STA 120
Einführung in die Statistik

Script

Reinhard Furrer
and the Applied Statistics Group

Version 12. Februar 2017, 15a17b3.

Inhaltsverzeichnis

Vorwort

Dieses Dokument wurde für das Modul "STA120 Einführung in die Statistik" geschrieben. Das Modul beinhaltet 14 Lektionen und in den Jahren 2013–2016 wurden dementsprechend 13 Kapitel, wie hier vorhanden, behandelt (plus eine Repetitionsstunde am Ende des Semesters).

Ein grosses Danke an all die externen "Beiträge", insbesondere von (alphabetisch) Julia Braun, Eva Furrer, Florian Gerber, Lisa Hofer, Mattia Molinaro, Franziska Robmann, u.v.m.

Das Skript enthält immer noch viele Tippfehler und einige Unklarheiten. Für Verbesserungsvorschläge und Beiträge bin ich dankbar. Wir investieren zur Zeit mehr Energie in die englische Version diese Dokuments, das wir auf das Frühjahrsemester 2017 erstellen.

Reinhard Furrer
Februar 2017

Kapitel 1

Daten graphisch darstellen

Die graphische Darstellung von Daten is ein zentraler Punkt einer statistischen Analyse. Graphische Darstellungen spannen oft eine Analyse: zum Begin werden die Beobachtungen oder Messwert dargestellt, man spricht von einer explorativen Analyse (*EDA*, *exploratory data analysis*). Am Ende werden die Resultate auch wieder oft graphisch zusammen gefasst; meistens ist es einfacher eine Grafik zu interpretieren oder verstehen als Werte einer Tabelle.

In diesem Kapitel betrachten wir wie verschiedene Arten von Daten dargestellt werden können.

1.1 Typen von Daten

Es gibt verschiedene Arten von Daten. Eine wichtiges Merkmal ist, ob Daten qualitativ oder quantitativ sind. Qualitative Daten sind nicht-numerische Daten, die oft in verschriftlichter oder in audiovisueller Form vorliegenden. Daher ergibt sich für qualitative Daten als einziges Skalenniveau die Nominalskalierung. Beispiele von quantitativ Daten sind kategorische oder metrische Daten, mit Ordinalskala, Intervallskala oder Verhältnisskala als möglichem Skalenniveau. Merkmalsausprägungen, die intervallskaliert sind, einen natürlicher Nullpunkt und eine willkürliche Masseinheit besitzen, sind verhältnisskaliert. Die Skalenniveaus sind nach Stevens (1946) klassifiziert und in Abbildung 1.1 zusammengefasst.

Bei diskreten Werten ist der Modus der “häufigste” Wert und zu dessen Berechnung benötigt man lediglich die Operatoren $\{=, \neq\}$. Der Median (eigentlich empirischer Median) ist derjenige Wert in der Mitte, dazu müssen die Daten sortiert werden (mit den Operatoren $\{<, >\}$.

Beispiel 1.1. Wenn für Temperaturen Kelvin verwendet wird (absoluten Nullpunkt bei $-273.15°C$) macht eine Aussage: ”Die Temperatur ist um 20% gestiegen” Sinn.

	Skalen			
	Nominal	Ordinal	Intervall	Verhältnis
Beispiele	Attribut: männlich/weiblich	Attribut: genügend/gut/ausgezeichnet	Temperatur gemessen in Celsius	Temperatur gemessen in Kelvin
Mathematische Operationen	= ≠	= ≠ < >	= ≠ < > + -	= ≠ < > + - * /
Statistische Kennzahlen	Modus	Modus Median	Modus Median Mittelwert	Modus Median Mittelwert Geometrisches Mittel
			Standard-abweichung Spannweite	Standard-abweichung Variations-koeffizient Studentisierte Spannweite

Abbildung 1.1: Skalenniveaus nach Stevens (1946).

1.2 Statistische Kennzahlen

Eine Kennzahl ist eine sehr reduzierte "Darstellung" der Daten, gibt aber dennoch einen ersten Eindruck über die Verteilung der Daten. Typische Kennzahlen für Lageparameter sind das arithmetische Mittel, gestutzte Mittel (*truncated mean*), Median, ..., für Skalenparameter sind es Varianz, Standardabweichung, Interquartilsabstand (*interquartile range*).

Der Modus (*mode*) ist der häufigste Wert bei einer empirischen Häufigkeitsverteilung. Bei intervallskalierten Daten werden diese (zuerst) in Klassen eingeteilt.

Beispiel 1.2. In R-Code 1.1 werden verschiedene Kennzahlen von 293 Proben berechnet. (Units?)

1.3 Klassische Diagrammarten für univariate Daten

Graphische Darstellung von univariaten Daten, wie zum Beispiel ein Wert pro Messung/Objekt oder eine Gruppe von Werten, erfolgen oft mit einem Histogramm, Boxplot, Q-Q-Plot oder Barplot.

Histogramme (*histograms*) stellen die Häufigkeitsverteilung von Beobachtungen graphisch dar. Histogramme sind einfach zu erstellen und interpretieren. Die Schwierigkeit ist die Wahl der Klassenanzahl.

Ein Stamm-Blatt-Diagramm (*stem-and-leaf*) ist einem Histogramm ähnlich, wird aber heutzutage quasi nicht mehr benutzt (Abbildung 1.3).

R-Code 1.1 Quecksilberdaten.

```
Hg <- c( read.csv('data/lemanHg.csv'))$Hg
str( Hg)
##  num [1:293] 0.17 0.21 0.06 0.24 0.35 0.14 0.08 0.26 0.23 0.18 ...
c( mean=mean(Hg), tr.mean=mean(Hg, trim=.1), median=median(Hg))
##      mean   tr.mean    median
## 0.4617747 0.4323830 0.4000000
c( var=var(Hg), sd=sd(Hg), iqr=IQR(Hg))
##        var         sd        iqr
## 0.09014615 0.30024349 0.38000000
```

R-Code 1.2 Histogramme. (Siehe Abbildung 1.2.)

```
hist( Hg)
hist( Hg, col=7, probability=TRUE, main='')
hist( Hg, col=7, breaks=90, main='')
hist( Hg, col=7, breaks=2, main='')
```

Ein Boxplot (*boxplot*) ist eine graphische Darstellung von fünf Kennzahlen der Häufigkeitsverteilung von Beobachtungen: kleinster/grösster Wert, unteres und obers Quartil und der Median. Vorteile von einem Boxplot sind:

- Quantitativ Darstellung von wichtigen Kennzahlen;
- Symmetrie ist visuell schnell erfasst;
- Ausreisser sind markiert.

R-Code 1.3 Boxplot. (Siehe Abbildung 1.4.)

```
boxplot( Hg)
boxplot( Hg, col="LightBlue", notch=TRUE, ylab="Hg",
        outlty=1, outpch='', outcol=2)
summary( Hg)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.2500  0.4000  0.4618  0.6300  1.7700
```

Quantile-Quantile-Diagramm (Q-Q-Plot, *Q-Q-plot*) ist eine graphische Darstellung, um empirische Datenquantile mit theoretischen Verteilungsquantilen zu vergleichen. Die

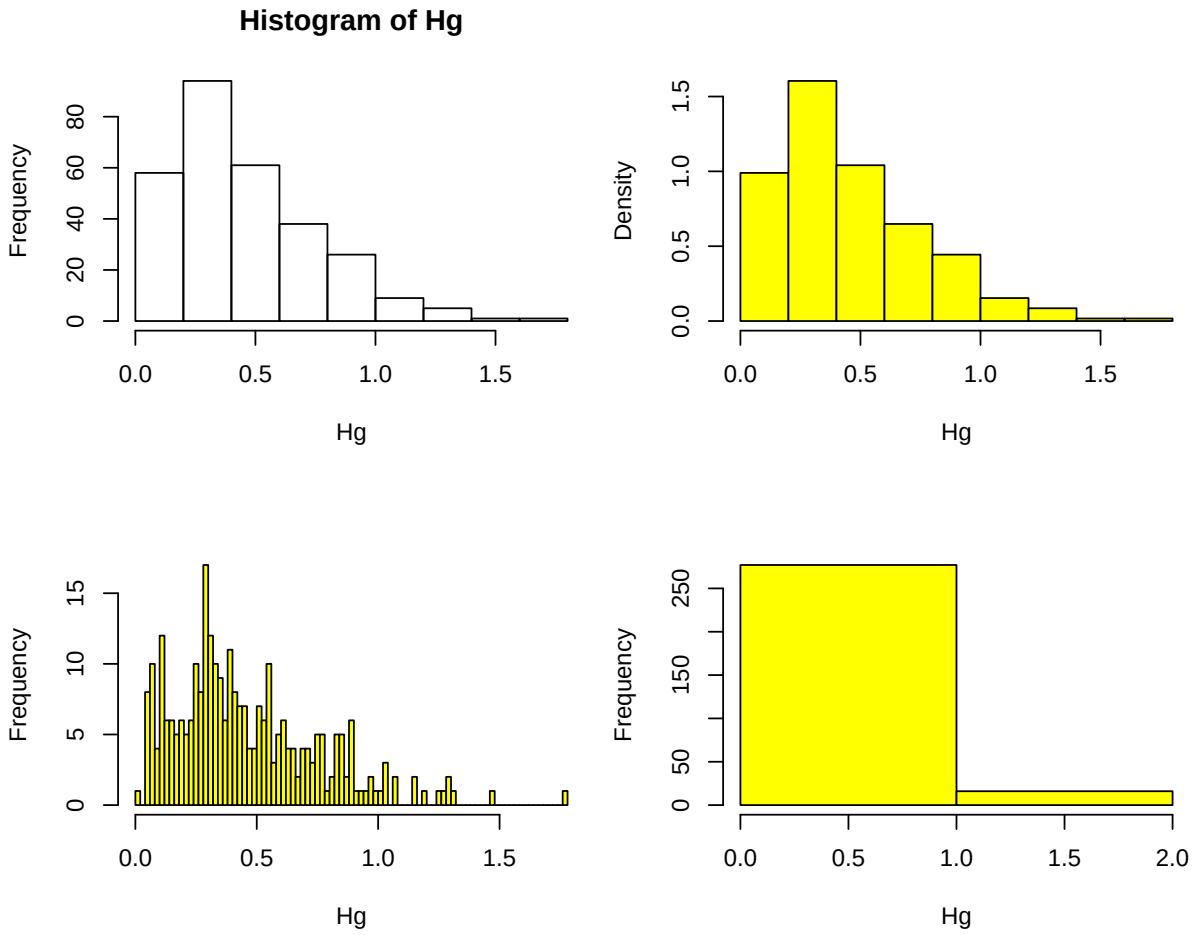


Abbildung 1.2: Histogramme. (Siehe R-Code 1.2.)

R-Code 1.4 QQplot. (Siehe Abbildung 1.5.)

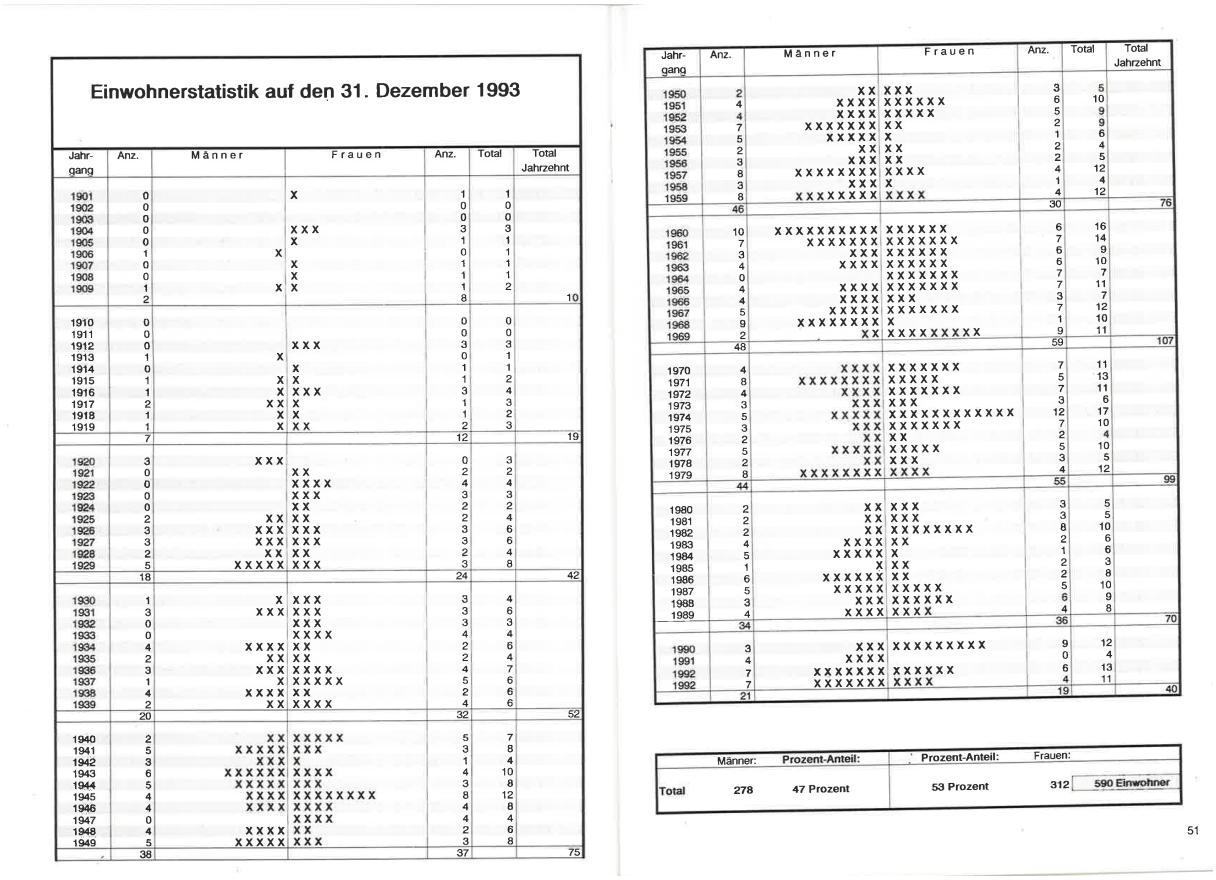
```
qqnorm( Hg)
qqline( Hg, col=2)

qqplot( qexp( ppoints( 293), rate=2.1), Hg)
qqline( Hg, distribution=function(p) qexp( p, rate=2.1),
        prob=c(0.1, 0.6), col=2)
```

geordneten Werte werden mit den $i/(n+1)$-Quantilen verglichen. In der Praxis wird, je nach Software, $(i-a)/(n+1-2a)$, $a \in [0,1]$, verwendet.

Die Häufigkeitsverteilung von diskreten Daten wird oft mit sogenannten Barplots dargestellt. Die Höhe eines "Bars" ist proportional zur Häufigkeit des entsprechenden Wertes/Variable. Die deutsche Sprache unterscheidet zwischen Balken- und Säulendiagramme (vertikal/horizontale Ausrichtung).

Keine Kuchendiagramme (Kreisdiagramm *pie chart*).



Einwohnerstatistik auf den 31. Dezember 1993

Jahrgang	Anz.	Männer	Frauen	Anz.	Total	Total Jahrzehnt
1901	0		x	1	1	
1902	0			0	0	
1903	0			0	0	
1904	0		xxx	3	3	
1905	0		x	1	1	
1906	1	x		0	1	
1907	0		x	1	1	
1908	0		x	1	1	
1909	1	x	x	1	2	
	2			8		10
1910	0			0	0	
1911	0			0	0	
1912	0		xxx	3	3	
1913	1	x		0	1	
1914	0		x	1	1	
1915	1	x	x	1	2	
1916	1	x	xxx	3	4	
1917	2	xx	x	1	3	
1918	1	x	x	1	2	
1919	1	x	xx	2	3	
	7			12		19
1920	3	xxx		0	3	
1921	0		xx	2	2	
1922	0		xxxx	4	4	
1923	0		xxx	3	3	
1924	0		xx	2	2	
1925	2	xx	xx	2	4	
1926	3	xxx	xxx	3	6	
1927	3	xxx	xxx	3	6	
1928	2	xx	xx	2	4	
1929	5	xxxxx	xxx	3	8	
	18			24		42
1930	1	x	xxx	3	4	
1931	3	xxx	xxx	3	6	
1932	0		xxx	3	3	
1933	0		xxxx	4	4	
1934	4	xxxx	xx	2	6	
1935	2	xx	xx	2	4	
1936	3	xxx	xxxx	4	7	
1937	1	x	xxxxx	5	6	
1938	4	xxxx	xx	2	6	
1939	2	xx	xxxx	4	6	
	20			32		52
1940	2	xx	xxxxx	5	7	
1941	5	xxxxx	xxx	3	8	
1942	3	xxx	x	1	4	
1943	6	xxxxxx	xxxx	4	10	
1944	5	xxxxx	xxx	3	8	
1945	4	xxxx	xxxxxxx	8	12	
1946	4	xxxx	xxxx	4	8	
1947	0		xxxx	4	4	
1948	4	xxxx	xx	2	6	
1949	5	xxxxx	xxx	3	8	
	38			37		75

Jahrgang	Anz.	Männer	Frauen	Anz.	Total	Total Jahrzehnt
1950	2	xx	xxx	3	5	
1951	4	xxxx	xxxxxx	6	10	
1952	4	xxxx	xxxxx	5	9	
1953	7	xxxxxxx	xx	2	9	
1954	5	xxxxx	x	1	6	
1955	2	xx	xx	2	4	
1956	3	xxx	xx	2	5	
1957	8	xxxxxxxx	xxxx	4	12	
1958	3	xxx	x	1	4	
1959	8	xxxxxxx	xxxx	4	12	
	46			30		76
1960	10	xxxxxxxxxx	xxxxxx	6	16	
1961	7	xxxxxxx	xxxxxxx	7	14	
1962	3	xxx	xxxxxx	6	9	
1963	4	xxxx	xxxxxx	6	10	
1964	0		xxxxxxx	7	7	
1965	4	xxxx	xxxxxx	7	11	
1966	4	xxxx	xxx	3	7	
1967	5	xxxxx	xxxxxxx	7	12	
1968	9	xxxxxxxx	x	1	10	
1969	2	xx	xxxxxxxxx	9	11	
	48			59		107
1970	4	xxxx	xxxxxxx	7	11	
1971	8	xxxxxxxx	xxxxx	5	13	
1972	4	xxxx	xxxxxxx	7	11	
1973	3	xxx	xxx	3	6	
1974	5	xxxxx	xxxxxxxxxxx	12	17	
1975	3	xxx	xxxxxxx	7	10	
1976	2	xx	xx	2	4	
1977	5	xxxxx	xxxxx	5	10	
1978	2	xx	xxx	3	5	
1979	8	xxxxxxxx	xxxx	4	12	
	44			55		99
1980	2	xx	xxx	3	5	
1981	2	xx	xxx	3	5	
1982	2	xx	xxxxxxxx	8	10	
1983	4	xxxx	xx	2	6	
1984	5	xxxxx	x	1	6	
1985	1	x	xxx	2	3	
1986	6	xxxxxx	xx	2	8	
1987	5	xxxxx	xxxxx	5	10	
1988	3	xxx	xxxxxx	6	9	
1989	4	xxxx	xxxx	4	8	
	34			36		70
1990	3	xxx	xxxxxxxxx	9	12	
1991	4	xxxx		0	4	
1992	7	xxxxxxx	xxxxxx	6	13	
1992	7	xxxxxxx	xxxx	4	11	
	21			19		40

	Männer:	Prozent-Anteil:	Prozent-Anteil:	Frauen:	
Total	278	47 Prozent	53 Prozent	312	590 Einwohner

51

Abbildung 1.3: Stamm-Blatt-Diagramm der Einwohner der Gemeinde Staldenried am 31. Dez. 1993 (Gemeinde Staldenried, 1994).

