgerade von $\mathfrak{x}$ bzgl. $\mathfrak{y}$ wieder dieselbe Steigung < 1! Plastisch wird dies an folgendem Beispiel: Wir betrachten Paare von Vätern und Söhnen aus einer Population, x_i sei die Körpergröße des i-ten Vaters, y_i die seines Sohnes. Angenommen, die obigen Voraussetzungen (positive, aber nicht perfekte Korrelation von $\vec{\mathfrak{x}}$ und $\vec{\mathfrak{y}}$, gleiche Streung der (x_i) und der (y_i)) sei erfüllt. Wenn wir die Körpergröße des Sohnes eines überdurchschnittlich großen Vaters vorhersagen sollen, kommen wir auf exakt dieselbe "regression towards the mean", wie wenn wir die Körpergröße des Vaters eines überdurchschnittlich großen Sohnes vorhersagen.

Hübsche Konturen gewinnt der eben beschriebene Sachverhalt auch im Licht der zweidimensionalen Normalverteilung:

U, V seien gemeinsam normalverteilt auf $\mathbb{R}^2$,

mit $\mathbb{E}U = \mathbb{E}V = 0 \quad , \quad \operatorname{Var} U = \operatorname{Var} V = 1$,

$\operatorname{Cov}(U, V) := \rho \in (0, 1)$.

Die Kovarianzmatrix von (U, V) ist also

$$K := \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} .$$

Mit (Z_1, Z_2) standard-normalverteilt auf $\mathbb{R}^2$ gilt

$$\mathcal{L}(U, V) = \mathcal{L}\left(Z_1, \rho Z_1 + \sqrt{1 - \rho^2} Z_2\right) \ ,$$

denn auch $\left(Z_1, \rho Z_1 + \sqrt{1 - \rho^2} Z_2\right)$ ist zentriert normalverteilt mit Kovarianzmatrix K.

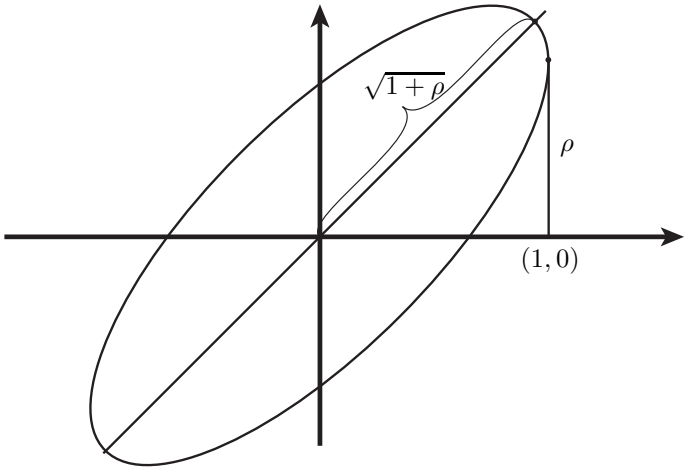
Die bedingte Verteilung von V gegeben U ist daher $\mathcal{N}\big(\rho U, 1-\rho^2\big)$, und für die bedingte Erwartung folgt:

$$\mathbb{E}\left[V\middle|U\right] = \rho U \ .$$

Genau so ist übrigens aus Symmetriegründen

$$\mathbb{E}\left[U\middle|V\right] = \rho V \ .$$

Dazu passt das folgende Bild:



Sei g die gemeinsame Verteilungsdichte von (U,V). Aus unserer Bemerkung zur bedingten Verteilung von V gegeben U folgt: Der Schnitt von g entlang $u=1$ ist eine Gauß'sche Glockenkurve mit Maximum bei $v=\rho$. Weil somit $g(u,v)$ unter der Nebenbedingung $u=1$ sein Maximum bei $v=\rho$ annimmt, muss die Gerade $u=1$ die Niveaulinie von g durch $(1,\rho)$ im Punkt $(1,\rho)$ berühren.

Die Niveaulinie von g sind von der Form

$$(u,v)K^{-1}\begin{pmatrix}u\\v\end{pmatrix} = r^2 \ , \tag{17.3}$$

mit

$$K^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix}1 & -\rho\\ -\rho & 1\end{pmatrix} \ .$$

Dies sind Ellipsen, deren Achsen in Richtung von Haupt- und Nebendiagonalen weisen.

Für $(u,v)=(1,\rho)$ ergibt sich $r=1$.

Die Achsenlängen der zugehörigen Ellipse (setze $r=1$ und $u=v$ bzw. $u=-v$ in (17.3)) ergeben sich als $\sqrt{1+\rho}$ bzw. $\sqrt{1-\rho}$.

In der Tat sind $1+\rho$ und $1-\rho$ die beiden Eigenwerte von K.

Bemerkung In Verallgemeinerung unserer obigen Aussage

$$\mathbb{E}\left[V\middle|U\right] = \rho U$$

ergibt sich (Beweis durch Standardisieren!)

Ist (X, Y) gemeinsam normalverteilt auf $\mathbb{R}^2$ mit Korrelation ρ, so ist

$$\mathbb{E}\left[Y\middle|X\right] = \mu_y + \rho\frac{\sigma_y}{\sigma_x}\left(X - \mu_x\right) \ .$$

17.2 Lineares Regressionsmodell mit Weißem Rauschen

Wir betrachten dasselbe Modell wie im vorigen Abschnitt,

$$\mathfrak{Y} = \vec{\mu} + \mathfrak{W} \quad , \quad \vec{\mu} \in M \ ,$$

wobei der k-dimensionale Unterraum M des $\mathbb{R}^n$ (genau wie im Abschnitt 17.1) von folgender Form ist:

$$M = \left\{C \cdot \beta \middle| \beta \in \mathbb{R}^\ell\right\}$$

mit einer festen $n \times \ell$-Matrix C.