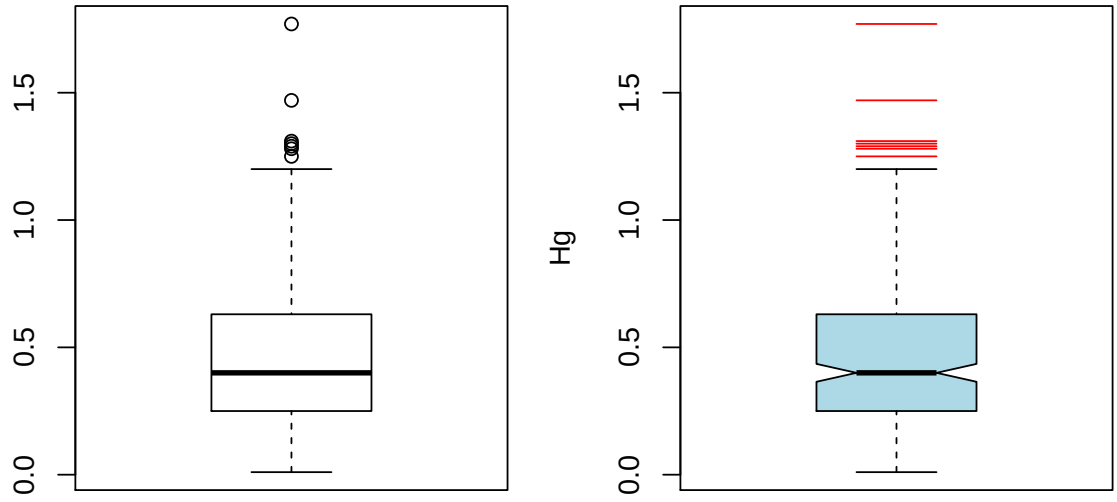


Abbildung 1.4: Boxplot: puristische und schicke Versionen. (Siehe R-Code 1.3.)

1.4 Einige Diagrammarten für multivariate Daten

Graphische Darstellungen von multivariaten Daten, wie zum Beispiel mehrere Variablen pro Objekt, erfolgen oft mit Punktwolkendiagrammen (mit `plot(x,y)` oder`pairs(...)`),

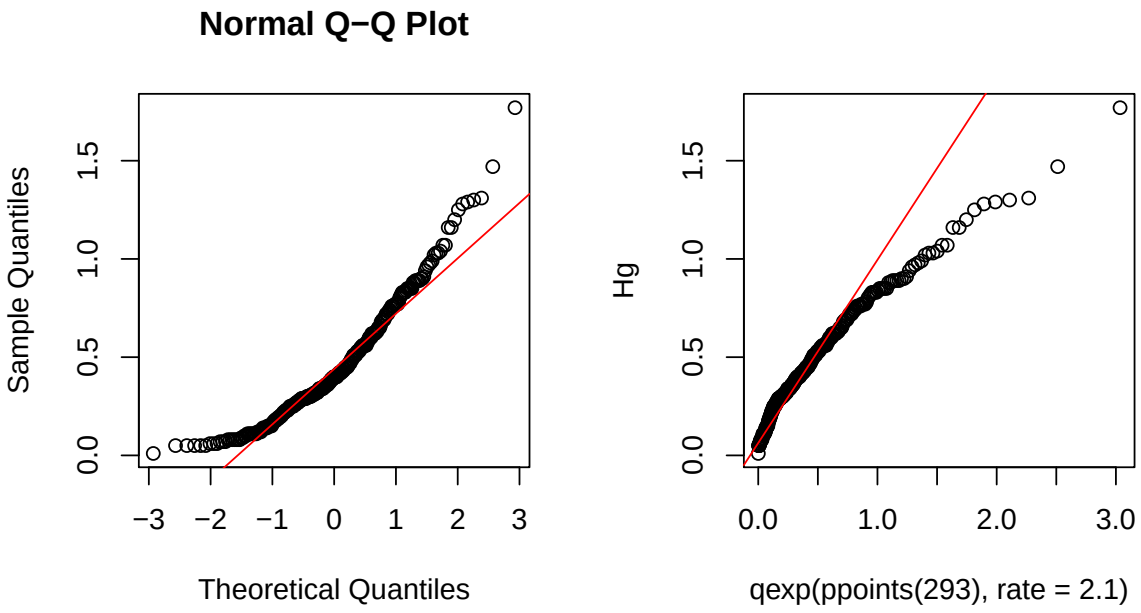


Abbildung 1.5: QQplot: Normalverteilung links und Exponentialverteilung rechts. (Siehe R-Code 1.4.)

R-Code 1.5 Alternative Quelle fürs Jahr 2005 von www.c2es.org/facts-figures/international-emissions/sector (Siehe Abbildung 1.6.)

```
dat <- c(2,15,16,32,25,10)
nam <- c('Air','Transp','Manufac','Electr','Deforest','Other')
barplot(dat,names=nam,ylab="Prozent",las=2)
dat2 <- c(2,10,12,28,26,22)
mat <- cbind(Quelle=dat,c2es=dat2)
rownames(mat) <- nam

barplot(mat, col=c(2,3,4,5,6,7), xlim=c(0.2,5), legend=nam,
        args.legend=list(bty='n') ,ylab='Prozent')
barplot(mat,beside=TRUE, col=c(2,3,4,5,6,7), xlim=c(1,30), legend=nam,
        args.legend=list(bty='n'), ylab='Prozent')
```

um die Zusammenhänge zwischen den Variablen besser aufzuzeigen (Abbildung 1.7).

Für diskrete Daten werden auch gruppierte Barplots verwendet (Diagramm unten rechts, Abbildung 1.5).

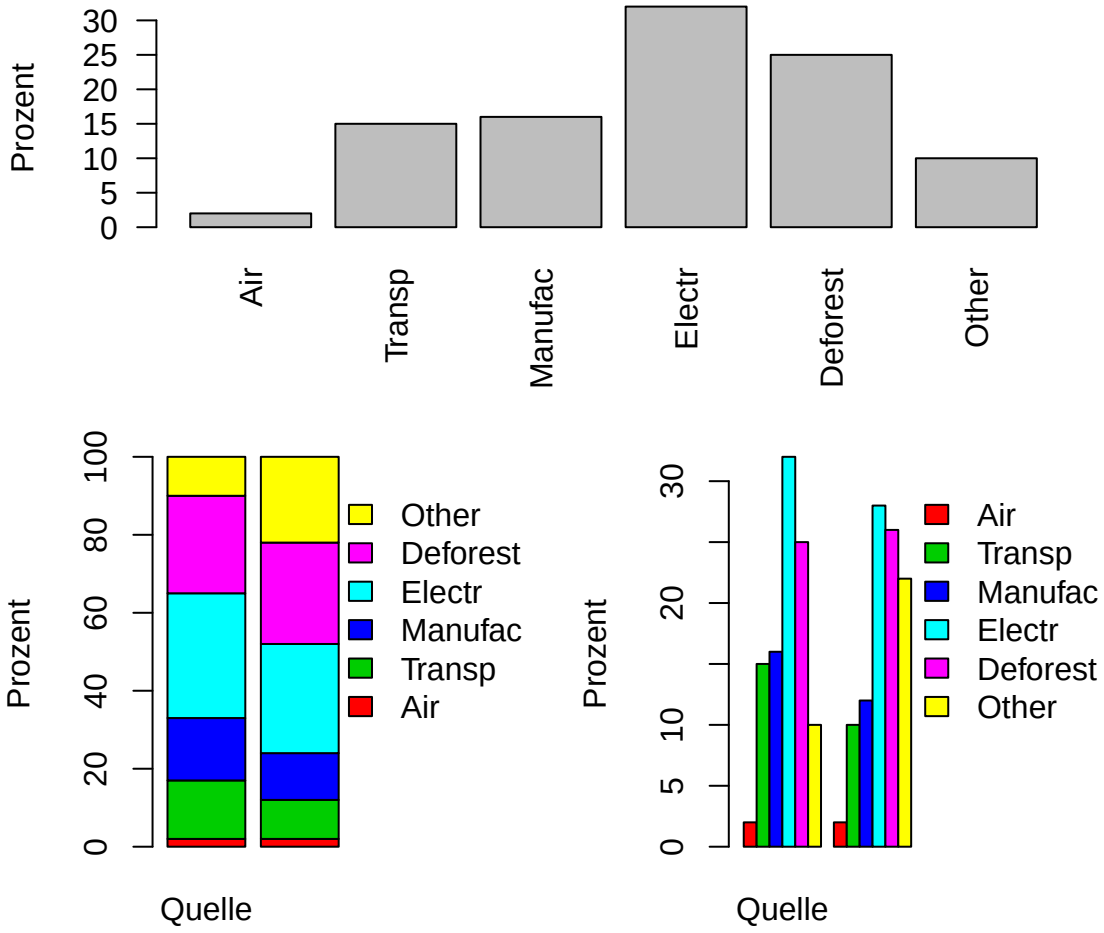


Abbildung 1.6: Barplots: gestapelt links unten, gruppiert rechts unten. (Siehe R-Code 1.5.)

1.5 Schlechte Beispiele

Abbildungen 1.8 bis 1.10 geben Beispiele von "schlechten" Darstellungen und Abbildungen. Weitere Beispiele aus renommierten Fachjournalen findet man auf www.biostat.wisc.edu/~kbroman/topten_worstgraphs/. Die einzelnen 'discussion-links' zeigen die Fehler und geben Vorschläge.

R-Code 1.6 Quecksilberdaten (multivariat).

```
metal <- read.csv( 'data/lemanHgCdZn.csv')
str( metal)
## 'data.frame': 293 obs. of  3 variables:
##  $ Hg: num  0.17 0.21 0.06 0.24 0.35 0.14 0.08 0.26 0.23 0.18 ...
##  $ Cd: num  0.23 0.37 0.14 0.3 0.56 0.3 0.17 0.44 0.39 0.26 ...
##  $ Zn: num  72.2 98.2 81.6 131 160 125 48 131 127 106 ...
pairs( metal, gap=0, main='')
# top left is as plot(Hg~Cd, data=metal)
```

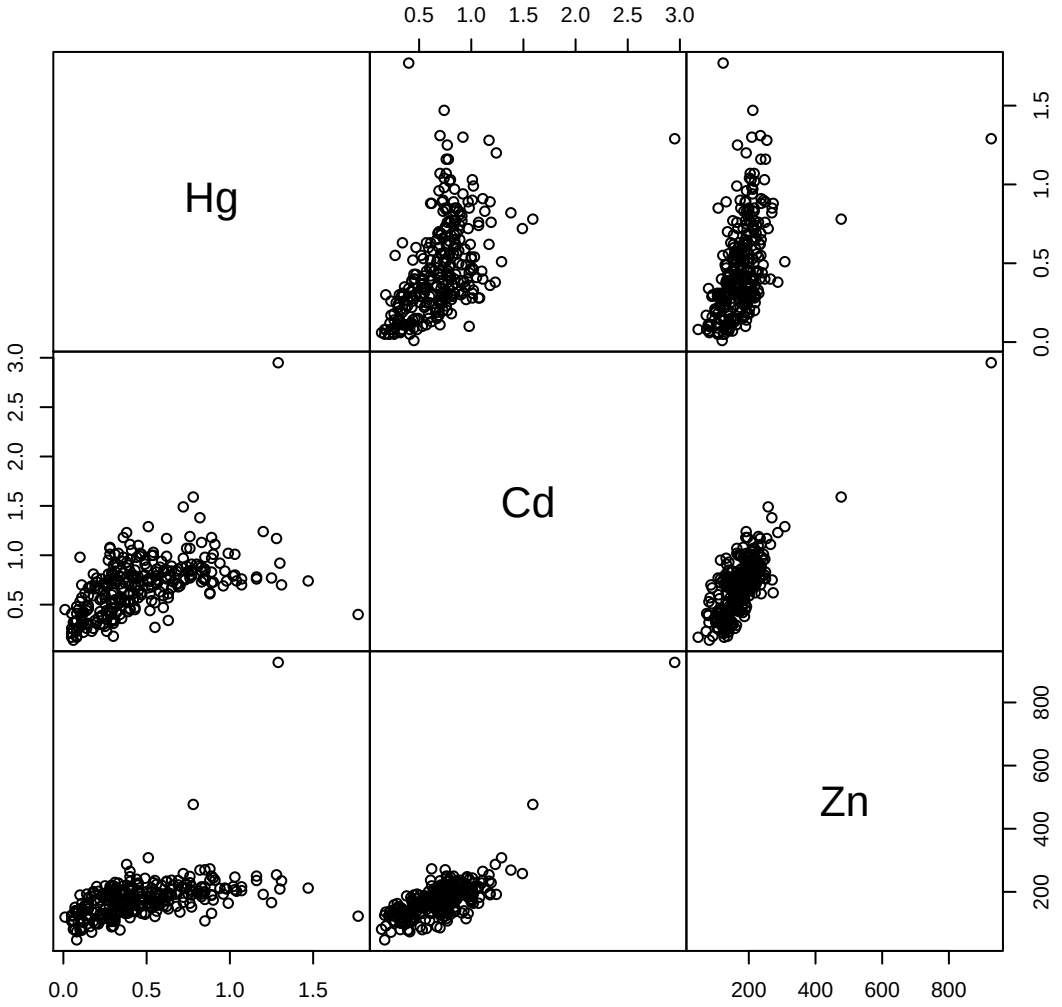


Abbildung 1.7: Punktwolkenmatrix. (Siehe R-Code 1.6.)

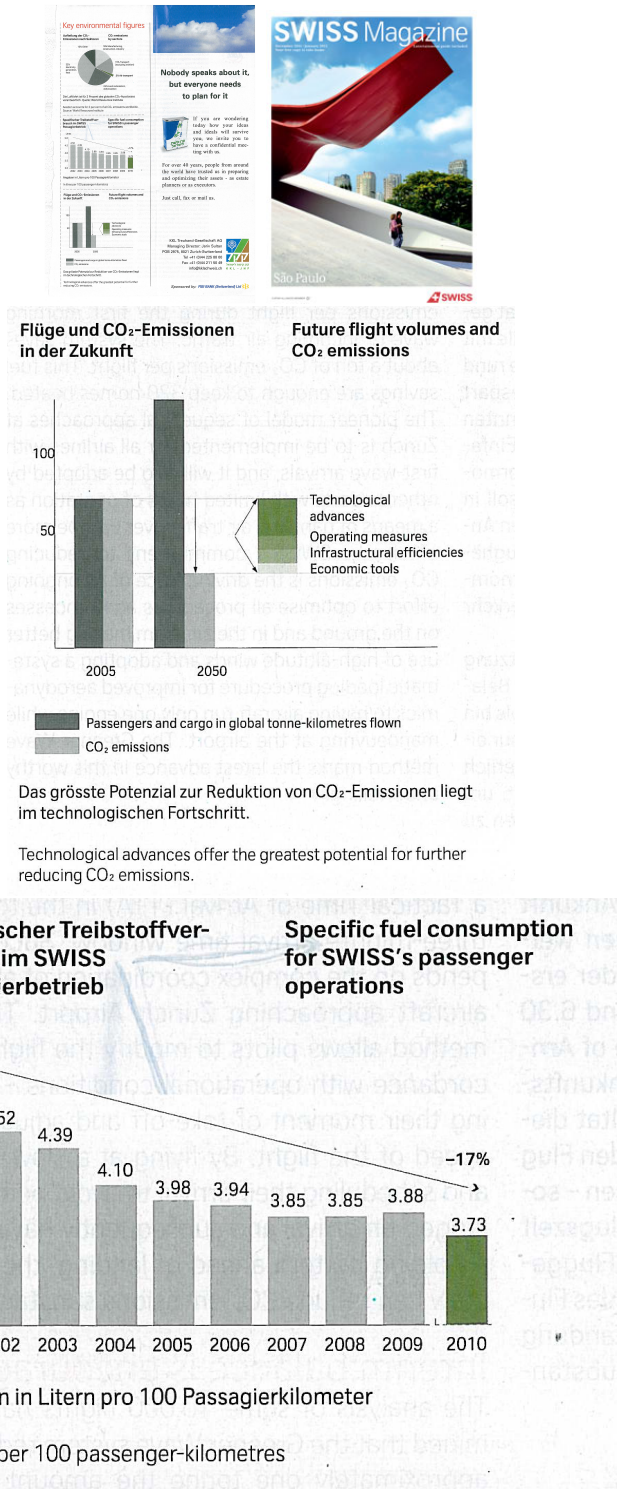


Abbildung 1.8: SWISS Magazine 10/2011,01/2012, 107

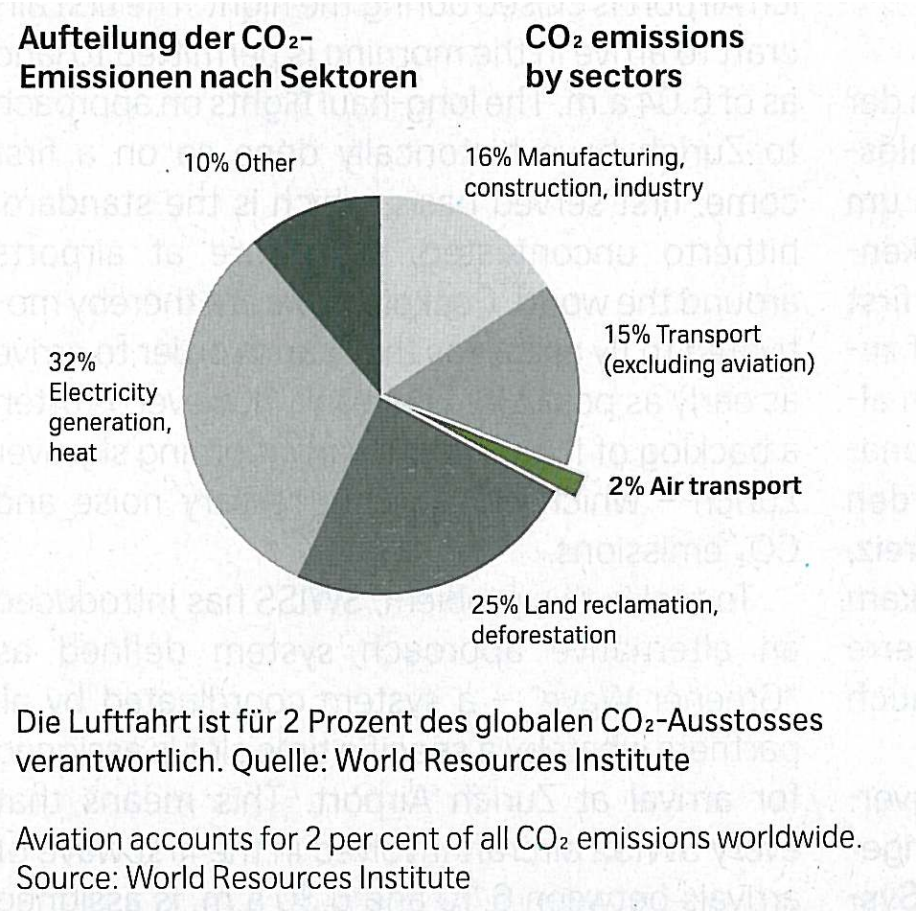


Abbildung 1.9: Kuchendiagram, aus SWISS Magazine 10/2011,01/2012, 107

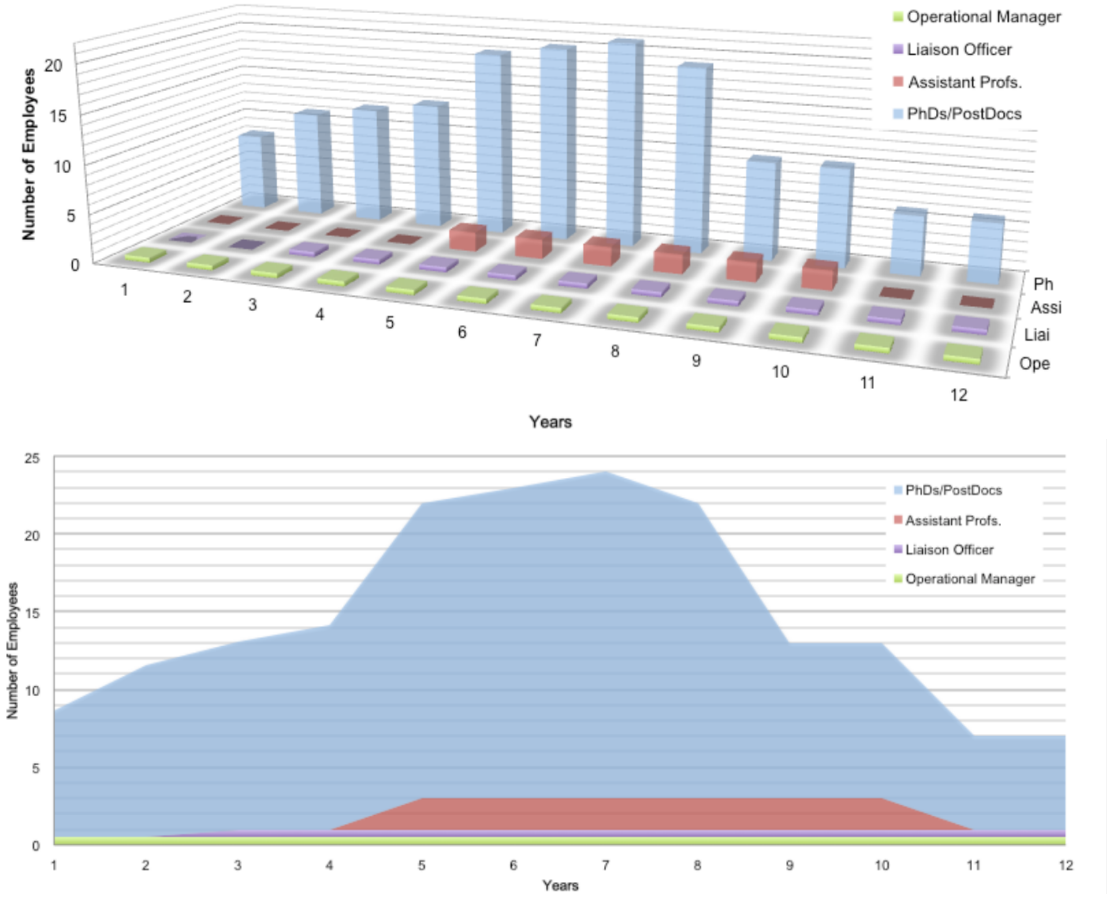


Abbildung 1.10: Schlechtes Beispiel (oben) und verbesserte, aber immer noch nicht optimale Darstellung (unten). Abbildungen aus universitären Dokumenten.

Kapitel 2

Zufallsvariablen

2.1 Grundmodell der Wahrscheinlichkeitstheorie

Um Zufallsvariablen zu definieren, benötigen wir zuerst einige Grundbegriffe.

Die Menge aller möglichen Ergebnisse eines Experiments heisst *Grundmenge* (oder *Ergebnisraum*, *sample space*). Diese Menge wird oft mit Ω bezeichnet. Eine Untermenge des Ergebnisraums heisst *Ereignis* (*event*), $A \in \Omega$.

Informell kann eine Wahrscheinlichkeit als Wert einer Abbildung P, die auf den Untermengen aus der Grundmenge definiert ist und Werte im Intervall $[0, 1]$ annimmt, betrachtet werden, d.h. $\mathrm{P}(A) \in [0, 1]$.

Um die Wahrscheinlichkeitstheorie formell einzuführen, braucht es aber noch einige technische Begriffe (*Massraum*, *σ-Algebra*, ...). Der axiomatische Aufbau von A. Kolmogorow kann aber auch zugänglich beschrieben werden:

Ein Wahrscheinlichkeitsmass muss folgende Axiome erfüllen:
$0 \leq P(A) \leq 1$, für alle Ereignisse A,
$P(\Omega) = 1$,
$P(\cup_i A_i) = \sum_i P(A_i)$, für $A_i \cap A_j = \emptyset$, $i \neq j$.

Wahrscheinlichkeiten werden oft mit sogenannten Venn-Diagrammen illustriert (Abbildung 2.1), welche auch komplexere Sachverhalte anschaulich erklären:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B), \tag{2.1}$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{2.2}$$

Wir betrachten eine Zufallsvariable als eine Funktion, die den Ergebnissen (Ereignisse) eines Zufallsexperiments Werte zuordnet, d.h. diese Werte oder Werte in Intervallen werden mit bestimmten Wahrscheinlichkeiten angenommen. Diese Werte werden als Realisationen der Zufallsvariable bezeichnet.

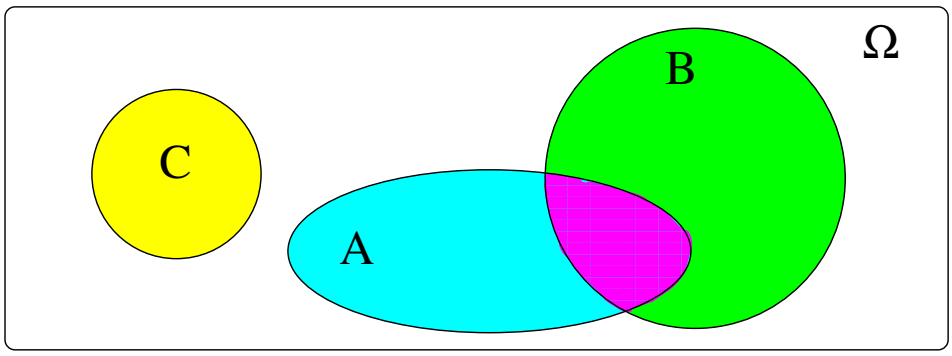


Abbildung 2.1: Venn-Diagram

Definition 2.1. Die Verteilungsfunktion (*cumulative distribution function*, *cdf*) einer Zufallsvariablen X ist

$$F(x) = F_X(x) = \mathrm{P}(X \leq x), \qquad \text{für alle } x. \tag{2.3}$$

♢

Eigenschaften 2.1. *Eine Verteilungsfunktion $F_X(x)$ ist*

i) monoton wachsend, d.h. für $x < y$ ist $F_X(x) \leq F_X(y)$;

ii) rechts-stetig, d.h. $\lim_{\epsilon \downarrow 0} F_X(x+\epsilon) = F_X(x)$, für alle $x \in \mathbb{R}$;

iii) normiert, d.h. $\lim_{x \to -\infty} F_X(x) = 0$ und $\lim_{x \to \infty} F_X(x) = 1$.

Jede Funktion $F : \mathbb{R} \to [0,1]$, welche die Eigenschaften i), ii) und iii) erfüllt, ist eine Verteilungsfunktion.

Die Zufallsvariablen werden üblicherweise mit Grossbuchstaben bezeichnet (z.B. X, Y), während man für die Realisationen die entsprechenden Kleinbuchstaben verwendet (hier somit x, y).

2.2 Diskrete Verteilungen

Eine Zufallsvariable heisst *diskret* (*discrete*), wenn sie endlich viele oder abzählbar unendlich viele Werte annehmen kann.

Beispiel 2.1. X = "Augensumme beim Wurf mit zwei Würfeln". Die Zufallsvariable X nimmt die Zahlen $2, 3, \ldots, 12$ an. Abbildung 2.2 zeigt die Wahrscheinlichkeiten und die Verteilungsfunktion. Die Verteilungsfunktion ist (wie für alle diskreten Zufallsvariablen) stückweise konstant mit Sprungstellen und den Werten, welche die Zufallsvariable annimmt. ♣

Definition 2.2. Die Wahrscheinlichkeitsfunktion (*probability mass function*, *pmf*) einer diskreten Zufallsvariablen X ist definiert durch $f_X(x) = \mathrm{P}(X = x)$. ♢

Eigenschaften 2.2. *Sei X eine diskrete Zufallsvariable mit Dichtefunktion $f_X(x)$ und Verteilungsfunktion $F_X(x)$.*

i) *Die Wahrscheinlichkeitsfunktion $f_X(x) \geq 0$ für alle $x \in \mathbb{R}$*

ii) $\sum_i f_X(x_i) = 1.$

iii) *Die Werte $f_X(x_i) > 0$ sind die "Sprünge" in x_i von $F_X(x)$.*

iv) $F_X(x_i) = \sum_{k;x_k \leq x_i} f_X(x_k).$

R-Code 2.1 Wahrscheinlichkeiten und die Verteilungsfunktion von X. (Siehe Abbildung 2.2.)

```
x <- 2:12
p <- c(1:6,5:1)/36
plot( x, p, type='h', ylim=c(0, .2),
     xlab=expression(x[i]), ylab=expression(p[i]))
points( x, p, pch = 19)
plot.ecdf( outer(1:6,1:6,"+"), ylab=expression(F[X](x)), main='')
```

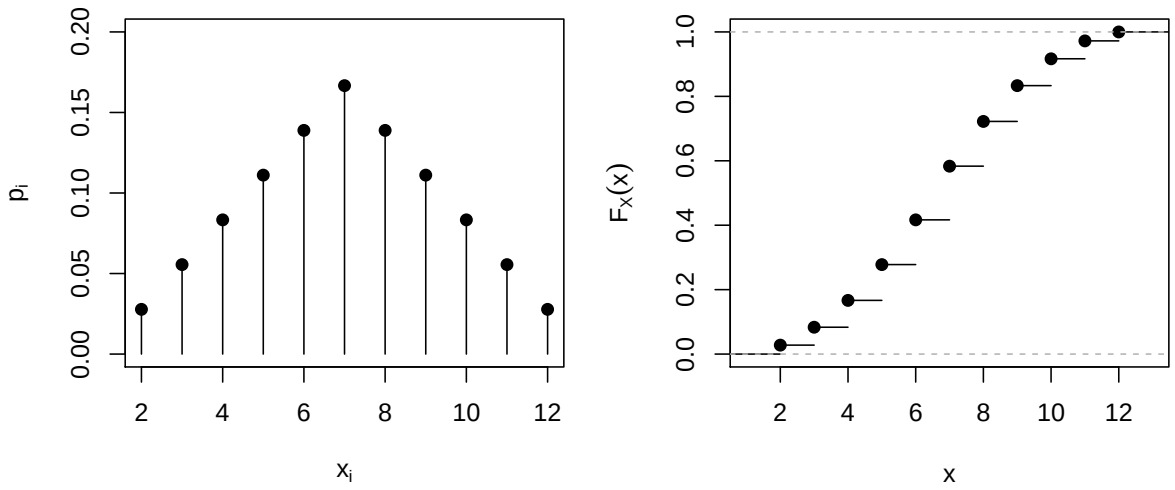


Abbildung 2.2: Wahrscheinlichkeiten (links) und Verteilungsfunktion (rechts) der Augensummen beim Wurf mit zwei Würfeln. (Siehe R-Code 2.1.)

2.3 Stetige Verteilungen

Eine Zufallsvariable heisst *stetig* (*continuous*), wenn sie (theoretisch) jeden Wert aus einem Intervall der reellen Zahlen annehmen kann.

Definition 2.3. Die Dichtefunktion (*probability density function*, *pdf*) $f_X(x)$ einer stetigen Zufallsvariablen X ist definiert durch

$$\mathrm{P}(a < X \leq b) = \int_a^b f_X(x)dx, \qquad a < b. \tag{2.4}$$

♢

Eigenschaften 2.3. *Sei X eine stetige Zufallsvariable mit Dichtefunktion $f_X(x)$ und Verteilungsfunktion $F_X(x)$.*

i) *Die Dichtefunktion erfüllt $f_X(x) \geq 0$ für alle $x \in \mathbb{R}$.*

ii) *Die Dichtefunktion $f_X(x)$ ist "fast überall" stetig.*

iii) $\displaystyle\int_{-\infty}^{\infty} f_X(x)dx = 1.$

iv) $f_X(x) = F_X'(x) = \dfrac{dF_X(x)}{dx}.$

v) $\displaystyle F_X(x) = \int_{-\infty}^{x} f_X(y)dy.$

vi) *Die Verteilungsfunktion $F_X(x)$ ist überall stetig.*

vii) $P(X = x) = 0.$

Definition 2.4. Die Quantilfunktion (*quantile function*) $Q_X(p)$ einer stetigen Zufallsvariablen X ist definiert durch

$$Q_X(p) = \inf\{x \mid F_X(x) = p\}, \quad 0 < p < 1. \tag{2.5}$$

♢

Die Quantilfunktion liefert den kleinsten Wert aus all denjenigen, die eine Wahrscheinlichkeit $\mathrm{P}(X \leq x) = p$ haben. Bei (streng) monotonen Verteilungsfunktionen entspricht die Quantilfunktion der Inversen der Verteilungsfunktion: $Q_X(p) = F_X^{-1}(p)$.

Die Schwierigkeit der Definition ist durch die möglichen Sprünge und Plateaus der Verteilungsfunktion bedingt.

Die Quantilfunktion einer diskreten Zufallsvariablen kann mathematisch zwar einfach definiert werden,

$$Q_X(p) = \inf\{x \in \mathbb{R} \mid F_X(x) \geq p\}, \tag{2.6}$$

ist aber nicht so intuitiv wie bei stetigen Zufallsvariablen.

Beispiel 2.2. Stetige Gleichverteilung $\mathcal{U}(a, b)$ ist definiert durch eine konstante Dichtefunktion über einem Intervall $[a, b]$, $a < b$, d.h. $f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{wenn } a \leq x \leq b, \\ 0, & \text{anderweitig.} \end{cases}$

Die Quantilfunktion ist $Q_X(p) = a + p(b - a)$ für $0 < p < 1$. Abbildung 2.3 zeigt Dichte- und Verteilungsfunktion der stetigen Gleichverteilung $\mathcal{U}(0, 1)$. ♣

R-Code 2.2 Dichte- und Verteilungsfunktion der stetigen Gleichverteilung (Siehe Abbildung 2.3.)

```
plot( c(-1,0,NA,0,1,NA,1,2), c(0,0,NA,1,1,NA,0,0), type='l',
     xlab='x', ylab=expression(f[X](x)))
plot( c(-1,0,1,2), c(0,0,1,1), type='l',
     xlab='x', ylab=expression(F[X](x)))
```

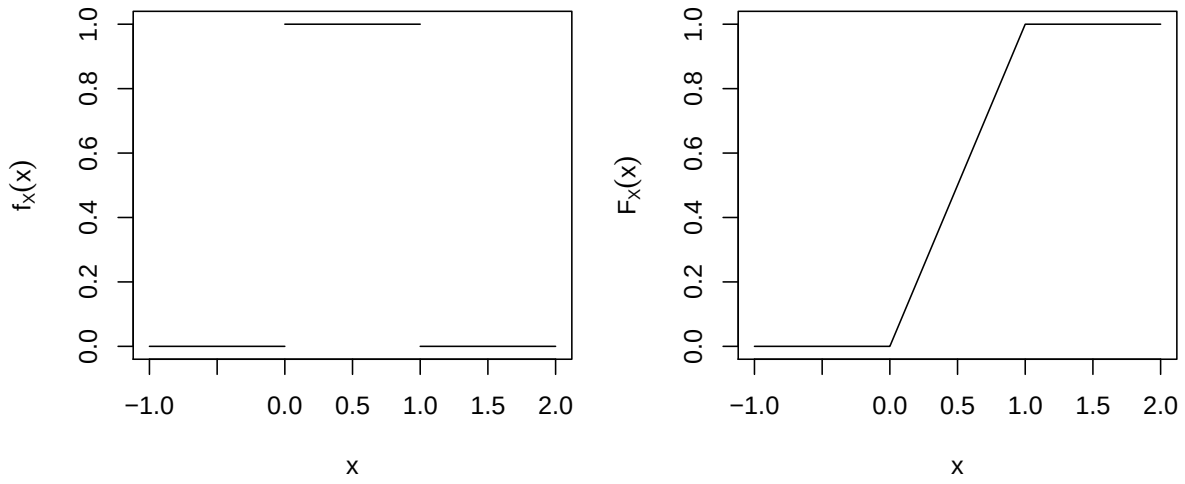


Abbildung 2.3: Dichte- und Verteilungsfunktion der stetigen Gleichverteilung (Siehe R-Code 2.2.)

2.4 Erwartungswert, Varianz und Momente

Definition 2.5. Erwartungswert (*expectation*) einer diskreten Zufallsvariablen X

$$\mathrm{E}(X) = \sum_i x_i \mathrm{P}(X = x_i)\,. \tag{2.7}$$

Erwartungswert einer stetigen Zufallsvariablen X

$$\mathrm{E}(X) = \int_{\mathbb{R}} x f_X(x) dx\,, \tag{2.8}$$

wobei $f_X(x)$ die Dichte von X. ♢

Definition 2.6. Varianz (*variance*) von X:

$$\mathrm{Var}(X) = \mathrm{E}\big((X - \mathrm{E}(X))^2\big)\,. \tag{2.9}$$

♢

Eigenschaften 2.4. *Für eine "beliebige" reelle Funktion g gilt:*

i) $\mathrm{E}(g(X)) = \sum_i g(x_i) P(X = x_i)$, *$X$ diskret,*

$\mathrm{E}(g(X)) = \int_{\mathbb{R}} g(x) f_X(x) dx$, *$X$ stetig.*

Unabhängig, ob X diskret oder stetig ist, gelten folgende Regeln:

ii) $\mathrm{Var}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2$.

iii) $\mathrm{E}(a + bX) = a + b\,\mathrm{E}(X)$.

iv) $\mathrm{Var}(a + bX) = b^2\,\mathrm{Var}(X)$.

v) $\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$, *für beliebige Zufallsvariablen* Y.

2.5 Unabhängige Zufallsvariablen

Definition 2.7. Zwei Zufallsvariablen X und Y sind unabhängig (*independent*), falls

$$\mathrm{P}(X \in A \cap Y \in B) = \mathrm{P}(X \in A)\mathrm{P}(Y \in B)\,. \tag{2.10}$$

♢

Eine Zufallsstichprobe $X_1, \ldots, X_n$ sind n unabhängige Zufallsvariablen mit derselben Verteilung F. Wir schreiben $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, wobei iid für "unabhängig und identisch verteilt" steht (*independent and identically distributed*). Die Anzahl n der Zufallsvariablen nennt man die Stichprobengrösse oder Stichprobenumfang.