Wieder sind die Normalgleichungen für den Kleinst-Quadrate-Schätzer:

$$C^T C\hat{\beta} = C^T\mathfrak{Y} \ .$$

Wir setzen im Folgenden voraus dass C vollen Rang hat (d.h. $\ell = k$). Dann gilt:

$$\begin{aligned}\hat{\beta} &= \left(C^T C\right)^{-1} C^T\mathfrak{Y} \\ &= \left(C^T C\right)^{-1} C^T(C\beta + \mathfrak{W}) \\ &= \beta + \left(C^T C\right)^{-1} C^T\mathfrak{W}.\end{aligned}$$

Also gilt insbesondere der

Satz: Hat C vollen Rang, dann ist der Kleinst-Quadrate-Schätzer $\hat{\beta}$ für β erwartungstreu und hat die Kovarianzmatrix $\sigma^2\left(C^T C\right)^{-1}$.

Beweis:

$$\begin{aligned}\mathbf{Var}\hat{\beta} &= \left(C^T C\right)^{-1} C^T\left(\sigma^2 I\right) C\left(C^T C\right)^{-1} \\ &= \sigma^2\left(C^T C\right)^{-1}. \qquad \square\end{aligned}$$

Definition: $J := \frac{1}{n}\frac{1}{\sigma^2}C^T C$ heißt *Informationsmatrix* des linearen Regressionsmodells.

Es gilt:

$$n\mathbf{Var}\hat{\beta} = J^{-1} \ .$$

Wir klären gleich einmal die Beziehung zwischen J und der Fisher-Informationsmatrix im normalen linearen Regressionsmodell. In diesem gilt:

$$\mathcal{L}(\mathfrak{Y}) = \mathcal{N}\left(C \cdot \beta, \sigma^2 I\right) \ .$$

Sei f_β die Dichte von $\mathcal{N}\big(C \cdot \beta, \sigma^2 I\big)$. Der Gradient der log-likelihood ist:

$$\begin{aligned}\frac{\partial}{\partial\beta}\log f_\beta(\mathfrak{Y}) &= \frac{\partial}{\partial\beta}\left(-\frac{1}{\sigma^2}\|\mathfrak{Y} - C\beta\|^2\right)\\ &= -\frac{1}{\sigma^2}2(\mathfrak{Y} - C\beta)^T(-C) = \frac{1}{\sigma}\mathfrak{Z}^T C\ .\end{aligned}$$

Die Fisher-Informationsmatrix ist somit

$$\begin{aligned}&\mathbb{E}_\beta\left[\frac{\partial}{\partial\beta}\log f_\beta(\mathfrak{Y})^T\frac{\partial}{\partial\beta}\log f_\beta(\mathfrak{Y})\right]\\ &= \mathbb{E}_\beta\left[\frac{1}{\sigma^2}C^T\mathfrak{Z}\mathfrak{Z}^T C\right] = \frac{1}{\sigma^2}C^T C = nJ.\end{aligned}$$

17.3 Normales lineares Regressionsmodell: Konfidenzellipsoid für β

Im normalen linearen Regressionsmodell

$$\mathfrak{Y} = C\beta + \sigma\mathfrak{Z}$$

hat (wegen des Satzes aus dem vorigen Abschnitt) der KQ-Schätzer $\hat\beta$ die Verteilung

$$\mathcal{N}\big(\beta, \sigma^2(C^T C)^{-1}\big) = \mathcal{N}\big(\beta, (nJ)^{-1}\big)\ .$$

(Wieder setzen wir voraus, dass C vollen Rang hat.)

Also gilt für jedes β:

$$(\hat\beta - \beta)^T nJ(\hat\beta - \beta) \quad \text{ist} \quad \chi^2(\ell)\text{-verteilt.} \tag{17.4}$$

Daraus läßt sich (zumindest für bekanntes σ^2) ein Konfidenzbereich für β gewinnen.

Wir setzen

$$\mathcal{E}(J, r) := \big\{z \in \mathbb{R}^\ell \big| z^T Jz \le r^2\big\}\ .$$

Satz: Sei $\gamma \in (0,1)$. Bei gegebenem σ^2 ist

$$B(\mathfrak{Y}) := \hat\beta + \frac{1}{\sqrt{n}}\mathcal{E}\left(J, \sqrt{\text{qchisq}(\gamma,\ell)}\right)$$

ein Konfidenzellipsoid für β zum Niveau γ.

Beweis:

$$\begin{aligned}B(\mathfrak{Y}) \ni \beta &\iff \sqrt{n}(\hat\beta - \beta) \in \mathcal{E}\left(J, \sqrt{\text{qchisq}(\gamma,\ell)}\right)\\ &\iff (\hat\beta - \beta)^T nJ(\hat\beta - \beta) \le \text{qchisq}(\gamma,\ell)\ .\end{aligned}$$

Wegen (17.4) trifft dies für jedes β mit Wahrscheinlichkeit γ zu. □

Das Volumen des Ellipsoids $\mathcal{E}(J,r)$ (und damit des Konfidenzellipsoids $B(\mathfrak{Y})$) hängt von der Determinante der Matrix J ab.

Satz: Das Volumen von $\mathcal{E}(J,r)$ ist

$$r^\ell \mu_\ell \frac{1}{\sqrt{\det J}} ,$$

wobei μ_ℓ das Volumen der ℓ-dimensionalen Einheitskugel ist.

Beweis: Sei

$$J = \sum_{j=1}^{\ell} \lambda_j^2 e_j e_j^T$$

wobei $0 < \lambda_1^2 \leq \ldots \leq \lambda_\ell^2$ die Eigenwerte und die $e_1, \ldots, e_\ell$ eine entsprechende ON-Basis von Eigenvektoren sind.

Es gilt dann:

$$\mathcal{E}(J,r) = \left\{ z \in \mathbb{R}^\ell \Big| \sum_{j=1}^{\ell} \left(\lambda_j z^T e_j\right)^2 \leq r^2 \right\} .$$

In den zu $e_1, \ldots, e_\ell$ gehörigen Koordinaten $(u_1, \ldots, u_\ell)$ ist $\mathcal{E}(J,r)$ also das Bild der Einheitskugel unter der Streckung $u_j \mapsto \frac{r}{\lambda_j} u_j$. Dieses hat Volumen

$$u_\ell r^\ell \frac{1}{\lambda_1, \ldots, \lambda_\ell} . \qquad \square$$

17.4 Lineare Regression mit Kontrollvariablen

Wir betrachten folgendes Modell:

$$Y_{\mathfrak{x}} = \mu(\mathfrak{x}) \qquad + \text{ Rauschen}$$

mit $\mu(\mathfrak{x}) = \sum_{j=1}^{\ell} \beta_j f_j(\mathfrak{x})$.

Die $f_1, \ldots, f_\ell$ sind dabei vorgegebene Funktionen von einem kompakten $K (\subseteq \mathbb{R}^d)$ nach $\mathbb{R}$; $\mathfrak{x} \in K$ bezeichnet man als *Kontrollvariable*, die Komponenten $x_1, \ldots, x_d$ von $\mathfrak{x}$ auch als *Faktoren*.

Für $\mathfrak{x}_1, \ldots, \mathfrak{x}_n \in K$ ist dann die Systemmatrix C unseres linearen Regressionsmodells gegeben durch

$$C = \begin{pmatrix} f^T(\mathfrak{x}_1) \\ \vdots \\ f^T(\mathfrak{x}_n) \end{pmatrix} ,$$

mit $f^T(\mathfrak{x}) := (f_1(\mathfrak{x}), \ldots, f_\ell(\mathfrak{x}))$.

Die Koordinaten μ_i der systematischen Komponente $\vec{\mu}$ sind dann von der Form

$$\mu_i = \sum_{j=1}^{\ell} \beta_j f_j(\mathfrak{x}_i)$$

wobei $\mathfrak{x}_i$ der Wert der Kontrollvariablen bei der i-ten Beobachtung ist.

Altbekannte Beispiele fügen sich in diesen Rahmen:

Beispiel 1. Gewöhnliche Varianzanalyse

$$K = \{1, \ldots, k\}$$
$$\mu_j = \beta_0 + \beta_j = \beta_0 + \sum_{p=1}^{k} \beta_p 1_{\{p\}}(j), \quad j \in K,$$
$$l = k + 1 \, .$$

Beispiel 2. Additives Modell mit k Behandlungen und m Blöcken:

$$K = \{1, \ldots, k\} \times \{1, \ldots, m\}$$
$$\mathfrak{x} = (u, v) \in K$$
$$\mu\left((u, v)\right) = \beta_0 + \beta_{1u} + \beta_{2v} = \beta_0 + \sum_{p=1}^{k} \beta_{1p} 1_{\{p\}}(u) + \sum_{q=1}^{m} \beta_{2q} 1_{\{q\}}(v)$$
$$l = k + m + 1 \, .$$

Beispiel 3. Zwei-Level-Faktorexperimente

$$K = \{-1, 1\}^d \quad , \quad \mathfrak{x} = (x_1, \ldots, x_d)$$
$$\mu(\mathfrak{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d + \beta_{12} x_1 x_2 + \ldots + \beta_{1\ldots d} x_1 x_2 \ldots x_d$$
$$l = 2^d \, .$$

Beispiel 4. Polynomiale Regression

$$K (\subseteq \mathbb{R}) \text{ ist ein Intervall.}$$
$$\mu(\mathfrak{x}) = \sum_{j=0}^{\ell-1} \beta_j \mathfrak{x}^j \, .$$

Kapitel 18

Versuchspläne

18.1 Ein Beispiel

Wir betrachten ein Wägeexperiment.

Die Aufgabe ist: Bestimme mit einer Federwaage die Gewichte von 3 Objekten mit 4 Messungen (wobei das Gewicht der Schale unbekannt ist).

Modell:

$$Y = \beta_{0+} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + W$$

wobei β_0 das Gewicht der Schale und β_r das Gewicht des r-ten Objektes ist..

$$\mathfrak{x} = (x_1, x_2, x_3) \in \{0,1\}^3 =: K$$

$x_r = 1$ bedeutet, dass sich das r-te Objekt auf der Waage befindet.