Eigenschaften 2.5. *Seien X und Y zwei unabhängige Zufallsvariablen*

i) $\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y)$.

Seien $X_1, \ldots, X_n \overset{iid}{\sim} F$ *mit* $\mathrm{E}(X_1) = \mu$ *und* $\mathrm{Var}(X_1) = \sigma^2$.

ii) $\mathrm{E}\big(\frac{1}{n}\sum_{i=1}^{n} X_i\big) = \mu$.

iii) $\mathrm{Var}\big(\frac{1}{n}\sum_{i=1}^{n} X_i\big) = \frac{1}{n}\,\mathrm{Var}(X)$.

2.6 Einige spezielle diskrete Verteilungen

2.6.1 Binomialverteilung

Ein Experiment, das sich in zwei Ausgängen manifestiert (zum Beispiel "Erfolg/Nichterfolg", "Kopf/Zahl", "männlich/weiblich"), wird ein Bernoulli-Versuch genannt. Einfachheitshalber codieren wir den Grundraum mit '1' (Erfolg) und '0' (Nichterfolg).

$$\mathrm{P}(X = 1) = p, \qquad \mathrm{P}(X = 0) = 1 - p, \qquad 0 < p < 1, \tag{2.11}$$

wobei die Fälle $p = 0, 1$ pathologischer Natur wären. Somit

$$\mathrm{E}(X) = p, \qquad \mathrm{Var}(X) = p(1-p). \tag{2.12}$$

Wird nun ein Bernoulli-Versuch n mal wiederholt, ist die Zufallsvariable $X =$ "Anzahl der Erfolge" naheliegend. Die Verteilung von X heisst Binomialverteilung, dies wird mit $X \sim \mathcal{B}in(n, p)$ notiert und es gilt:

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad 0 < p < 1, \quad k = 0, 1, \ldots, n, \tag{2.13}$$

und

$$\mathrm{E}(X) = np, \qquad \mathrm{Var}(X) = np(1-p). \tag{2.14}$$

2.6.2 Poissonverteilung

Eine Zufallsvariable X, deren Wahrscheinlichkeiten mit

$$\mathrm{P}(X = k) = \frac{\lambda^k}{k!} \exp(-\lambda), \qquad 0 < \lambda, \quad k = 0, 1, \ldots, \tag{2.15}$$

gegeben sind, wird mit $X \sim \mathcal{P}oisson(n, p)$ bezeichnet. Es gilt:

$$\mathrm{E}(X) = \lambda, \qquad \mathrm{Var}(X) = \lambda. \tag{2.16}$$

Die Poissonverteilung ist eine gute Näherung für eine Binomialverteilung mit grossem n und kleinem p.

2.7 Einige spezielle stetige Verteilungen

2.7.1 Normalverteilung

Definition 2.8. Die Zufallsvariable X is normalverteilt, falls gilt

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx \tag{2.17}$$

mit Dichte

$$f(x) = f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left[\frac{(x-\mu)^2}{\sigma^2}\right]\right),$$

für alle x ($\mu \in \mathbb{R}$, $\sigma_x > 0$). Dies wird mit $X \sim \mathcal{N}(\mu, \sigma^2)$ notiert.
Die Zufallsvariable $Z = (X - \mu)/\sigma$ (z-Transformation) ist standardnormalverteilt und deren Dichte und Verteilungsfunktion wird üblicherweise mit $\phi(z)$ und $\Phi(z)$ bezeichnet. $\diamondsuit$

Eigenschaften 2.6. *i) Wenn* $X \sim \mathcal{N}(\mu, \sigma^2)$, *dann* $\dfrac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$. *Umgekehrt, wenn* $Z \sim \mathcal{N}(0,1)$, *dann* $\sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2)$.

ii) Wenn $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ *und* $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, *unabhängig, dann* $aX_1 + bX_2 \sim \mathcal{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$.

Die Stammfunktion besitzt keine geschlossene Form und die entsprechenden Wahrscheinlichkeiten müssen numerisch bestimmt werden. Früher wurden oft sogenannte "Normaltabellen" benutzt (Tabelle 2.1 gibt einen Auszug einer solchen). In der Zwischenzeit besitzen sogar "einfache" Taschenrechner die entsprechenden Funktionen.

Tabelle 2.1: Standardnormalverteilung. Die Tabelle gibt die Werte $\Phi(z_p)$.

z_p	0.00	0.02	0.04	0.06	0.08
0.0	0.500	0.508	0.516	0.524	0.532
0.1	0.540	0.548	0.556	0.564	0.571
0.2	0.579	0.587	0.595	0.603	0.610
0.3	0.618	0.626	0.633	0.641	0.648
0.4	0.655	0.663	0.670	0.677	0.684
0.5	0.691	0.698	0.705	0.712	0.719
⋮					
1.0	0.841	0.846	0.851	0.855	0.860
⋮					
1.6	0.945	0.947	0.949	0.952	0.954
1.7	0.955	0.957	0.959	0.961	0.962
1.8	0.964	0.966	0.967	0.969	0.970
1.9	0.971	0.973	0.974	0.975	0.976
2.0	0.977	0.978	0.979	0.980	0.981
⋮					
3.0	0.999	0.999	...		

R-Code 2.3 Berechnung der "z-Tabelle", siehe Tabelle 2.1.

```
y <- seq( 0, by=.02, length=5)
x <- c( seq( 0, by=.1, to=.5), 1, seq(1.6, by=.1, to=2), 3)
round( pnorm( outer( x, y, "+")), 3)
```

Beispiel 2.3. Sei $X \sim \mathcal{N}(4, 9)$. Dann

i) $$\begin{aligned} \mathrm{P}(X \le -2) &= \mathrm{P}\Big(\frac{X-4}{3} \le \frac{-2-4}{3}\Big) \\ &= \mathrm{P}(Z \le -2) = \Phi(-2) = 1 - \Phi(2) = 1 - 0.977 = 0.023\,. \end{aligned}$$

ii) $\mathrm{P}(|X-3|>2) = 1-\mathrm{P}(|X-3|\leq 2) = 1-\mathrm{P}(-2\leq X-3\leq 2)$
$= 1-(\mathrm{P}(X-3\leq 2)-\mathrm{P}(X-3\leq -2)) = 1-\Phi\left(\frac{5-4}{3}\right)+\Phi\left(\frac{1-4}{3}\right)\approx$ 0.5281. ♣

R-Code 2.4 Dichte, Verteilungsfunktion, Quantilfunktion der Standardnormalverteilung. (Siehe Abbildung 2.4.)

```
plot( dnorm, -2, 2, ylim=c(-1,2))
abline( c(0, 1), h=c(0,1), col='gray') # diag and horizontal lines
plot( pnorm, -2, 2, col=3, add=TRUE)
plot( qnorm, 0, 1, col=4, add=TRUE)
```

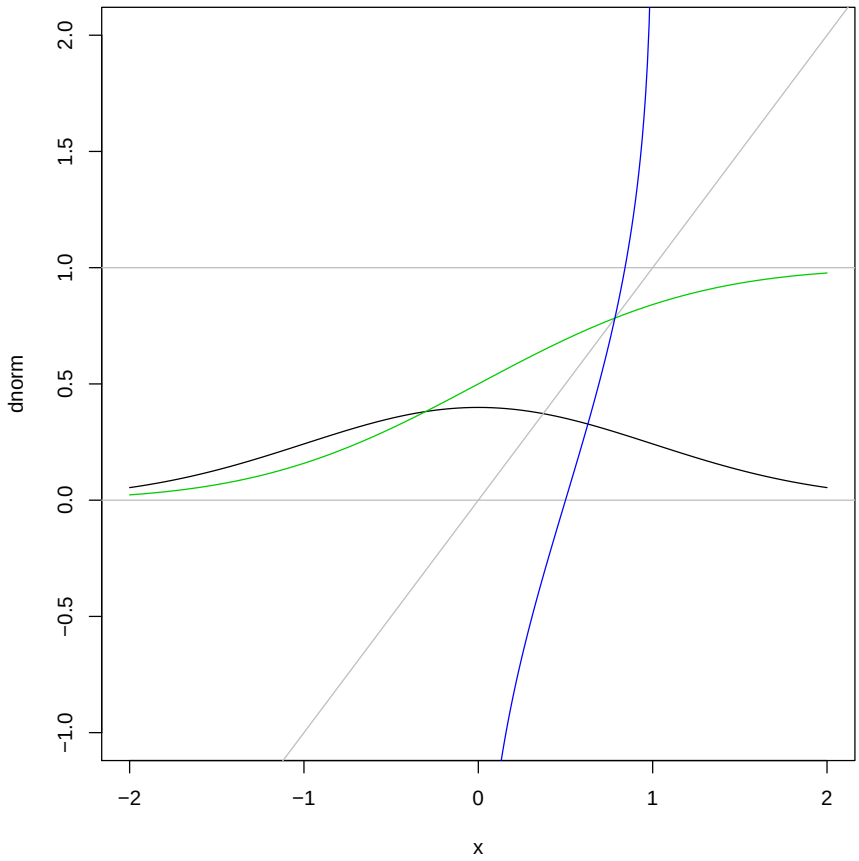


Abbildung 2.4: Dichte (schwarz), Verteilungsfunktion (grün), Quantilfunktion (blau) der Standardnormalverteilung. (Siehe R-Code 2.4.)

Wenn $n\to\infty$ konvergiert die Binomialverteilung gegen eine Normalverteilung. Somit kann die Normalverteilung $\mathcal{N}(np, np(1-p))$ als Annäherung für die Binomialverteilung $\mathcal{B}in(n,p)$ verwendet werden.

2.7.2 Chi-Quadrat-Verteilung

Es seien $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Die Verteilung

$$\chi^2_n = \sum_{i=1}^{n} Z_i^2 \tag{2.18}$$

heisst Chi-Quadrat-Verteilung (χ^2-Verteilung) mit n Freiheitsgraden. Es gilt:

$$\mathrm{E}(\chi^2_n) = n \qquad\qquad \mathrm{Var}(\chi^2_n) = 2n. \tag{2.19}$$

Die Chi-Quadrat-Verteilung wird in zahlreichen statistischen Tests verwendet.

In vielen Fällen kann die Verteilung mit $n > 50$ mit einer Normalverteilung angenähert werden, d.h. χ^2_n "ist" in etwa $\mathcal{N}(n, 2n)$. Zudem ist für χ^2_n mit $n > 30$ die Zufallsvariable $X = \sqrt{2\chi^2_n}$ näherungsweise normalverteilt, mit Erwartungswert $\sqrt{2n-1}$ und Standardabweichung 1.

R-Code 2.5 Chi-Quadrat-Verteilung für verschiedene Freiheitsgrade. (Siehe Abbildung 2.5.)

```
x <- seq( 0, to=50, length=150)
plot(x, dchisq( x, df=1), type='l', ylab='Dichte')
for (i in 1:6)
    lines( x, dchisq(x, df=2^i), col=i+1)
legend( "topright", legend=2^(0:6), col=1:7, lty=1, bty="n")
```

2.7.3 Studentsche t-Verteilung

Es seien $Z \sim \mathcal{N}(0,1)$ und $X \sim \chi^2_m$ unabhängig. Die Verteilung

$$T_m = \frac{Z}{\sqrt{X/m}} \tag{2.20}$$

heisst studentische t-Verteilung mit m Freiheitsgraden. Es gilt:

$$\mathrm{E}(T_m) = 0, \qquad \text{für } m > 1, \tag{2.21}$$

$$\mathrm{Var}(T_m) = \frac{m}{(m-2)}, \qquad \text{für } m > 2. \tag{2.22}$$

Die Dichte ist symmetrisch (um Null) und für $m \to \infty$ nähert sich die Dichte der Dichte der Normalverteilung an, siehe Abbildung 2.6.

Die t-Verteilung wird vor allem gebraucht, wenn Mittelwerte verglichen werden.

Bemerkung 2.1. Für $m = 1, 2$ besitzt die Dichte schwere "Schwänze" (*heavy-tailed*) und die Varianz existiert nicht. Realisationen dieser Zufallsvariablen manifestieren dies mit zum Teil extrem grossen Werten. Natürlich können trotzdem noch empirische Varianzen berechnet werden, siehe R-Code 2.7. □

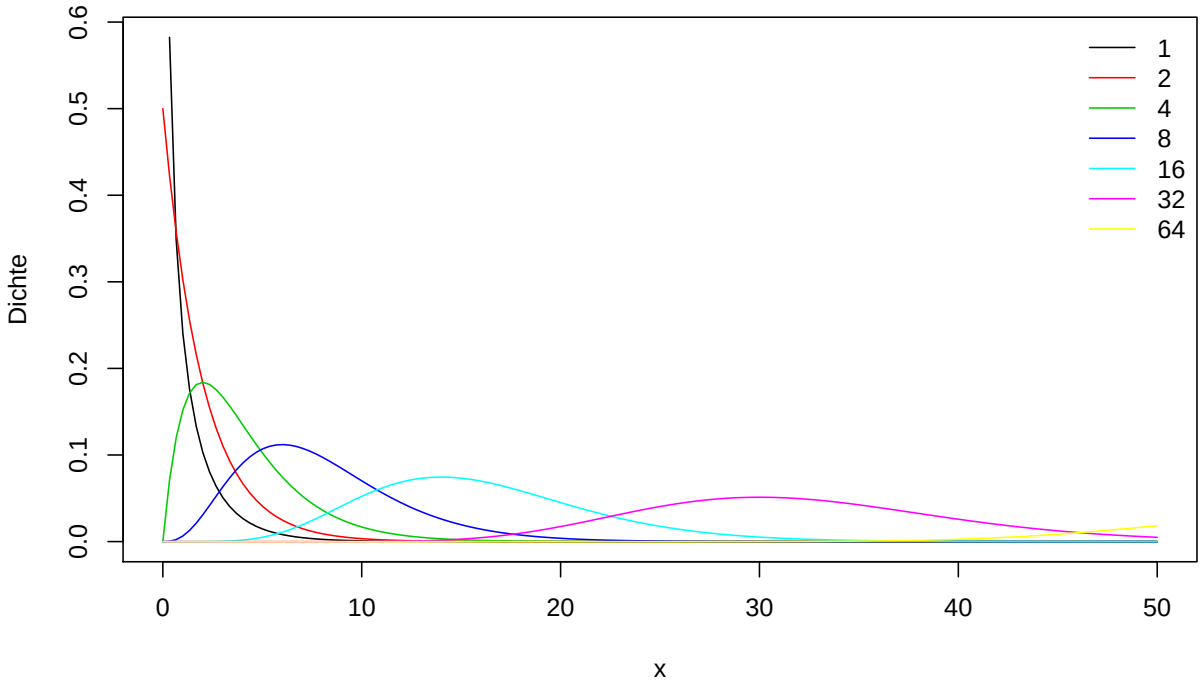


Abbildung 2.5: Chi-Quadrat-Verteilung für verschiedene Freiheitsgrade. (Siehe R-Code 2.5.)

R-Code 2.6 *t*-Verteilung für verschiedene Freiheitsgrade. (Siehe Abbildung 2.6.)

```
x <- seq( -3, to=3, length=100)
plot( x, dnorm(x), type='l', ylab='Dichte')
for (i in 0:6)
    lines( x, dt(x, df=2^i), col=i+2)
legend( "topright", legend=2^(0:6), col=2:8, lty=1, bty="n")
```

2.7.4 F-Verteilung

Es seien $X \sim \chi^2_m$ und $Y \sim \chi^2_n$ unabhängig. Die Verteilung

$$F_{m,n} = \frac{X/m}{Y/n} \tag{2.23}$$

heisst F-Verteilung mit m und n Freiheitsgraden. Es gilt:

$$\mathrm{E}(F_{m,n}) = \frac{n}{n-2}, \qquad \text{für } n > 2, \tag{2.24}$$

$$\mathrm{Var}(F_{m,n}) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, \qquad \text{für } n > 4. \tag{2.25}$$

Abbildung 2.7 zeigt für verschiedene Freiheitsgrade einige Dichten auf.

Die F-Verteilung wird vor allem gebraucht, wenn zwei empirische Varianzen miteinander verglichen werden, wie wir in Kapitel 11 sehen werden.

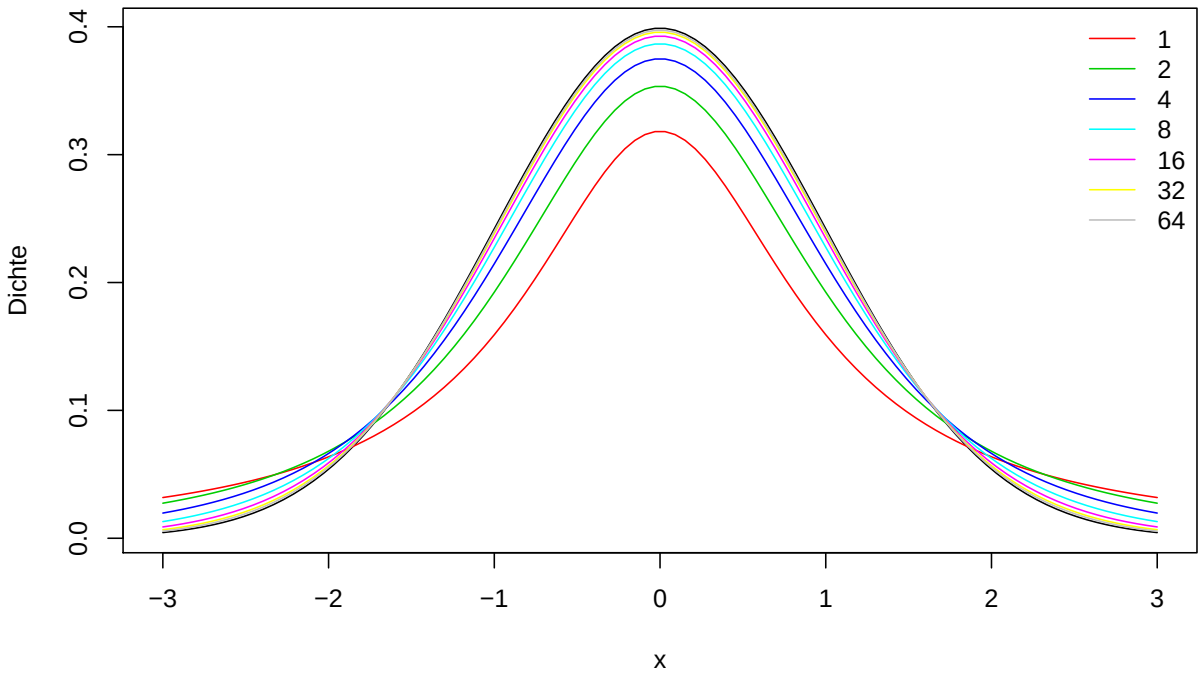


Abbildung 2.6: t-Verteilung für verschiedene Freiheitsgrade. Normalverteilung ist in schwarz. Eine Dichte mit $2^7 = 128$ Freiheitsgraden liesse die Normalverteilungskurve dicker erscheinen. (Siehe R-Code 2.6.)