Ein naiver Versuchsplan ist

$$\begin{array}{l} \mathfrak{x}_1 = (1, \quad 0, \quad 0) \\ \mathfrak{x}_2 = (0, \quad 1, \quad 0) \\ \mathfrak{x}_3 = (0, \quad 0, \quad 1) \\ \mathfrak{x}_4 = (0, \quad 0, \quad 0) \end{array}$$

Die vierte Messung dient also zum Messen des Gewichtes der Waage. Dieses Design liefert (Übung)

$$\begin{aligned} \hat{\beta}_0 &= Y_4 \\ \hat{\beta}_r &= Y_r - Y_4 \quad , \quad r = 1,2,3\,. \end{aligned}$$

Ein zweiter, etwas subtilerer Versuchsplan ist

$$\begin{array}{l} \mathfrak{x}_1 = (1, \quad 0, \quad 0) \\ \mathfrak{x}_2 = (0, \quad 1, \quad 0) \\ \mathfrak{x}_3 = (0, \quad 0, \quad 1) \\ \mathfrak{x}_4 = (1, \quad 1, \quad 1) \end{array}$$

Dieser hat den Vorteil, dass auch bei der 4. Wägung Information über die Gewichte der Objekte gewonnen wird.

Der zweite Plan liefert (Übung)

$$\hat{\beta}_0 = \frac{1}{2}(Y_1 + Y_2 + Y_3 - Y_4)$$
$$\hat{\beta}_1 = \frac{1}{2}(Y_1 - Y_2 - Y_3 + Y_4)$$
$$\hat{\beta}_2 = \frac{1}{2}(-Y_1 + Y_2 - Y_3 + Y_4)$$
$$\hat{\beta}_3 = \frac{1}{2}(-Y_1 - Y_2 + Y_3 + Y_4)$$

Damit ergeben sich folgende Eintragungen für $\mathbf{Var}\ \hat{\beta}$:

	$\operatorname{Var}\hat{\beta}_0$	$\operatorname{Var}\hat{\beta}_1$	$\operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1)$	$\operatorname{Cov}(\hat{\beta}_1, \hat{\beta}_2)$
1. Design	σ^2	$2\sigma^2$	$-\sigma^2$	σ^2
2. Design	σ^2	σ^2	$-\sigma^2/2$	0

18.2 Zwei Schätzprobleme

Wir bleiben beim Modell des Abschnitts 17.4. Die Funktionen $f_1, \ldots, f_\ell : K \to \mathbb{R}$ seien vorgegeben,

$$\mu(\mathfrak{x}) := \sum_{j=1}^{\ell} \beta_j f_j(\mathfrak{x}) = f^T(\mathfrak{x})\beta \quad , \quad \mathfrak{x} \in K\ .$$

Für $\mathfrak{x}_1, \ldots, \mathfrak{x}_n \in K$ sei

$$Y_i := \mu(\mathfrak{x}_i) + W_i\ ,\ \text{mit } (W_1, \ldots, W_n) \text{ weißes Rauschen mit Varianz } \sigma^2.$$

Frage 1. Wie gut läßt sich $\beta = (\beta_1, \ldots, \beta_\ell)$ aus den Beobachtungen $Y_1, \ldots, Y_n$ schätzen?

Zur Erinnerung: Die Systemmatrix unseres Modells ist

$$C = \begin{pmatrix} f^T(\mathfrak{x}_1) \\ \vdots \\ f^T(\mathfrak{x}_n) \end{pmatrix}\ .$$

Damit ist $C^T C = \sum_{i=1}^n f(\mathfrak{x}_i) f^T(\mathfrak{x}_i)$, und die Informationsmatrix ist

$$J = \frac{1}{\sigma^2}\frac{1}{n}\sum_{i=1}^{n} f(\mathfrak{x}_i) f^T(\mathfrak{x}_i)\ .$$

Die Kovarianzmatrix des Kleinst-Quadrate-Schätzers $\hat{\beta}$ ist

$$\mathbf{Var}\hat{\beta} = \frac{1}{n} J^{-1}\ .$$

Ein Maß für die Größe von J ist seine Determinante $\det J$. Für normales weißes Rauschen ist das Volumen des Konfidenzellipsoids für β invers proportional zu $\sqrt{\det J}$ (vgl. Abschnitt 17.3). Ein "kleines" J ergibt sich, wenn die Spaltenvektoren $c_1, \ldots, c_\ell$ von C allesamt klein sind, oder wenn sie "fast linear abhängig" sind.

Frage 2. Wie gut lässt sich $\mu(\mathfrak{x})$ aus den Beobachtungen $Y_1, \ldots, Y_n$ schätzen? (Dabei fassen wir ein bestimmtes $\mathfrak{x} \in K$ ins Auge).

$\mu(\mathfrak{x}) = \sum_{j=1}^{\ell} f_j(\mathfrak{x})\beta_j$ ist ein lineares Funktional von β; sein Kleinst-Quadrate-Schätzer ist

$$\hat{\mu}(\mathfrak{x}) := \sum_{j=1}^{\ell} f_j(\mathfrak{x})\hat{\beta}_j = f^T(\mathfrak{x})\hat{\beta} \; .$$

Dessen Varianz ist

$$\operatorname{Var} \hat{\mu}(\mathfrak{x}) = \frac{1}{n} f^T(\mathfrak{x}) J^{-1} f(\mathfrak{x}) \; .$$

18.3 Optimale Versuchspläne

Wir betrachten in dem im Abschnitt 17.4 definierten Modell die folgende Situation:

Wir haben n Messungen frei. An welchen Stellen $\mathfrak{x}_1, \ldots, \mathfrak{x}_n$ sollten wir messen, damit wir

(i) den Parametervektor β

(ii) die systematische Komponente $\mu(\mathfrak{x})$ gleichmäßig über alle $\mathfrak{x} \in K$

möglichst gut schätzen können?

Im Sinn des im vorigen Abschnitt Gesagten läuft dies auf zwei Optimierungsprobleme hinaus:

(D) finde $\mathfrak{x}_1, \ldots, \mathfrak{x}_n$ so, dass $\sqrt{\det J}$ maximal wird,

(G) finde $\mathfrak{x}_1, \ldots, \mathfrak{x}_n$ so, dass $\sup_{\mathfrak{x} \in K} f^T(\mathfrak{x}) J^{-1} f^T(\mathfrak{x})$ minimal wird.

Weil J offensichtlich nicht von der Reihenfolge der $\mathfrak{x}_i$ abhängt, ist dies eine Frage an die empirische Verteilung ξ der $\mathfrak{x}_1, \ldots, \mathfrak{x}_n$.

Im Allgemeinen haben (D) und (G) nicht diesselbe Lösung.

Beispiel: (vgl. J. Kiefer, Optimum experimental designs, J. Royal Stat. Soc. Ser. B, 21 (1959), 272-304), abgedruckt und kommentiert in: Breakthroughs in Statistics, Bd. 1 (S. Kotz, N.L. Johnson (eds), Springer 1992).

Quadratische Regression auf $[-1, 1], n := 4$:

D-optimal ist $\left\{\pm 1, \pm\frac{1}{\sqrt{2}}\right\}$

G-optimal ist $\left\{\pm 1, \pm\frac{1}{\sqrt{3}}\right\}$.

Wie wir im nächsten Abschnitt zeigen werden, führen die beiden Probleme (D) und (G) in der Asymptotik $n \to \infty$ auf dieselben Lösungen.

18.4 Zur Äquivalenz von D- und G-Optimalität

Wieder sind wir im Szenario der vorigen Abschnitte. Jedes W-Maß ξ auf K fassen wir als einen *asymptotischen Versuchsplan (Design)* auf.

Die Matrix

$$J(\xi) := \frac{1}{\sigma^2} \int_K f(\mathfrak{x}) f^T(\mathfrak{x}) \xi(d\mathfrak{x})$$

heißt Informationsmatrix zum Design ξ.

Wir betrachten zwei Optimierungsprobleme:

(D) Finde ξ so, dass $\det J(\xi)$ maximal wird.

(G) Finde ξ so, dass $\sup_{\mathfrak{x} \in K} f(\mathfrak{x})^T J(\xi) f(\mathfrak{x})$ minimal wird.

Satz (Äquivalenztheorem von Kiefer)[1]

Die Funktionen $f_1, \ldots, f_\ell$ seien linear unabhängig. Dann gilt: Ein Design ξ ist D-optimal genau dann, wenn es G-optimal ist.

Der Rest des Abschnittes ist dem Beweis des Äquivalenztheorems gewidmet.

Wir setzen

$$\begin{aligned}
\mathfrak{M} &:= \text{ Menge der } \ell \times \ell\text{-Matritzen} \\
\mathfrak{M}_+ &:= \left\{ M \in \mathfrak{M} \middle| M \text{ regulär} \right\} \\
\mathcal{S} &:= \left\{ M \in \mathfrak{M} \middle| M \text{ symmetrisch} \right\} \\
\mathcal{S}_+ &:= \left\{ M \in \mathcal{S} \middle| M \text{ positiv definiert} \right\} \\
\mathfrak{J} &:= \left\{ J(\xi) \middle| \xi\, W - \text{Maß auf } K \right\} \\
\mathfrak{J}_+ &:= \mathfrak{J} \cap \mathfrak{M}_+ \, .
\end{aligned}$$

Wir betrachten zwei Abbildungen auf M:

$$\begin{aligned}
\Psi(M) &:= -\log\det M \\
\Phi(M) &:= \begin{cases} \sup_{x \in K} f(\mathfrak{x})^T M^{-1} f(\mathfrak{x}) & \text{falls } M \text{ regulär} \\ \infty & \text{falls } M \text{ singulär} \end{cases}
\end{aligned}$$

Unsere beiden Optimierungsprobleme lauten

[1] formuliert in: J. Kiefer (1959), loco citato, Beweis erschienen in J. Kiefer, J. Wolfowitz, The equivalence of two extremum problems, Canad. J. Math 12 (1980), 363-366

(D) maximiere Ψ auf $\mathfrak{J}$

(G) minimiere Φ auf $\mathfrak{J}$

Die Menge $\Pi(K)$ der W-Maße auf K, ausgestattet mit der von den Abbildungen $\pi \mapsto \int g(\mathfrak{x})\pi(dx), g : K \to \mathbb{R}$ stetig, erzeugten Topologie, ist kompakt (siehe etwa Billingsley, Convergence of probability measures).

Die Abbildung $\pi \to J(\pi)$ ist stetig von $\Pi(K)$ nach $\mathfrak{J}$. Außerdem sind Ψ und Φ stetig von $\mathfrak{J}$ nach $\bar{\mathbb{R}}$. Damit ist die Existenz von Lösungen für (D) und (G) schon mal gesichert.

Wir analysieren jetzt die Abbildung

$$\begin{aligned} \Psi : \mathfrak{M}_+ &\to \mathbb{R} \\ M &\mapsto -\log \left| \det M \right| . \end{aligned}$$

Lemma 1: Für alle $M \in \mathfrak{M}$ und alle $V(\neq 0) \in \mathfrak{M}$ gilt für die Richtungsableitung $\frac{\partial \Psi}{\partial V}(M)$

a) $\frac{\partial \Psi}{\partial V}(M) = \left\langle -M^{-1}, V \right\rangle$

b) $\frac{\partial^2 \Psi}{\partial V^2}(M) = \left\langle M^{-1} V M^{-1}, V \right\rangle$

Dabei setzen wir für $A, B \in \mathfrak{M}$:

$$\langle A, B \rangle := \sum_{i,j=1}^{\ell} a_{ij} b_{ij} = \text{tr } AB^T$$

wobei

$$\text{tr } H := \sum_{i=1}^{\ell} h_{ii}$$

die sogenannte Spur (trace) der Matrix H bezeichnet.