R-Code 2.7 Empirische Varianzen

```
set.seed( 14)
tmp <- rt( 1000, df=1)
print( c(summary( tmp), Var=var( tmp)))
##        Min.     1st Qu.      Median        Mean     3rd Qu.
##  -190.90000    -1.12300    -0.02003     8.11100     1.00700
##        Max.         Var
##  5727.00000 37390.66447
sort( tmp)[1:10]  # many "large" values, but 2 exceptionally large
##  [1] -190.92895 -168.91968  -60.60277  -53.73566  -47.76399
##  [6]  -43.37674  -36.25175  -31.49846  -30.02890  -25.59560
sort( tmp, decreasing=TRUE)[1:10]
##  [1] 5726.53133 2083.68159  280.84799  239.75150  137.36306
##  [6]  119.15724  102.70172   47.37576   37.88695   32.44290
```

2.7.5 Beta-Verteilung

Eine Zufallsvariable X mit Dichte

$$f_X(x) = c \cdot x^{\alpha-1}(1-x)^{\beta-1}, \qquad x \in [0,1], \alpha > 0, \beta > 0 \tag{2.26}$$

R-Code 2.8 F-Verteilung für verschiedene Freiheitsgrade. (Siehe Abbildung 2.7.)

```
x <- seq(0, to=4, length=500)
df1 <- c( 1, 2, 5, 10, 50, 100, 250)
df2 <- c( 1, 50, 10, 50, 50, 300, 250)
plot( x, df( x, df1=1, df2=1), type='l', ylab='Dichte')
for (i in 2:length(df1))
    lines( x, df(x, df1=df1[i], df2=df2[i]), col=i)
legend( "topright", legend=c(expression(F[list(1,1)]),
          expression(F[list(2,50)]), expression(F[list(5,10)]),
          expression(F[list(10,50)]),expression(F[list(50,50)]),
          expression(F[list(100,300)]),expression(F[list(250,250)])),
       col=1:7, lty=1, bty="n")
```

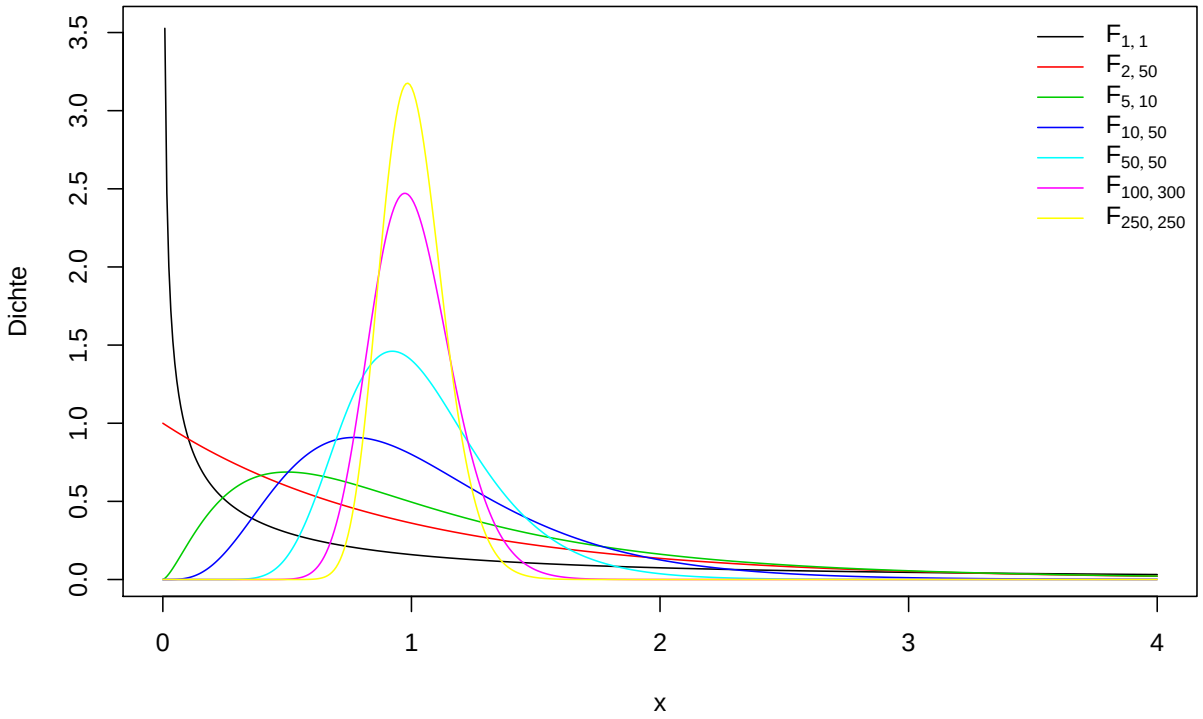


Abbildung 2.7: F-Verteilung für verschiedene Freiheitsgrade. (Siehe R-Code 2.8.)

mit c der Normalisierungskonstante, heisst Beta-Verteilung mit Parameter α und β. Wir schreiben $X \sim \mathcal{B}eta(\alpha, \beta)$. Die Normalisierungskonstante kann nicht für alle Parameter α und β in geschlossener Form geschrieben werden. Für $\alpha = \beta$ ist die Dichte symmetrisch um $1/2$ und für $\alpha > 1$ $\beta > 1$ ist diese konkav mit Modus $(\alpha - 1)/(\alpha + \beta - 2)$. Es gilt für

beliebige $\alpha > 0$, $\beta > 0$:

$$\mathrm{E}(X) = \frac{\alpha}{\alpha + \beta}, \tag{2.27}$$

$$\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}. \tag{2.28}$$

Abbildung 2.8 zeigt Dichten für verschiedene Paare (α, β).

Die Beta-Verteilung wird vor allem gebraucht, um Wahrscheinlichkeiten zu modellieren. Wir treffen diese Verteilung im Kapitel 12 wieder an.

R-Code 2.9 Dichten von betaverteilten Zufallsvariablen für verschiedene Parameterpaare (α, β). (Siehe Abbildung 2.8.)

```
p <- seq( 0, to=1, length=100)
a.seq <- c( 1:6, .8, .4, .2, 1, .5, 2)
b.seq <- c( 1:6, .8, .4, .2, 4, 4, 4)
col <- c( 1:6, 1:6)
lty <- rep( 1:2, each=6)
plot( p, dbeta( p, 1, 1),  type='l', ylab='Dichte',  xlab='x',
     xlim=c(0,1.3), ylim=c(0,3))
for ( i in 2:length(a.seq))
    lines( p, dbeta(p, a.seq[i], b.seq[i]), col=col[i], lty=lty[i])

legend("topright", legend=c(expression( list( alpha, beta)),
            paste(a.seq, b.seq)),
       col=c(NA,col), lty=c(NA, lty), cex=.9, bty='n')
```

2.8 Funktionen von Zufallsvariablen

Sei X eine Zufallsvariable mit Verteilungsfunktion $F_X(x)$. Wir definieren eine Zufallsvariable $Y = g(X)$. Die Verteilungsfunktion von Y schreibt sich

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y). \tag{2.29}$$

In vielen Fällen ist $g(\cdot)$ invertierbar (und differenzierbar) und wir erhalten

$$F_Y(y) = \begin{cases} \mathrm{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)), & \text{wenn } g^{-1} \text{ monoton wachsend,} \\ \mathrm{P}(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)), & \text{wenn } g^{-1} \text{ monoton fallend.} \end{cases} \tag{2.30}$$

Die Dichtefunktion wird durch Eigenschaft 2.3.iv) hergeleitet und ist somit

$$f_Y(y) = \left| \frac{d}{dy} g^{-1}(y) \right| f_X(g^{-1}(y)). \tag{2.31}$$

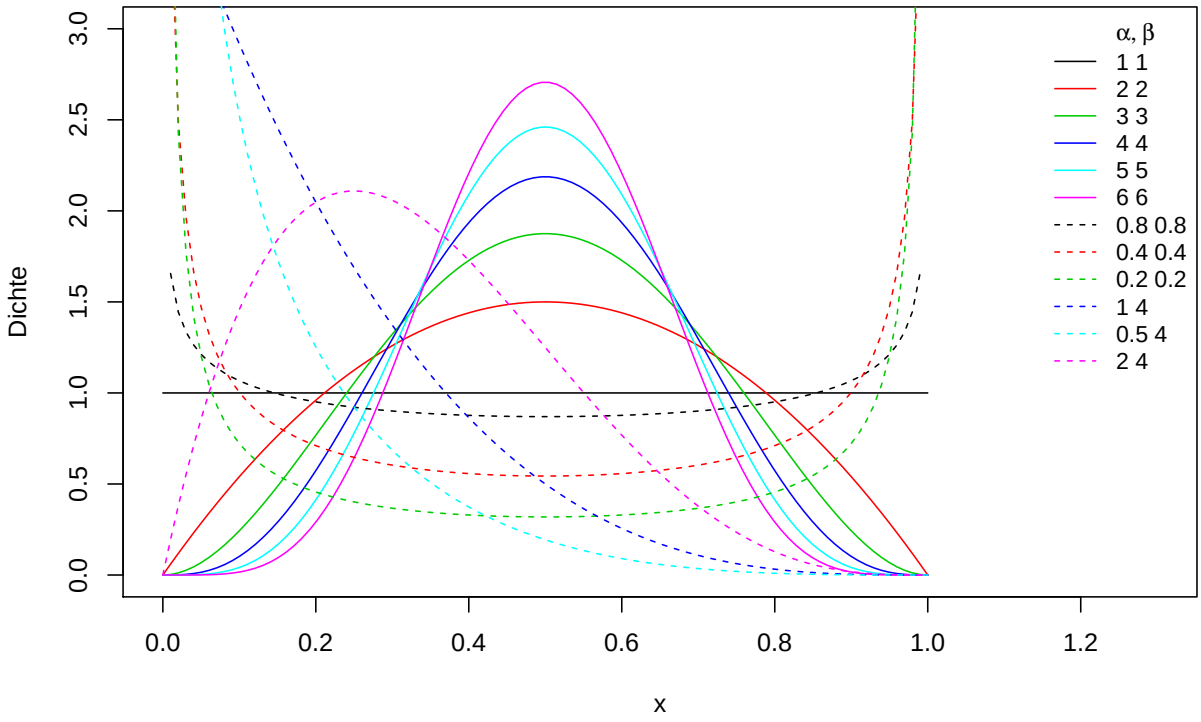


Abbildung 2.8: Dichten von betaverteilten Zufallsvariablen für verschiedene Parameterpaare (α, β). (Siehe R-Code 2.9.)

Der Erwartungswert und die Varianz von Y können mit der sogenannten Deltamethode angenähert werden. Die Idee besteht aus einer Taylorentwicklung um den Erwartungswert $\mathrm{E}(X)$:

$$g(X) \approx g(\mathrm{E}(X)) + g'(\mathrm{E}(X)) \cdot (X - \mathrm{E}(X)) \tag{2.32}$$

(zwei Terme der Taylorreihe mit Entwicklungsstelle $\mathrm{E}(X)$). Somit

$$\mathrm{E}(Y) \approx g(\mathrm{E}(X)), \qquad \mathrm{Var}(Y) \approx g'(\mathrm{E}(X))^2 \cdot \mathrm{Var}(X). \tag{2.33}$$

Beispiel 2.4. Sei X Bernoulli, und $Y = X/(1-X)$. Somit ist

$$\mathrm{E}(Y) \approx p/(1-p), \qquad \mathrm{Var}(Y) \approx \left(\frac{1}{(1-p)^2}\right)^2 \cdot p(1-p) = \frac{p}{(1-p)^3}. \tag{2.34}$$

♣

Beispiel 2.5. Sei X Bernoulli, und $Y = \log(X)$. Somit ist

$$\mathrm{E}(Y) \approx \log(p), \qquad \mathrm{Var}(Y) \approx \left(\frac{1}{p}\right)^2 \cdot p(1-p) = \frac{1-p}{p}. \tag{2.35}$$

♣

Kapitel 3

Schätzen

Ein zentraler Punkt der Statistik ist Informationen aus Beobachtungen (Messungen, Daten) zu ziehen. Dazu sind Daten (notwendigerweise) und ein *statistisches Modell* notwendig. Ein solches Modell beschreibt anhand von Verteilungen die Daten, d.h. die Daten stellen eine Realisation der Verteilungen dar. Das Modell beinhaltet unbekannte Grössen, sogenannte Parameter. Ziel einer *statistischen Schätzung* ist es, aus beobachteten Daten plausible Werte für die Parameter des Modells zu bestimmen.

Gegeben ist folgendes statistisches Modell

$$Y_i = \mu + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{3.1}$$

wobei Y_i die Beobachtung, μ eine unbekannte Konstante und ε_i sind Zufallsvariablen, die den Messfehler darstellen. Es ist sinnvoll anzunehmen, dass ε_i eine symmetrischen Dichte besitzt und dass $\mathrm{E}(\varepsilon_i) = 0$. Wir nehmen hier sogar an, dass $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Somit sind $Y_1, \ldots, Y_n$ normalverteilt mit Mittelwert μ und Varianz σ^2.

Im Normalfall sind μ und σ^2 nicht bekannt. Wie können wir aus den beobachteten Daten plausible Werte für die Parameter des Modells bestimmen?

Beispiel 3.1. Im R-Code 3.1 sind Hämoglobingehalte von Blutproben von Patienten mit Hb SS und Hb S/β Sichelzellenanämie (*sickle cell disease*) gegeben (Quelle: Hüsler and Zimmermann, 2010). Die Daten sind in Abbildung 3.1 zusammengefasst.

Für beide Krankheitstypen basiert ein einfaches statistisches Modell auf Gleichung (3.1), wobei μ den Populationsmittelwert darstellt und ε_i die indivduelle Abweichung.

Natürliche Fragen sind: Was sind plausible Populationsmittelwerte? Wie stark variieren die indivduellen Abweichungen? ♣

3.1 Punktschätzer

Für eine Schätzung werden Daten als eine Realisierung einer Zufallsstichprobe mit einer bestimmten (parametrisierten) Verteilung gesehen. Das Ziel einer Schätzung besteht darin, Aussagen über die Parameter der Verteilung zu gewinnen.

R-Code 3.1 Hämoglobingehalte von Sichelzellenanämiepatienten. (Siehe Abbildung 3.1.)

```
HbSS <- c( 7.2, 7.7, 8, 8.1, 8.3, 8.4, 8.4, 8.5, 8.6, 8.7,
           9.1, 9.1, 9.1, 9.8, 10.1, 10.3)
HbSb <- c(8.1, 9.2, 10,10.4,10.6,10.9,11.1,11.9,12.0,12.1)

par( mfcol=c(1,2))
boxplot( list(HbSS=HbSS, HbSb=HbSb), col=c(3,4))
qqnorm( HbSS, xlim=c(-2,2), ylim=c(7,12), col=3, main='')
qqline( HbSS, col=3)
tmp <- qqnorm( HbSb, plot.it=FALSE)
points( tmp, col=4)
qqline( HbSb, col=4)
# Note: it is very difficult to superimpose qqplots with car::qqPlot.
```

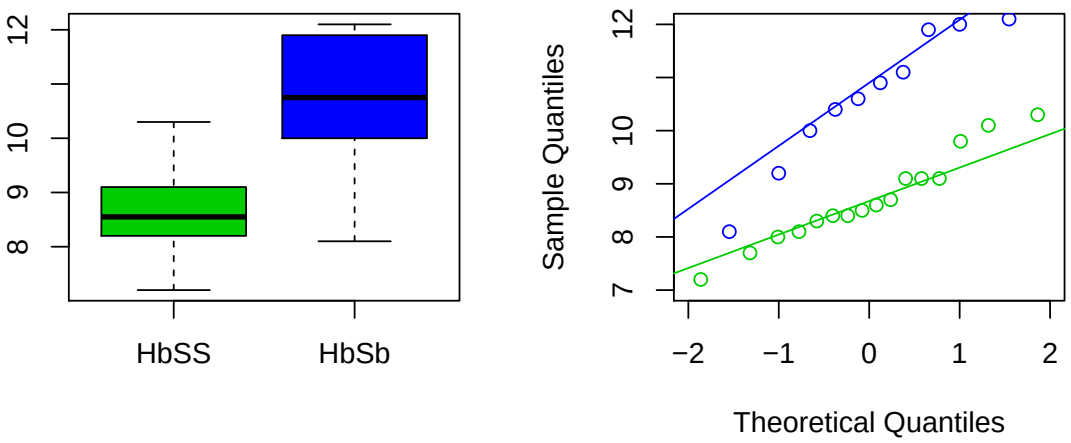


Abbildung 3.1: Hämoglobingehalte von Patienten mit Hb SS und Hb S/β Sichelzellenanämie. (Siehe R-Code 3.1.)

Definition 3.1. Eine Stichprobenfunktion ist eine (beliebige) Funktion der Zufallsstichprobe $Y_1, \ldots, Y_n$ und ist somit auch eine Zufallsvariable.

Schätzfunktion ist eine Stichprobenfunktion, die benutzt wird, um Informationen in einer Stichprobe über einen Parameter zu extrahieren.

Ein Punktschätzwert (Schätzwert, Schätzung) ist der Wert der Schätzfunktion evaluiert in $y_1, \ldots, y_n$, der Realisation der Zufallsstichprobe. ♢

Beispiel 3.2. i) $\overline{Y} = \frac{1}{n}\sum_i Y_i$ ist eine Schätzfunktion.

$\overline{y} = 8.7$ ist ein Punktschätzwert.

```
mean( HbSS)
## [1] 8.7125
mean( HbSb)
## [1] 10.63
```

ii) $S^2 = \frac{1}{n-1}\sum_i (Y_i - \overline{Y})^2$ ist eine Schätzfunktion.

$s^2 = 4.41$ oder $s = 2.1$ sind Punktschätzwerte.

```
var( HbSS)
## [1] 0.7131667
c( var(HbSb), sum( (HbSb-mean(HbSb))^2)/(length(HbSb)-1) )
## [1] 1.649 1.649
c( sd( HbSS), sqrt( var( HbSS)))
## [1] 0.844492 0.844492
```

♣

Im Englischen wird eine Schätzfunktion als *estimator*, ein Schätzwert als *estimate* und der Vorgang als *estimation* bezeichnet.

Die Schätzfunktion als auch der Schätzwert eines Parameter θ wird mit $\widehat{\theta}$ bezeichnet. Durch den Zusammenhang wird klar, welcher der beiden Fälle gemeint ist. Für unspezifische Fälle benutzt man oft θ als Parameter.

3.2 Konstruktion von Schätzfunktionen

Wir betrachten nun drei Beispiele, wie man für "beliebige" Verteilungen Schätzfunktionen konstruiert.

3.2.1 Keinste Quadrate

Seien $Y_1, \ldots, Y_n$ iid mit $\mathrm{E}(Y_i) = \mu$. Der Kleinste-Quadrate-Schätzer (*least squares estimator*) basiert auf

$$\widehat{\mu} = \widehat{\mu}_{\mathrm{KQ}} = \operatorname*{argmin}_{\mu} \sum_{i=1}^{n} (Y_i - \mu)^2, \tag{3.2}$$

somit die Schätzfunktion $\widehat{\mu}_{\mathrm{KQ}} = \overline{Y}$ und Schätzwert $\widehat{\mu}_{\mathrm{KQ}} = \overline{y}$.