Beweis von Lemma 1:

a) Sei $M = (m_{ij})_{1 \le i,j \le \ell}$. Wir zeigen

$$\text{grad } \Psi(M) = \left(\frac{\partial}{\partial m_{ij}} \Psi(M) \right)_{1 \le i,j \le \ell} = -M^{-1} .$$

In der Tat: Die Determinantenentwicklungsformel ergibt mit Entwicklung nach der j-ten Spalte

$$\det M = \sum_{i=1}^{\ell} (-1)^{i+j} m_{ij} \det M_{ij} .$$

(Dabei entsteht die Matrix M_{ij} aus M durch Streichen der i-ten Zeile und der j-ten Spalte.)

Es folgt

$$\frac{\partial}{\partial m_{ij}} \det M = (-1)^{i+j} \det M_{ij} =: \tilde{m}_{ij} \; .$$

Aus der Linearen Algebra kennt man folgende Darstellung von M^{-1} :

$$M^{-1} = \frac{1}{\det M} (\tilde{m}_{ij}) \; .$$

Die Kombination der beiden letzten Formeln ergibt

$$(M^{-1})_{ij} = \frac{1}{\det M} \frac{\partial}{\partial m_{ij}} \det M = \frac{\partial}{\partial m_{ij}} \log \left| \det M \right| \; .$$

b) Es reicht zu zeigen:

$$\lim_{\varepsilon \to \infty} \frac{1}{\varepsilon} \left((M + \varepsilon V)^{-1} - M^{-1} \right) = -M^{-1} V M^{-1} \; .$$

Dies folgt wegen:

$$\begin{aligned} (M + \varepsilon V)^{-1} - M^{-1} &= (M + \varepsilon M M^{-1} V)^{-1} - M^{-1} \\ = \left(M \left(I + \varepsilon M^{-1} V \right) \right)^{-1} - M^{-1} &= \left(I + \varepsilon M^{-1} V \right)^{-1} M^{-1} - M^{-1} \\ = \left(I - \varepsilon M^{-1} V \right) M^{-1} - M^{-1} + o(\varepsilon) &= -\varepsilon M^{-1} V M^{-1} + o(\varepsilon) \qquad \square \end{aligned}$$

Lemma 2. Für $A, B \in \mathfrak{M}$ gilt:

$$\operatorname{tr}(AB) = \operatorname{tr}(BA) \, .$$

Beweis: $\operatorname{tr}(AB) = \sum_i \sum_j a_{ij} b_{ij} = \sum_j \sum_i b_{ij} a_{ij} = \operatorname{tr}(BA)$. $\qquad \square$

Lemma 3. Ist $M \in \mathfrak{M}$ symmetrisch und positiv definit, und $V (\neq 0) \in \mathfrak{M}$, so ist

$$\frac{\partial^2}{\partial V^2} \Psi(M) > 0 \; .$$

Beweis: Es sei $M^{-1} = AA$, mit A symmetrisch und positiv definit. Wegen Lemma 1b) und Lemma 2 gilt:

$$\begin{aligned} \frac{\partial^2}{\partial V^2} \Psi(M) &= \operatorname{tr} \, \left(M^{-1} V M^{-1} V^T \right) \\ &= \operatorname{tr} \, \left(A V A A^T V^T A^T \right) > 0 \; . \qquad \square \end{aligned}$$

Als nächstes analysieren wir die Einschränkung von Ψ auf $\mathfrak{J}_+$.

Es gilt: $\mathfrak{J} \subseteq \mathcal{S}$ und $\mathcal{S} \cap \mathfrak{M}_+ = \mathcal{S}_+$.

Also ist

$$\mathfrak{J}_+ = \mathfrak{J} \cap \mathcal{S}_+ .$$

Weil sowohl $\mathfrak{J}$ als auch $\mathcal{S}_+$ konvexe Teilmengen von $\mathfrak{M}$ sind, ist auch $\mathfrak{J}_+$ konvex.

Aus Lemma 3 folgt nun:

Lemma 4. Ψ ist strikt konvex auf $\mathfrak{J}_+$ (und konvex auf $\mathfrak{J}$).

Als nächstes charakterisieren wir das globale Minimum von Ψ auf $\mathfrak{J}$ als diejenige Informationsmatrix, von der aus es überallhin nach $\mathfrak{J}$ aufwärts geht:

Lemma 5. Folgende Aussagen sind äquivalent:

a) J ist globale Minimalstelle von $\Psi\big|\mathfrak{J}$;

b) $\langle \text{grad}\Psi(J), H - J\rangle \geq 0 \quad \forall H \in \mathfrak{J}$;

c) $\langle \text{grad}\Psi(J), \frac{1}{\sigma^2} f(\mathfrak{x}) f(\mathfrak{x})^T - J\rangle \geq 0 \quad \forall \mathfrak{x} \in K$.