3.2.2 Momentenmethode

Die Momentenmethode (*method of moments*) basiert auf folgender Idee. Man drückt die Parameter der Verteilung in Abhängigkeit von den Momenten der Verteilung aus. Anschliessend setzt man die empirischen Momente anstatt der theoretischen Momente in die Gleichungen ein und erhält so die Momentenschätzer. Indem man die beobachteten Werte einer Stichprobe in den Momentenschätzer einsetzt, erhält man eine Schätzung des entsprechenden Parameters. Somit haben wir

$$\mu = \mathrm{E}(Y)\,, \qquad \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}\,, \tag{3.3}$$

$$\mu_2 = \mathrm{E}(Y^2)\,, \qquad \widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2\,. \tag{3.4}$$

Beispiel 3.3. Sei $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{E}(\lambda)$

$$\mathrm{E}(Y) = 1/\lambda, \qquad \overline{Y} = 1\Big/\widehat{\lambda} \qquad \widehat{\lambda} = \widehat{\lambda}_{\mathrm{ML}} = \frac{1}{\overline{Y}}\,. \tag{3.5}$$

Somit ist der Schätzungwert von λ der Wert $1/\overline{y}$. ♣

Beispiel 3.4. Sei $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F$ mit Erwartungswert μ und Varianz σ^2. Da $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}(Y)^2$ haben wir die Schätzfunktion

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2. \tag{3.6}$$

♣

3.2.3 Likelihood Methode

Wir betrachten die Dichtefunktion (bei stetigen Zufallsvariablen) oder die Wahrscheinlichkeitsfunktion (bei diskreten Zufallsvariablen) als Funktion des Parameters θ, d.h.

$$f_Y(y) = f_Y(y;\theta) \qquad \longrightarrow \qquad L(\theta) := f_Y(y;\theta) \tag{3.7}$$

$$p_i = \mathrm{P}(Y = y_i) = \mathrm{P}(Y = y_i;\theta) \qquad \longrightarrow \qquad L(\theta) := \mathrm{P}(Y = y_i;\theta). \tag{3.8}$$

Für eine gegebene Verteilung nennt man $L(\theta)$ die *Likelihoodfunktion* (*likelihood*).

Definition 3.2. Der Maximum-Likelihood-Schätzer (*maximum likelihood estimator*) $\widehat{\theta}_{\mathrm{ML}}$ eines Parameters θ basiert auf Maximierung der Likelihoodfunktion, d.h.

$$\widehat{\theta}_{\mathrm{ML}} = \operatorname*{argmax}_{\theta} L(\theta). \tag{3.9}$$

♢

Da eine Zufallsstichprobe unabhängig und identisch verteilte Zufallsvariablen beinhaltet, ist die Likelihoodfunktion das Produkt der individuellen Dichten. Da $\widehat{\theta}_{\text{ML}} = \text{argmax}_\theta\, L(\theta) = \text{argmax}_\theta\, \ell(\theta)$ wird bei der Maximierung oft mit der Log-Likelihoodfunktion $\ell(\theta) := \log(L(\theta))$ gearbeitet.

Beispiel 3.5. Sei $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{E}(\lambda)$, somit

$$L(\lambda) = \prod_{i=1}^{n} f_Y(y_i) = \prod_{i=1}^{n} \lambda \exp(-\lambda y_i) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} y_i)\,. \tag{3.10}$$

Da

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{dn\log(\lambda) - \lambda \sum_{i=1}^{n} y_i}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i \overset{!}{=} 0 \tag{3.11}$$

$$\widehat{\lambda} = \widehat{\lambda}_{\text{ML}} = \frac{n}{\sum_{i=1}^{n} y_i} = \frac{1}{\overline{y}}. \tag{3.12}$$

In diesem Fall (wie auch in anderen) ist $\widehat{\lambda}_{\text{ML}} = \widehat{\lambda}_{\text{MM}}$. ♣

3.3 Vergleich von Schätzfunktionen

Ein Schätzer $\widehat{\theta}$ eines Parameters θ ist erwartungstreu (unverzerrt, *unbiased*), wenn

$$\text{E}(\widehat{\theta}) = \theta, \tag{3.13}$$

ansonsten ist er verzerrt (*biased*).

Beispiel 3.6. $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$

i) $\overline{Y}$ ist erwartungstreu für μ, denn

$$\text{E}(\overline{Y}) = \frac{1}{n} n\, \text{E}(Y_i) = \mu\,. \tag{3.14}$$

ii) $S^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$ ist erwartungstreu für σ^2, denn

$$\sum_{i=1}^{n} (Y_i - \mu)^2 = (n-1)S^2 + n(\overline{Y} - \mu)^2 \tag{3.15}$$

$$\Rightarrow n \cdot \sigma^2 = (n-1)\,\text{E}(S^2) + n \cdot \frac{\sigma^2}{n} \tag{3.16}$$

$$\Rightarrow n \cdot \sigma^2 - \frac{n\sigma^2}{n} = (n-1)\,\text{E}(S^2) \tag{3.17}$$

$$\Rightarrow \sigma^2 = \text{E}(S^2)\,. \tag{3.18}$$

iii) $\widehat{\sigma^2} = \frac{1}{n}\sum(Y_i - \overline{Y})^2$ ist verzerrt, denn

$$E(\widehat{\sigma^2}) = \frac{1}{n}(n-1)\underbrace{\mathrm{E}\Big(\frac{1}{n-1}\sum(Y_i-\overline{Y})^2\Big)}_{\mathrm{E}(S^2)=\sigma^2} = \frac{n-1}{n}\sigma^2. \tag{3.19}$$

Man nennt den Wert

$$\mathrm{E}(\widehat{\sigma^2}) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2 \tag{3.20}$$

den *Bias* (*bias*). ♣

Eine weitere Möglichkeit, Schätzer zu vergleichen, ist die *mittlere quadratische Abweichung* (*mean squared error*)

$$\mathrm{MSE}(\widehat{\theta}) = \mathrm{E}((\widehat{\theta} - \theta)^2). \tag{3.21}$$

Die mittlere quadratische Abweichung kann auch als $\mathrm{MSE}(\widehat{\theta}) = \mathrm{bias}(\widehat{\theta})^2 + \mathrm{Var}(\widehat{\theta})$ geschrieben werden.

Beispiel 3.7. $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$

$$\mathrm{MSE}(\overline{Y}) = \mathrm{bias}(\overline{Y})^2 + \mathrm{Var}(\overline{Y}) = 0 + \frac{\sigma^2}{n} \tag{3.22}$$

$$\mathrm{MSE}(S^2) = \mathrm{Var}(S^2) = \frac{\sigma^4}{(n-1)^2}\,\mathrm{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = \frac{\sigma^4}{(n-1)^2}(2n-2) = \frac{2\sigma^4}{n-1}. \tag{3.23}$$

wobei wir $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ benutzten.

Analog zeigt man, dass $\mathrm{MSE}(\widehat{\sigma^2})$ kleiner ist als (3.23). Der Schätzer $\frac{n-1}{n+1}S^2$ besitzt den kleinsten MSE. ♣

3.4 Intervallschätzer

Seien $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, σ^2 bekannt und somit $\dfrac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. Daher

$$1 - \alpha = \mathrm{P}\Big(z_{\alpha/2} \le \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \le z_{1-\alpha/2}\Big) \tag{3.24}$$

$$= \mathrm{P}\Big(z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \overline{Y} - \mu \le z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\Big) \tag{3.25}$$

$$= \mathrm{P}\Big(-\overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le -\mu \le -\overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\Big) \tag{3.26}$$

$$= \mathrm{P}\Big(\overline{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \ge \mu \ge \overline{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\Big). \tag{3.27}$$

Definition 3.3. Seien $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, σ^2 bekannt. Das Intervall

$$\Big[\overline{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\Big] \tag{3.28}$$

ist ein exaktes $(1-\alpha)$-Konfidenzintervall für den Parameter μ. ♢

Die Interpretation eines (exakten) Konfidenzintervalls ist folgendermassen: Wenn sehr viele Realisationen aus derselben Zufallsstichprobe gezogen werden, so überdecken im *Mittel* $(1-\alpha)\cdot 100\%$ der Konfidenzintervalle den wahren Parameter μ.

Das Konfidenzintervall beinhaltet keine Zufallsvariablen und es können somit keine Wahrscheinlichkeitsaussagen gemacht werden.

Falls die Standardabweichung σ unbekannt ist, muss der Ansatz geändert werden, und zwar indem für σ ein Punktschätzwert verwendet wird, typischerweise $S = \sqrt{S^2}$, $S^2 = \frac{1}{n-1}\sum_i (Y_i - \overline{Y})^2$. Da aber $\dfrac{\overline{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}$, müssen die entsprechenden Quantile geändert werden:

$$1-\alpha = \mathrm{P}\Big(t_{n-1,\alpha/2} \le \frac{\overline{Y}-\mu}{S/\sqrt{n}} \le t_{n-1,1-\alpha/2} \Big). \tag{3.29}$$

Definition 3.4. Seien $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Das Intervall

$$\Big[\overline{Y} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}},\ \overline{Y} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} \Big] \tag{3.30}$$

ist ein exaktes $(1-\alpha)$-Konfidenzintervall für den Parameter μ. ♢

Die Konfidenzintervalle sind, wie in den zwei letzten Definitionen aufgezeigt, sind durch Zufallsvariablen gegeben (Funtionen von $Y_1, \ldots, Y_n$). Ähnlich zu Schätzfunktionen und Schätzwerten, die empirischen Konfidenzintervalle werden mit der entsprechenden Realisation $y_1, \ldots, y_n$ der Zufallsstichprobe berechnet. In der Folge werden wir (empirische) Konfidenzintervalle mit blau unterlegten Textblöcken darstellen, wie hier gezeigt wird.

Konfidenzintervall für den Mittelwert μ

Unter Annahme der Normalverteilung ist

$$\Big[\overline{y} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}},\ \overline{y} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \Big] \tag{3.31}$$

ein exaktes $(1-\alpha)$-Konfidenzintervall und

$$\Big[\overline{y} - z_{1-\alpha/2}\frac{s}{\sqrt{n}},\ \overline{y} + z_{1-\alpha/2}\frac{s}{\sqrt{n}} \Big] \tag{3.32}$$

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

Beispiel 3.8. Seien $Y_1, \ldots, Y_4 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Der R-Code 3.2 und die Abbildung 3.2 zeigen 100 Konfidenzintervalle basierend auf (3.28) (oben) und (3.30) (unten). Da n klein ist, ist die Differenz zwischen der Normal- und der t-Verteilung ausgeprägt. Dies wird deutlich, wenn man

$$\left[\overline{Y} - z_{1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{Y} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right] \tag{3.33}$$

als Annäherung benutzt (mitte). ♣

Konfidenzintervall für die Varianz σ^2

Unter Annahme der Normalverteilung ist

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right] \tag{3.34}$$

ein exaktes $(1-\alpha)$-Konfidenzintervall für σ^2.

R-Code 3.2: 100 Konfidenzintervalle für den Parameter $\mu = 0$ mit $\sigma = 1$ bekannt und unbekannt. (See Figure 3.2.)

```
set.seed( 1)
ex.n <- 100    #  100 Konfidenzintervalle
alpha <- .05   #  95\% Konfidenzintervalle
n <- 4         #  Stichprobengroesse
mu <- 0
sigma <- 1
sample <- array( rnorm( ex.n * n, mu, sigma), c(n,ex.n))
yl <- c( -2.7, 2.7)              # Gleiche y-Achsen fuer alle
ybar <- apply( sample, 2, mean)  # Mittelwerte
# Erster Teil:
sigmaybar <- sigma/sqrt(n)
plot( 1:ex.n, 1:ex.n, type='n', ylim=yl, xaxt='n',
     ylab=expression(sigma~bekannt),
     main='n = 4, alpha = 0.05, sigma bekannt')
abline(h=0)
for ( i in 1:ex.n){
  ci <- ybar[i] + sigmaybar * qnorm(c(alpha/2,1-alpha/2))
```

```
  lines( c(i,i), ci, col=ifelse( ci[1]>0|ci[2]<0, 2, 1))
}
# Zweiter Teil:
sybar <- apply(sample, 2, sd)/sqrt(n)
plot( 1:ex.n, 1:ex.n, type='n', ylim=yl, xaxt='n',
     ylab="Gauss'sche Annaeherung",
     main="n = 4, alpha = 0.05, Gauss'sche Annaeherung")
abline(h=0)
for ( i in 1:ex.n){
  ci <- ybar[i] + sybar[i] * qnorm(c(alpha/2, 1-alpha/2))
  lines( c(i,i), ci, col=ifelse( ci[1]>0|ci[2]<0, 2, 1))
}
# Dritter Teil:
plot(1:ex.n, 1:ex.n, type='n', ylim=yl, xaxt='n',
     ylab='t-Verteilung',
     main="n = 4, alpha = 0.05, t-Verteilung")
abline(h=0)
for ( i in 1:ex.n){
  ci <- ybar[i] + sybar[i] * qt(c(alpha/2,1-alpha/2), n-1)
  lines( c(i,i), ci, col=ifelse( ci[1]>0|ci[2]<0, 2, 1))
}
```

Die Überdeckungswahrscheinlichkeit eines exakten Konfidenzinterval beträgt genau $1-\alpha$. Bei approximativen Konfidenzintervallen ist das nicht der Fall.

Beispiel 3.9. Seien $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Die Schätzfunktion S^2 für den Parameter σ^2 ist so, dass $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, d.h. eine Chi-Quadrat-Verteilung mit $n-1$ Freiheitsgraden. Daher

$$1-\alpha = \mathrm{P}\Big(\chi^2_{n-1,\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,1-\alpha/2} \Big) \tag{3.35}$$

$$= \mathrm{P}\Big(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \geq \sigma^2 \geq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} \Big), \tag{3.36}$$

wobei $\chi^2_{n-1,p}$ das p-Quantil der Chi-Quadrat-Verteilung mit $n-1$ Freiheitsgraden ist. Das entsprechende exakte $(1-\alpha)$-Konfidenzintervall besitzt nicht mehr die Form $\widehat{\theta} \pm q_{1-\alpha/2}\mathrm{sd}(\widehat{\theta})$, auch weil die Chi-Quadrat-Verteilung nicht symmetrisch ist.

Für die Hb SS Daten mit empirischer Varianz 0.71, wir haben das Konfidenzintervall $[0.39, 1.71]$, berechnet mit `(16-1)*var(HbSS)/qchisq( c(.975, .025), df=16-1)`. ♣

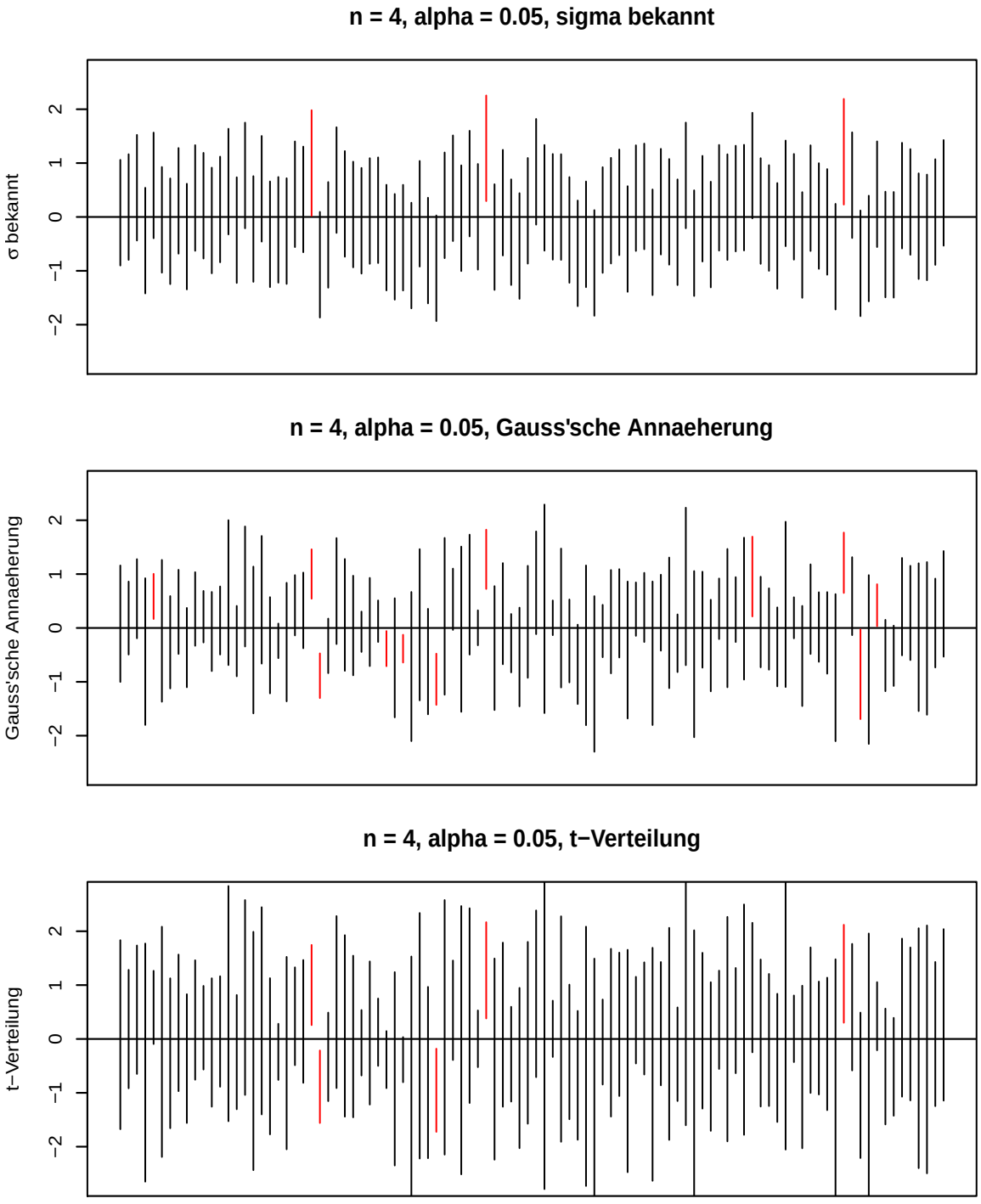


Abbildung 3.2: Normale- und t-Konfidenzintervalle für den Parameter $\mu = 0$ mit $\sigma = 1$ bekannt (oben) und unbekannt (mitte und unten). Stichprobengrösse ist $n = 4$. Konfidenzintervalle, welche die Null nicht abdecken, sind in rot. (Siehe R-Code 3.2.)

Kapitel 4

Statistische Testverfahren

Selbst mit einer idealen Münze werfen wir nicht immer genau $n/2$ mal Zahl (`rbinom(1, size=1, prob=1/2)` ist eine ideale "Münze" und `rbinom(1, size=n, prob=1/2)` entspricht n Würfen). Umgekehrt kann man sich fragen, ob wir aus 10 mal Zahl in 17 Würfen schliessen können, dass es sich nicht um eine ideale Münze handelt?

Wir werden einen formellen statistischen Ablauf formulieren, um solche Fragen zu beantworten.

4.1 Allgemeines Konzept eines statistischen Tests

Beispiel 4.1. Eine Maturandin hat durch Beobachtung ihres Essverhaltens im Verlauf eines Menstruationszyklus fest gestellt, dass sie einige Tage vor der Menstruation mehr Hunger hatte als sonst, jedoch nach der Menstruation manchmal fast keinen Appetit hatte. Basierend auf diesen Beobachtungen wurde die Hypothese aufgestellt, dass ein wiederkehrendes Muster im Essverhalten einer Frau erkennbar ist, welches sich durch den Menstruationszyklus erklären lässt

Im Rahmen der Maturitätsarbeit wurde eine Erhebung erstellt und während zwei Monaten haben 17 Frauen täglich notiert, was und wie viel sie im Laufe des Tages gegessen hatten. Für die Erfassung wurde ein Fragebogen verwendet, welchen die Probandinnen täglich ausfüllten. Darin wurde festgehalten von welcher Nahrungsmittelgruppe (Gemüse, Früchte, Kohlenhydrate, Milchprodukte, Proteine) wie viele Portionen zum Frühstück, Mittagessen und Abendessen eingenommen wurden. Mit Hilfe eines Merkblatts wurde den Probandinnen erklärt, wie sie die Nahrungsmittel einteilen müssen und wie viel als eine Portion gilt.