Beweis:

a) $\Rightarrow$ **b)** : $\quad \langle \text{ grad } \Psi(J), H - J\rangle$

$= \lim_{\varepsilon \to 0} \varepsilon^{-1} \Big(\underbrace{\Psi(J + \varepsilon(H - (J))}_{\in J, \text{ weil } J \text{ konvex}} - \Psi(J) \Big)$

$\geq 0 \quad .$

b) $\Rightarrow$ **a)** Für jedes $H \in \mathfrak{J}$ ist

$g : \alpha\,(\in [0,1]) \longrightarrow \Psi\,(J + \alpha(H - J))$

konvex und differenzierbar in $\alpha = 0$.

Also:

$$\begin{aligned} \Psi(H) - \Psi(J) = g(1) - g(0) &\geq g'(0) \\ &= \langle \text{ grad } \Psi(J), H - J\rangle \ . \end{aligned}$$

b) $\Rightarrow$ **c)** : klar

c) $\Rightarrow$ **b)** : Für alle $H \in \mathfrak{J}$ der Form

$$H = \sum_{i=1}^{n} \gamma_i \frac{1}{\sigma^2} f(\mathfrak{x}_i) f^T(\mathfrak{x}_i) \ , \ n \in \mathbb{N} \ , \ \gamma_i \leq 0 \ , \ \sum \gamma_i = 1 \tag{18.1}$$

gilt

$$\langle \text{ grad } \Psi(J), H - J\rangle \geq 0$$

wegen c) und aus Linearitätsgründen. Weil die Menge aller H der Form (18.1) dicht in $\mathfrak{J}$ ist, überträgt sich die Ungleichung auf alle $H \in \mathfrak{J}$. $\square$

Bemerkung: Für $J \in \mathfrak{J}_+$ und $\mathfrak{x} \in K$ ist

$$\left\langle \text{ grad } \Psi(J) \ , \frac{1}{\sigma^2} f(\mathfrak{x}) f(\mathfrak{x})^T - J \right\rangle$$

$$\begin{aligned} &= \left\langle -J^{-1}, \frac{1}{\sigma^2} f(\mathfrak{x}) f(\mathfrak{x})^T \right\rangle + \left\langle J^{-1}, J \right\rangle \\ &= -\frac{1}{\sigma^2} \text{ tr } \left(J^{-1} f(\mathfrak{x}) f(\mathfrak{x})^T \right) + \text{ tr } \left(J^{-1} J \right) \\ &= -\frac{1}{\sigma^2} \text{ tr } \Big(\underbrace{f^T(\mathfrak{x}) J^{-1} f(\mathfrak{x})}_{\in \mathbb{R}} \Big) + \text{ tr } (I) \\ &= -\frac{1}{\sigma^2} f^T(\mathfrak{x}) J^{-1} f(\mathfrak{x}) \quad + \ell \ . \end{aligned}$$

Korollar aus der Bemerkung und Lemma 5:

Sei $J \in \mathfrak{J}_+$. Dann gilt:

J ist globale Minimalstelle von $\Psi\big|\mathfrak{J} \iff$

$$\iff \frac{1}{\sigma^2} f^T(\mathfrak{x}) J^{-1} f(\mathfrak{x}) \leq \ell \quad \forall \mathfrak{x} \in K$$
$$\iff \Phi(J) \leq \sigma^2 \ell \; .$$

Finale des Beweises des Äquivalenztheorems:

Sei ξ D-optimal und η G-optimal. Dann folgt aus dem vorigen Korollar:

$$\Phi\left(J(\eta)\right) \leq \Phi\left(J(\xi)\right) \leq \sigma^2 \ell \; .$$

Eine nochmalige Anwendung des Korollars impliziert, dass η auch D-optimal ist. Also sind sowohl $J(\xi)$ als auch $J(\eta)$ Minimalstellen von Ψ auf $\mathfrak{J}_+$. Aus der strikten Konvexität von Ψ auf $\mathfrak{J}_+$ folgt:

$$J(\eta) = J(\xi) \; .$$

Damit ist ξ auch G-optimal. $\square$

18.5 Polynomiale Regression auf $[-1, 1]$

$$K = [-1, 1] \; , \; f_0(x) \equiv 1 \; , f_1(x) \equiv x \; , \ldots f_{\ell-1}(x) \equiv x^{\ell-1} \; , \; \xi := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

$$C\left(x_1, \ldots, x_n\right) := \begin{pmatrix} 1 & x_1 & \ldots & x_1^{\ell-1} \\ \vdots & & & \\ 1 & x_n & \ldots & x_n^{\ell-1} \end{pmatrix} \quad , \quad J(\xi) = \frac{1}{\sigma^2 n} C^T C \; .$$

Beispiel: $\ell = 3$

Behauptung: $\xi = \frac{1}{3}\left(\delta_{-1} + \delta_0 + \delta_1\right)$ ist D-optimal

Begründung: Nach dem Korollar aus dem vorigen Abschnitt reicht es zu zeigen:

$$f(x)^T J(\xi)^{-1} f(x) \leq \sigma^2 \cdot \ell \tag{18.2}$$

Dabei ist:

$$J(\xi) = \frac{1}{\sigma^2} \frac{1}{3} \left[\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} (1, -1, 1) + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} (1, 0, 0) + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} (1, 1, 1) \right]$$
$$= \frac{1}{\sigma^2} \frac{1}{3} \begin{pmatrix} 3 & 0 & 2 \\ 0 & 2 & 0 \\ 2 & 0 & 2 \end{pmatrix} \; .$$

In der Tat gilt:

$$\frac{1}{\sigma^2}\left(1,x,x^2\right)3\sigma^2\begin{pmatrix}3&0&2\\0&2&0\\2&0&2\end{pmatrix}^{-1}\begin{pmatrix}1\\x\\x^2\end{pmatrix}=\frac{3}{4}\left(4-6x^2\left(1-x^2\right)\right)\leq 3\quad\text{für } x\in[-1,1]\,.$$

Nach dem Äquivalenztheorem ist ξ auch G-optimal. Es ist also vorteilhaft, jeweils ein Drittel der Messungen bei $-1, 0$ und 1 durchzuführen. Will man partout die Messungen gleichmäßig auf **vier** Stellen $x_1,\ldots,x_4$ aufteilen, dann klafft, wie wir im vorletzten Abschnitt festgestellt haben, eine Lücke zwischen D- und G-Optimalität.

Wir betrachten nun den Fall eines allgemeinen ℓ:

Satz Das D-optimale Design ist gegeben durch

$$\xi=\frac{1}{\ell}\sum_{i=1}^{\ell}\delta_{x_i^\star}\,,$$

wobei $x_1^\star,\ldots,x_\ell^\star$ die Nullstellen des Polynoms

$$\left(1-x^2\right)\frac{d^\ell}{dx^\ell}\left(x^2-1\right)^{\ell-1}=\left(1-x^2\right)\frac{d}{dx}\tilde{P}_{\ell-1}(x)$$

sind, mit

$$\tilde{P}_k(x):=\frac{d^k}{dx^k}\left(x^2-1\right)^k\,.$$

$\tilde{P}_0,\tilde{P}_1,\ldots$ sind (bis auf Normierungskonstante) gleich den sogenannten Legendre-Polynomen P_n, die ein vollständiges Orthonormalsystem in $L^2\left([-1,1]\right)$ bilden.

Beweis: siehe S. Karlin, W.J. Studden, Optimum Experimental Designs, Ann. Math. Statist. 37 (1966), 783-815.

Bemerkung: Die $x_i^\star$ sind dichter in der Nähe von ±1 als in der Nähe von 0.

Skizze des Beweises: Man zeigt zuerst, dass man sich auf Designs ξ der Form $\xi=\frac{1}{\ell}\sum_{i=1}^{\ell}\delta_{x_j}$ zurückziehen kann.

Dann bleibt das Problem der Bestimmung der Maximalstelle von

$$\prod_{i<j}\left|x_i-x_j\right|\quad,\quad -1\leq x_1,\ldots,x_\ell\leq 1\,.$$

(Beachte: $\prod_{i<j}\left|x_i-x_j\right|$ ist die Determinante der Van der Monde Matrix $C(x_1,\ldots,x_\ell)$.)

Unter den x_j sind in jedem Fall die Punkte -1 und +1 (sonst wird was verschenkt!).

Also geht es um die Maximierung von

$$g(x_1,\ldots,x_{\ell-2}):=\prod_{j=1}^{\ell-2}\left[(x_j-1)(x_j+1)\right]\prod_{1\leq j<k\leq\ell-2}\left|x_j-x_k\right|\,.$$

Man prüft leicht nach:

grad $g(x_1, \ldots, x_{\ell-2}) = 0$

$$\iff \forall j : \sum_{k \neq j} \frac{1}{x_j - x_k} + \frac{1}{x_j - 1} + \frac{1}{x_j + 1} = 0 \; . \tag{18.3}$$

Definieren wir das Polynom

$$\varphi(x) = \prod_{j=1}^{\ell-2} (x - x_j),$$

so wird die rechte Seite von (18.3) zu

$$\forall j : \frac{1}{2} \frac{\varphi''(x_j)}{\varphi'(x_j)} + \frac{1}{x_j - 1} + \frac{1}{x_j + 1} = 0 \; .$$

Dies ist gleichbedeutend mit

$$\forall j : \left(1 - x_j^2\right) \varphi''(x_j) - 4 x_j \varphi'(x_j) = 0 \; .$$

Also ist $x \mapsto (1 - x^2)\varphi''(x) - 4x\varphi'(x)$ ein Polynom vom Grad $\ell - 2$, welches dieselben Nullstellen wie φ hat. Somit gilt:

$$(1 - x^2)\varphi''(x) - 4x\varphi'(x) + \gamma\varphi(x) = 0 \quad \text{für ein } \gamma \in \mathbb{R} \; .$$

Nach G. Szegö, Orthogonal Polynomials, AMS Coll. Publ. 1939, Thm 4.2.2 folgt:

φ ist bis auf eine Konstante das Jacobi-Polynom $P_{\ell-2}^{(1,1)}$

(Die Jacobi-Polynome $P_n^{(\alpha,\beta)}(x), x \in [-1, 1]$, sind orthogonal bzgl. der Gewichtsfunktion $w(x) := (1 - x)^\alpha (1 + x)^\beta$, d.h. $\int_{-1}^{1} P_m^{(\alpha,\beta)}(x) P_n^{(\alpha,\beta)}(x) w(x) dx = \delta_{mn}$).

Nach Szegö, (4.7.27) zusammen mit (4.7.1) und (4.7.2) folgt

$P_{\ell-2}^{(1,1)}$ ist bis auf eine Konstante gleich der Ableitung des Legendre Polynoms $P_{\ell-1}$. (Die Legendre-Polynome $P_n, x \in [-1, 1]$, sind orthogonal bzgl. der Gewichtsfunktion 1).

Die Behauptung ergibt sich nun mit der Formel von Rodrigues (Abramowitz-Stegun, Handbook of Mathematical Functions, p.334)

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n (x^2 - 1)^n}{dx^n} \; .$$