Der Menstruationszyklus wird in vier Phasen aufgeteilt (vorher/während/nacher und normal), und die Daten der jeweiligen Phase sind in `Essmenge.normal`, `Essmenge.waehrend` usw. gegeben.

Die Hypothese ist, dass der Populationsmittelwert von 9 Portionen verschieden ist. (Der beobachtete Mittelwert ist 7.73 mit einer Standardabweichung von 2.06.) Der R-

Code zur Berechnung der Kennzahlen und zur Konstruktion der Abbildung 4.1 ist in 4.1 gegeben. ♣

R-Code 4.1 Essdaten (Siehe Abbildung 4.1.)

```
Essmenge.normal
##  [1]  5.17  9.00 12.69  6.50  5.76  7.60  6.40  5.19  9.90  7.60
## [11]  6.40  8.39  6.06  8.40 10.80  8.80  6.70
print( me <- mean( Essmenge.normal))  # Zuweisung als "Argument"
## [1] 7.727059
(se <- sd( Essmenge.normal))          # equivalent zu print(se <- ... )
## [1] 2.064551
plot( cbind( Essmenge.normal, 0))
rug( en, ticksize=0.3)
abline( v=me, col=3)
abline( v=9, lwd=3, col=2)
 # t- KI in gruen, Normal-KI (approximation!) in blau
lines( me + qt(c(.025,.975),16)*se/sqrt(17), c(.1,.1), col=3, lwd=5)
lines( me + qnorm(c(.025,.975))*se/sqrt(17), c(-.1,-.1), col=4, lwd=2)
```

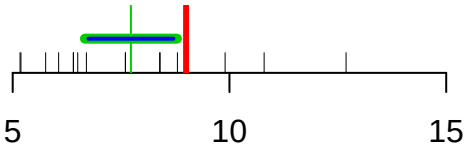


Abbildung 4.1: Daten mit Mittelwert, mit Konfidenzintervallen. (Siehe R-Code 4.1.)

Die Idee von statistischen Testverfahren ist inhaltliche Hypothesen statistisch zu formulieren und basierend auf den Daten Schlüsse daraus zu ziehen. Inhaltliche Hypothesen sind zum Beispiel zweiseitige (ungerichtete, *two-sided test*) Unterschiedshypothesen ("Hb SS und Hb Sβ haben einen unterschiedlichen mittleren Hämoglobingehalt") oder einseitige (gerichtete, *one-sided test*) Unterschiedshypothesen ("Hb SS hat einen kleineren mittleren Hämoglobingehalt als Hb Sβ"). Weitere Beispiele inhaltlicher Hypothesen sind in Rudolf and Kuhlisch (2008) gegeben.

Vereinfacht ausgedrückt, berechnet ein statistischer Test aus den Daten einen "Wert" (der von der Hypothese abhängt), und vergleicht diesen mit der hypothetischen Dichte. Wenn der Wert eine kleine "Wahrscheinlichkeit" hat aufzutreten, argumentieren wir, dass

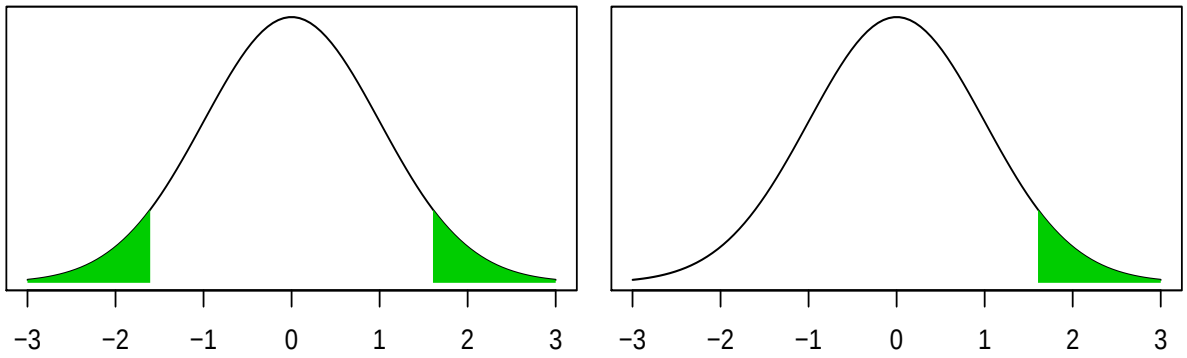


Abbildung 4.2: p-Wert für $H_0 : \mu = \mu_0 = 0$ (links) und $H_0 : \mu \leq \mu_0 = 0$ (rechts) für $t_{\text{Vers}} = 1.6$.

die Hypothese eher unwahrscheinlich ist und verwerfen diese. Eine ähnliche Argumentation gilt auch für diskrete Verteilungen.

Die Hypothese wird mit Nullhypothese H_0 bezeichnet und die hypothetische Verteilung als Nullverteilung. Im Zusammenhang mit statistischen Tests wird der oben erwähnte Wert als Wert der Teststatistik genannt und die Wahrscheinlichkeit als p-Wert. Die Teststatistik wird durch die Testsituation und die statistischen Voraussetzungen gegeben und wird in den kommenden Abschnitten diskutiert.

Eine formellere Definition des p-Werts lautet:

Definition 4.1. Der p-Wert (*p-value*) ist die Wahrscheinlichkeit, unter der Nullverteilung den beobachteten Wert der Teststatistik oder einen noch extremeren Wert zu beobachten. ♢

Die Entscheidung "klein" oder nicht, basiert auf dem sogenannten Signifianzniveau, α. Das Signifikanzniveau definiert (induziert) durch die Nullverteilung bei einem zweiseitigen Test Quantile, $Q(\alpha/2)$ und $Q(1-\alpha/2)$, welche eine besondere Rolle einnehmen:

Definition 4.2. Der Ablehnungsbereich eines Tests (*rejection region*) sind alle Werte der Teststatistik mit einen kleineren p-Wert als das Signifianzniveau. Die Grenze des Ablehnungsbereichs ist der kritische Wert. ♢

Für gerichtete, d.h. einseitige Hypothesen wie zum Beispiel $H_0 : \mu \leq \mu_0$ oder $H_0 : \mu \geq \mu_0$, wird trotzdem der "="-Fall in der Nullhypothese verwendet.

Testverfahren induzieren zwei verschiedene Fehler: α-Fehler (*type I error*) und β-Fehler (*type II error*) (siehe Tabelle 4.1). Der α-Fehler:

i) wird im Voraus durch die Wahl des Signifikanzniveaus festgelegt (oft 5%, 1%)

ii) wird nicht durch den Stichprobenumfang beeinflusst

iii) wird durch mehrfaches Testen derselben Daten erhöht

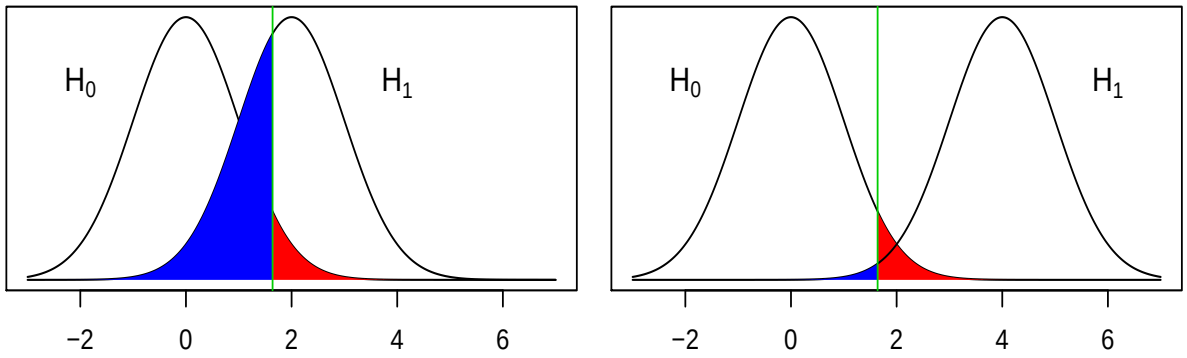


Abbildung 4.3: α-Fehler (rot) und β-Fehler (blau) für zwei verschiedene Alternativen ($\mu = 2$ links, $\mu = 4$ rechts) mit $H_0 : \mu \leq \mu_0 = 0$.

Tabelle 4.1: Fehler eines Tests: $\boldsymbol{\alpha}$-Fehler/Fehler 1. Art und $\boldsymbol{\beta}$-Fehler/Fehler 2. Art.

		Wahrer Zustand	
		H_0 wahr	H_1 wahr
Testresultat	H_0 nicht verwerfen	$1 - \alpha$	$\boldsymbol{\beta}$
	H_0 verwerfen	$\boldsymbol{\alpha}$	$1 - \beta$

und der β-Fehler:

iv) hängt vom Stichprobenumfang und von α ab

v) kann nur als Funktion der Alternative bestimmt werden.

Der α- und β-Fehler ist in der Abbildung 4.3 für zwei Alternativen eingezeichnet. Der Wert $1 - \beta$ wird auch als Macht oder Power des Tests bezeichnet (*power*). Eine grosse Macht ist wünschenswert. R-Code 4.2 berechnet die Macht unter einer Gauss'schen Annahme.

Der Ablauf eines statistischen Tests kann mit folgendem Ablauf beschrieben werden:

i) Inhaltliche Hypothese oder Fragestellung

ii) Statistisches Modell (Voraussetzungen)

iii) Statistische Hypothese

iv) Wahl des Signifikanzniveaus

v) Wahl des Tests

vi) Berechnung des Teststatistikwerts und/oder des p-Werts

vii) Entscheidung

viii) Interpretation

R-Code 4.2 Powerkurve einseitig und zweiseitig für einen z-Test. (Siehe Abbildung 4.4.)

```
alpha <- 0.05                       # significance level
mu0 <- 0                             #  H_0  mean
mu1 <- seq(-1, to=4, by=.5)          #  H_1  mean
power1 <- 1-pnorm( qnorm(1-alpha, mean=mu0), mean=mu1)
power2 <- pnorm( qnorm(alpha/2, mean=mu0), mean=mu1)+
    pnorm( qnorm(1-alpha/2, mean=mu0), mean=mu1, lower.tail=FALSE)
plot( mu1, power1, type='l', ylim=c(0,1), xlab=expression(mu[1]-mu[0]),
     ylab="Power", col=4)
lines( mu1, power2,lty=2)
abline( h=alpha, col='gray')        # Signifikanz
abline( v=c(2,4), lty=4, col=3)     # Werte aus Abbildung 4.3
```

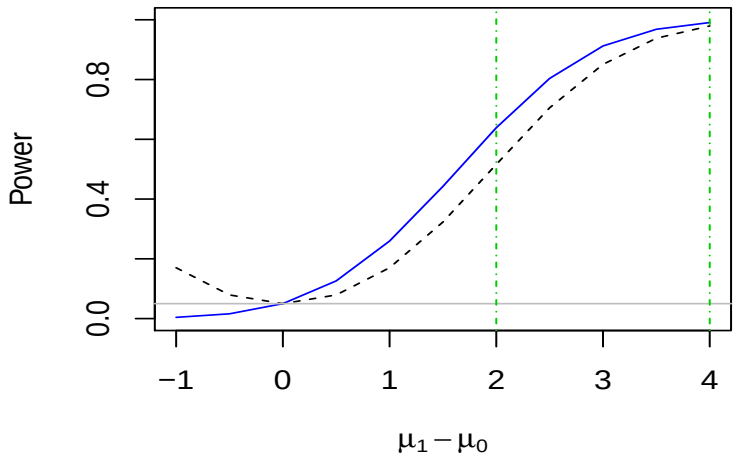


Abbildung 4.4: Power einseitig (blau durchgezogen) und zweiseitig (schwarz gestrichelt). Die vertikalen Linen entsprechen den Alternativen $\mu = 2$ und $\mu = 4$ der Abbildung 4.3. (Siehe R-Code 4.2.)

Je nach Voraussetzungen ist die Wahl des Tests eingeschränkt. Das Signifikanzniveau muss jedoch immer vor der Berechnung gewählt werden.

Im Punkt vi), wenn der Teststatistikwert t_{Vers} berechnet wird, wird dieser in vii) mit dem tabellarisierten kritischen Wert t_{Tab} verglichen. Die Berechnung des p-Werts beinhaltet einen Vergleich mit α. Der p-Wert ist oft schwieriger zu berechnen, hat aber als Mehrwert auch eine direkte Interpretation, wie stark (oder schwach) die Evidenz gegen die Nullhypothese ist.

Wir betrachten hier verschiedene Situationen, in denen wir testen. Die Wahl des Tests hängt in erster Linie vom Parameter ab, in zweiter Linie von den statistischen Voraussetzungen ab. Die folgende Auflistung kann auch als Entscheidungsbaum verwendet werden.

Allgemeine Bemerkungen zu statistischen Tests

In der Folge werden wir einige wichtige statistische Tests in Textblöcken wie diesem darstellen.
Allgemein bezeichnen wir mit

- n, n_x, n_y, ... den Stichprobenumfang
- $\overline{x}$, $\overline{y}$, ... das arithmetische Mittel
- s^2, s_x^2, s_y^2, ... die geschätzte Varianz, zum Beispiel

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n_x}(x_i - \overline{x})^2.$$

  In den betrachteten Tests ist die Varianz unbekannt;

- α das Signifikanzniveau

Die inhaltliche Hypothese kann nur allgemein formuliert werden und wird daher der einfachheitshalber *Fragestellung* genannt. Generell werden zweiseitige Tests ausgeführt. Bei einseitigen Tests kann das Signifikanzniveau entsprechend angepasst werden.
Bei Unklarheiten wird die *statistische Hypothese* spezifiziert.
Für die meisten Tests gibt es eine entsprechende R Funktion. Die Argumente `x`, `y` bezeichnen meist die Vektoren mit den Realisationen und `alpha` das Signifikanzniveau.

- Mittelwerte
  - Eine Stichprobe (Test 1)
  - Zwei Stichproben
    - Zwei unabhängige Stichproben (Test 2 und Test 7 in Kapitel 6)
    - Zwei gepaarte Stichproben (Test 3 und Test 8 in Kapitel 6)
  - Mehrere Stichproben (Test 13 in Kapitel 11)
- Varianzen (Test 4)
- Proportionen (Test 6 in Kapitel 5)
- Verteilungen (Test 5)
- ...

Verteilungstest (oder auch Anpassungstests) unterscheiden sich von den restlichen, weil diese nicht nur einen einzigen Parameter testen. In den folgenden Abschnitten betrachten wir einzelne Test im Detail. Die Tests werden immer mit dem selben Schema präsentiert (allgemeine Bemerkungen dazu sind im gelben Textblock gegeben).

4.2 Vergleich von Mittelwerten

In diesem Abschnitt vergleichen wir Mittelwerte aus normalverteilten Stichproben. Die Teststatistiken sind t-verteilt (siehe Abschnitt 2.7.3) und so wird zur Berechnung des p-Werts die Funktion `pt` verwendet.

Wir präsentieren nun die Tests mit je einem Beispiel basierend auf den Essdaten und einem entsprechenden R-Code.

Test 1: Vergleich eines Mittelwertes mit einem theoretischen Wert

Fragestellung: Weicht der experimentell gefundene Mittelwert $\overline{x}$ der Stichprobe signifikant vom unbekannten ("theoretischen") Mittelwert μ_T ab?

Voraussetzungen: Die Grundgesamtheit, aus der die Stichprobe stammt, ist normalverteilt mit dem unbekannten Mittelwert μ_T. Die gemessenen Daten sind intervallskaliert und Varianz unbekannt.

Berechnung: $t_{\text{Vers}} = \dfrac{|\overline{x} - \mu_T|}{s} \cdot \sqrt{n}\,.$

Entscheidung: Vergleiche t_{Vers} und $t_{\text{Tab}}(n-1;\alpha/2)$: verwerfe $H_0 : \mu = \mu_T$, wenn $t_{\text{Vers}} > t_{\text{Tab}}$.

Berechnung in R: `t.test( x, mu=muT, conf.level=1-alpha)`

Beispiel 4.2. Wir testen die Hypothese, ob die normale Essmenge signifikant weniger als 9 ist (hier: eine Stichprobe, weicht der Mittelwert von einem unbekannten Wert ab). Folgende Werte sind gegeben: Mittelwert: 7.73, Standardabweichung: 2.06, Stichprobenumfang: 17.

$H_0 : \mu = 9$

$t_{\text{Vers}} = \dfrac{|7.73 - 9|}{2.06/\sqrt{17}} = 2.54$

$t_{\text{Tab}}(16; 0.05/2) = 2.12$ $\qquad$ p-Wert: 0.022.

Siehe R-Code 4.3 und Test 1. ♣

R-Code 4.3 Eine Stichprobe (siehe Beispiel 4.2 und Test 1)

```
t.test( Essmenge.normal, mu=9)
##
##  One Sample t-test
##
## data:  Essmenge.normal
## t = -2.5422, df = 16, p-value = 0.02174
## alternative hypothesis: true mean is not equal to 9
## 95 percent confidence interval:
##  6.665564 8.788553
## sample estimates:
## mean of x
##  7.727059
```

Test 2: Vergleich zweier Mittelwerte unabhängiger Stichproben

Fragestellung: Sind die Mittelwerte $\overline{x}$ und $\overline{y}$ zweier Stichproben X und Y signifikant verschieden?

Voraussetzungen: Beide Grundgesamtheiten sind normalverteilt mit gleichen, unbekannten Varianzen. Die Stichproben sind unabhängig.

Berechnung: $t_{\text{Vers}} = \dfrac{|\overline{x} - \overline{y}|}{s_d} \cdot \sqrt{\dfrac{n_x \cdot n_y}{n_x + n_y}}$,

wobei $s_d = \sqrt{\dfrac{1}{n_x + n_y - 2} \cdot \left[(n_x - 1)s_x^2 + (n_y - 1)s_y^2\right]}$.

Entscheidung: Vergleiche t_{Vers} und $t_{\text{Tab}}(n_x + n_y - 2; \alpha/2)$: werfe $H_0 : \mu_x = \mu_y$ wenn $t_{\text{Vers}} > t_{\text{Tab}}$.

Berechnung in R: `t.test( x, y, var=TRUE, conf.level=1-alpha)`

Beispiel 4.3. Wir vergleichen Mittelwerte von zwei unabhängigen Stichproben (siehe R-Code 4.4 und Test 2). Gegeben sind folgende Werte: Mittelwerte: 7.73 und 8.21, Standardabweichungen: 2.06 und 1.9 und Stichprobenumfänge: 17 und 17:

$H_0 : \mu_X = \mu_Y$

$$t_{\text{Vers}} = \frac{|7.73 - 8.21|}{\sqrt{(16 \cdot 2.06^2 + 16 \cdot 1.90^2)/(17 + 17 - 2)}} \sqrt{\frac{n}{2}} = 0.71$$

$t_{\text{Tab}}(32; 0.05/2) = 2.12$ $\quad$ p-Wert: 0.48. ♣

R-Code 4.4 Zweistichproben-t-Test mit unabhängigen Stichproben (siehe Beispiel 4.3 und Test 2).

```
t.test( Essmenge.normal, Essmenge.waehrend, var=TRUE)
##
##  Two Sample t-test
##
## data:  Essmenge.normal and Essmenge.waehrend
## t = -0.70939, df = 32, p-value = 0.4832
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.869660  0.903778
## sample estimates:
## mean of x mean of y
##  7.727059  8.210000
```

Test 3: Vergleich zweier Mittelwerte verbundener Stichproben

Fragestellung: Sind die Mittelwerte $\overline{x}$ und $\overline{y}$ zweier verbundener Stichproben X und Y signifikant verschieden?

Voraussetzungen: Die Stichproben sind verbunden, die Messwerte intervallskaliert. Die Differenzen sind normal verteilt mit dem unbekannten Mittelwert δ. Varianz unbekannt.

Berechnung: $t_{\text{Vers}} = \dfrac{|\overline{d}|}{s_d} \cdot \sqrt{n}$, wobei

- $d_i = x_i - y_i$ die i-te Messwert-Distanz,
- $\overline{d}$ das arithmetische Mittel und s_d die Standardabweichung der Differenzen d_i.

Entscheidung: Vergleiche t_{Vers} und $t_{\text{Tab}}(n-1;\alpha/2)$: verwerfe $H_0 : \delta = 0$ wenn $t_{\text{Vers}} > t_{\text{Tab}}$.

Berechnung in R: `t.test(x-y, conf.level=1-alpha)` oder

`t.test(x, y, paired=TRUE, conf.level=1-alpha)`

Natürlich kann der gepaarte Zweistichproben-t-Test auch als einfacher t-Test der Differenzen mit Mittelwert $\mu_T = 0$ betrachtet werden.

Beispiel 4.4. Wir vergleichen die Mittelwerte zweier gepaarten Stichproben (siehe R-Code 4.5 und Test 3). Gegeben sind folgende Werte: Mittelwert der Differenzen: -0.48, Standardabweichungen: 0.91 und Stichprobenumfang: 17:

$H_0 : d = 0$ oder $H_0 : \mu_X = \mu_Y$

$$t_{\text{Vers}} = \frac{|-0.48|}{0.91/\sqrt{17}} = 2.17$$

$t_{\text{Tab}}(16; 0.05) = 2.12$ p-Wert: 0.045.

Mit einem Signifikanzniveau von $\alpha = 5\%$ würde man die Hypothese verwerfen. Die Evidenz dazu ist aber sehr schwach und der p-Wert muss hier auch kommuniziert werden. ♣

R-Code 4.5 Zwei Stichprobentests gepaarte Stichproben (siehe Beispiel 4.4 und Test 3).

```
t.test( Essmenge.normal, Essmenge.waehrend, paired=TRUE)
##
##  Paired t-test
##
## data:  Essmenge.normal and Essmenge.waehrend
## t = -2.1862, df = 16, p-value = 0.04401
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.95124273 -0.01463962
## sample estimates:
## mean of the differences
##               -0.4829412
# Gleiches Resultat mit
# t.test( Essmenge.normal - Essmenge.waehrend)
```

Die "t-Tests" setzen normalverteilte Daten voraus. Die Tests sind aber relativ robust, solange nicht extreme Ausreisser vorhanden sind. Ansonsten können auch Rang-basierte Tests gebraucht werden (siehe Kapitel 6):

- U-Test (Wilcoxon-Mann-Whitney-Test),
- Wilcoxon-Test.

Natürlich kann man diese Voraussetzung quantitativ mit *Normalitäts-Tests* überprüfen (χ^2-, Shapiro-, Kolmogorov–Smirnov-Test). Oft reicht jedoch eine qualitative Überprüfung wie zum Beispield mit Hilfe von Q-Q-Plots aus.

4.3 Dualität Test und Konfidenzintervalle

Verwerfen von $H_0 : \theta = \theta_0$ mit einem Signifikanzniveau α ist equivalent zu θ_0 liegt nicht im $(1-\alpha)$-Konfidenzintervall von θ. Die Dualität von Test und Konfidenzintervall wird schnell deutlich, wenn man das Kriterium zur Testentscheidung umformuliert, hier am Beispiel "Vergleich eines Mittelwerts mit einem theoretischen Wert".

Wie oben gezeigt, wird H_0 verworfen, wenn $t_{\text{Vers}} > t_{\text{Tab}}$, also wenn

$$t_{\text{Vers}} = \frac{|\overline{x} - \mu_T|}{s} \cdot \sqrt{n} > t_{\text{Tab}}. \tag{4.1}$$

Im umgekehrten Fall wird H_0 beibehalten, falls

$$t_{\text{Vers}} = \frac{|\overline{x} - \mu_T|}{s} \cdot \sqrt{n} \leq t_{\text{Tab}}. \tag{4.2}$$

Dies kann man nun umschreiben zu

$$|\overline{x} - \mu_T| \leq t_{\text{Tab}} \frac{s}{\sqrt{n}}, \tag{4.3}$$

was also bedeutet, dass man H_0 nicht verwirft, wenn entweder

$$\overline{x} - \mu_T \geq -t_{\text{Tab}} \frac{s}{\sqrt{n}} \tag{4.4}$$

oder

$$\overline{x} - \mu_T \leq t_{\text{Tab}} \frac{s}{\sqrt{n}}. \tag{4.5}$$

Dies kann man wiederum umschreiben zu

$$\mu_T \leq \overline{x} + t_{\text{Tab}} \frac{s}{\sqrt{n}} \tag{4.6}$$

und

$$\mu_T \geq \overline{x} - t_{\text{Tab}} \frac{s}{\sqrt{n}}, \tag{4.7}$$

was den Grenzen des $(1-\alpha)$-Konfidenzintervalls für μ_T entspricht. Analog kann man diese Dualität auch bei allen anderen hier gezeigten Tests herleiten.

Beispiel 4.5. Wir betrachten die Situation aus Beispiel 4.2. Statt des p-Werts kann man auch das Konfidenzintervall betrachten, dessen Grenzen 6.67 und 8.79. Da die 9 nicht in diesem Bereich liegt, wird die Nullhypothese abgelehnt.

Grafisch ist dies in Abbildung 4.1 dargestellt. ♣

In R liefern die allermeisten Testfunktionen gleichzeitig die entsprechenden Konfidenzintervalle mit. Bei einigen muss zusätzlich das Argument `conf.int=TRUE` gesetzt werden.

4.4 Multiples Testen

In vielen Fällen will man nicht nur einen einzigen, sondern eine ganze Reihe von Tests durchführen. Hier muss man allerdings beachten, dass das Signifikanzniveau α nur für einen einzigen Test gilt. Das heisst, bei der Durchführung eines einzigen Tests ist die Wahrscheinlichkeit für ein falsch signifikantes Testresultat gleich dem Signifikanzniveau, zum Beispiel $\alpha = 0.05$. Die Wahrscheinlichkeit, dass bei diesem Test die Nullhypothese H_0 korrekterweise nicht abgelehnt wird, ist dann $1 - 0.05 = 0.95$.

Betrachten wir nun aber die Situation, dass $m > 1$ Tests durchgeführt werden. Die Wahrscheinlichkeit, dass mindestens 1 falsch signifikantes Testresultat beobachtet wird, ist dann gleich 1− der Wahrscheinlichkeit, dass kein falsch signifikantes Resultat herauskommt. Es gilt:

$$\begin{aligned} \text{P(mind. 1 falsch signif. Resultat)} &= 1 - \text{P(kein falsch signif. Resultat)} && (4.8)\\ &= 1 - (1-\alpha)^m && (4.9)\\ &= 1 - 0.95^m. && (4.10) \end{aligned}$$

In Tabelle 4.2 sind die Wahrscheinlichkeiten für mindestens ein falsch signifikantes Resultat bei unterschiedlichen m und $\alpha = 0.05$ aufgeführt. Schon bei wenigen Tests erhöht sich diese drastisch, was nicht toleriert werden sollte.

Tabelle 4.2: Wahrscheinlichkeiten für mindestens ein falsch signifikantes Testresultat bei Durchführung von m Tests

m	P(mind. 1 falsch signif. Resultat)
1	0.050
2	0.098
3	0.143
4	0.185
5	0.226
6	0.265
7	0.302
8	0.337
9	0.370
10	0.401

Es gibt eine Reihe von Verfahren, die es erlauben, mehrere Tests durchzuführen und dabei das vorgegebene Signifikanzniveau einzuhalten. Das einfachste und bekannteste ist die so genannte Bonferroni-Korrektur. Dabei verwendet man für jeden einzelnen Test ein neues Signifikanzniveau $\alpha_{neu} = \frac{\alpha}{m}$. Allerdings finden sich eine Reihe von alternativen Verfahren, die je nach Situation besser geeignet sein können, siehe hierzu zum Beispiel Farcomeni (2008).

4.5 Weitere Tests

Wie wir Mittelwerte vergleichen können, gibt es auch Tests, die zwei Varianzen vergleichen. Der "klassische" F-Test ist im Test 4 gegeben.

Die Quantil-, Dichte- und Verteilungsfunktion der Teststatistik des F-Test ist in R mit `[q,d,p]f` implementiert (siehe auch Abschnitt 2.7.4).

Test 4: Vergleich zweier Varianzen

Fragestellung: Sind die Varianzen s_x^2 und s_y^2 der beiden Stichproben X und Y signifikant verschieden?

Voraussetzungen: Beide Grundgesamtheiten, aus denen die Stichproben entnommen wurden, sind normalverteilt. Die Stichproben sind unabhängig, die Messwerte intervallskaliert.

Berechnung: $F_{\text{Vers}} = \dfrac{s_x^2}{s_y^2}$ (die grössere Varianz steht im Zähler, also $s_x^2 > s_y^2$).

Entscheidung: Vergleiche F_{Vers} mit $F_{\text{Tab}}(n_x - 1, n_y - 1; \alpha)$: verwerfe H_0: $\sigma_x^2 = \sigma_y^2$ wenn $F_{\text{Vers}} > F_{\text{Tab}}$.

Berechnung in R: `var.test( x, y, conf.level=1-alpha)`

Beispiel 4.6. Der Zweistichproben-t-Test (Test 2) setzt gleiche Varianzen voraus. Die Daten `Essmenge` widersprechen der Nullhypothese nicht, wie R-Code 4.6 zeigt (vergleiche Beispiel 4.3 und R-Code 4.4). ♣

Der sogenannte Chi-Quadrat-Test (χ^2-Test, *Chi-squared test*) prüft, ob die vorliegenden Daten auf eine bestimmte Weise verteilt sind. Diese Test basiert auf einem Vergleich von beobachteten mit erwarteten Häufigkeiten und wird daher auch noch in anderen Zusammenhängen gebraucht, zum Beispiel, um zu testen, ob zwei Merkmale (stochastisch) unabhängig sind (Test 6).

Unter der Nullhypothese ist der Chi-Quadrat-Test χ^2-verteilt und die Quantil-, Dichte- und Verteilungsfunktion ist in R mit `[q,d,p]chisq` implementiert (siehe Abschnitt 2.7.2).

In Test 5 sollten die Merkmalsklassen so zusammengefasst werden, dass alle $E_1 \geq 1$ sind. Ausserdem muss $N - k > 1$ sein.

Beispiel 4.7. Bei nur 17 Beobachtungen ist es oft nutzlos, die Normalität der Daten zu überprüfen. Selbst bei grösseren Stichproben ist ein Q-Q-Plot aufschlussreicher. Der Vollständigkeit halber wird die Vorgehensweise im R-Code 4.7 mit Essmenge normal und während (Menstruation) aufgezeigt. ♣

R-Code 4.6 Vergleich von Varianzen zweier Stichproben (siehe Beispiel 4.6 und Test 4).

```
var.test( Essmenge.normal, Essmenge.waehrend)
##
##  F test to compare two variances
##
## data:  Essmenge.normal and Essmenge.waehrend
## F = 1.1785, num df = 16, denom df = 16, p-value = 0.7465
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4267976 3.2543763
## sample estimates:
## ratio of variances
##           1.178541
```

Test 5: Vergleich von beobachteten mit erwarteten Häufigkeiten

Fragestellung: Weichen die beobachteten Häufigkeiten B_i einer Stichprobe signifikant von erwarteten Häufigkeiten E_i einer Verteilung ab?

Voraussetzungen: Stichprobe mit mindestens nominalskalierten Daten.

Berechnung: Berechne zu den beobachteten Werten B_i die erwarteten absoluten Häufigkeiten E_i mit Hilfe der erwarteten Verteilung und bilde dann

$$\mathcal{X}^2_{\text{Vers}} = \sum_{i=1}^{n} \frac{(B_i - E_i)^2}{E_i}$$

wobei N die Anzahl der Merkmalsklassen.

Entscheidung: Vergleiche $\mathcal{X}^2_{\text{Vers}}$ und $\mathcal{X}^2_{\text{Tab}}(N-1-k;\alpha)$ mit k die Anzahl aus den Daten geschätzter Parameter: verwerfe H_0: "keine Abweichung zwischen Beobachtung und Erwartung" wenn $\mathcal{X}^2_{\text{Vers}} > \mathcal{X}^2_{\text{Tab}}$,.

Berechnung in R: `chisq.test( x, p=E/sum(x))` oder `chisq.test( x, p=E, rescale.p=TRUE)`

R-Code 4.7 Normalität testen (siehe Beispiel 4.7 und Test 5).

```
Essmenge.normal.waehrend <- c(Essmenge.normal, Essmenge.waehrend)
observed <- hist( Essmenge.normal.waehrend, plot=FALSE, breaks=4)
       # ohne 'breaks' Argument gibts zu viele Klassen
observed[1:2]
## $breaks
## [1]  4  6  8 10 12 14
##
## $counts
## [1]  4 13 13  2  2
m <- mean( Essmenge.normal.waehrend)
s <- sd( Essmenge.normal.waehrend)
p <- pnorm( observed$breaks, mean=m, sd=s)
chisq.test( observed$counts, p=diff( p), rescale.p=TRUE)
## Warning in chisq.test(observed$counts, p = diff(p), rescale.p =
TRUE): Chi-squared approximation may be incorrect
##
##  Chi-squared test for given probabilities
##
## data:  observed$counts
## X-squared = 4.3613, df = 4, p-value = 0.3593
# Freiheitsgrade sollten ajustiert werden da eigendlich 2 und nicht 1
```

Kapitel 5

Vertiefung: Anteile

Präeklampsie ist eine hypertensive Erkrankung in der Schwangerschaft (Schwangerschaftshypertonie) mit Leitsymptomen (Ödeme, Bluthochdruck und Proteinurie). In einer doppelblinden (*double-blinded*) randomisierten kontrollierten Studie (RCT, *randomized controlled trial*) wurden 2706 Schwangere mit einem Diuretikum oder mit einem Scheinpräprat behandelt Landesman *et al.* (1965). In 138 von 1370 Fällen in der Behandlungsgruppe wurde Präeklampsie diagnostiziert, in der Kontrollgruppe jedoch in 175 Fällen. Die medizinische Frage ist, ob die Diuretika-Medikamente, welche Wasser ausschwemmen, das Risiko einer Präeklampsie verringern.

In diesem Kapitel werden wir Anteile schätzen und Anteile miteinander vergleichen.

5.1 Schätzen

Wir betrachten $X \sim \mathcal{B}in(n, p)$. Intuitiv ist x/n ein Schätzwert von p und dementsprechend X/n eine Schätzfunktion.

Mit der Momentenmethode erhalten wir die Schätzfunktion $\widehat{p} = X/n$, da $np = \mathrm{E}(X)$ und wir nur eine Beobachtung (Anzahl Fälle) haben.

Der Likelihoodschätzer ist wie folgt:

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{5.1}$$

$$\ell(p) = \log\big(L(p)\big) = \log\binom{n}{x} + x\log(p) + (n-x)\log(1-p) \tag{5.2}$$

$$\frac{d\ell(p)}{dp} = \frac{x}{p} - \frac{n-x}{1-p} \tag{5.3}$$

Und $\widehat{p}_{\mathrm{ML}}$ erfüllt $\dfrac{x}{\widehat{p}_{\mathrm{ML}}} = \dfrac{n-x}{1-\widehat{p}_{\mathrm{ML}}}$ und daher $\widehat{p}_{\mathrm{ML}} = x/n$

In unserem Beispiel: $\widehat{p}_{\mathrm{B}} = 138/1370 \approx 10\%$ und $\widehat{p}_{\mathrm{K}} = 175/1336 \approx 13\%$. Schon hier stellt sich die Frage, ob die beiden Anteile verschieden "genug" sind, um von einer Wirkung des Medikamentes sprechen zu können.

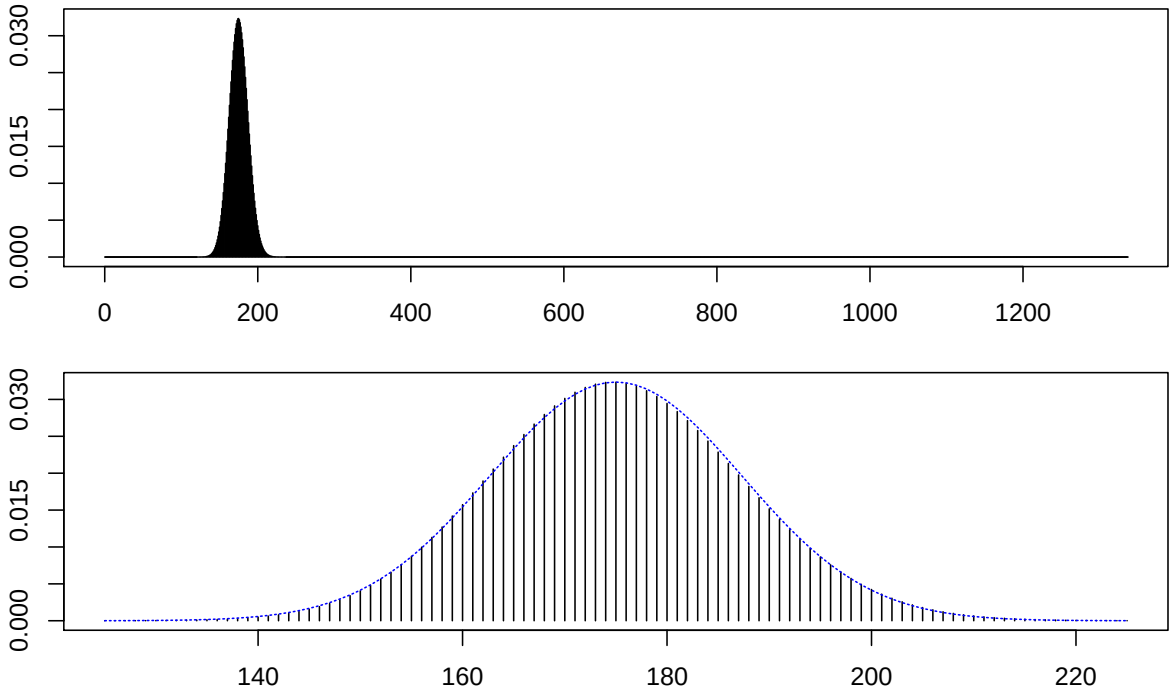


Abbildung 5.1: Wahrscheinlichkeitsfunktion mit Normalapproximation.

Fragen, die man jetzt beantworten kann, sind z.B.: Wieviele Falle von Präeklampsie muss man unter 100 Schwangerschaften erwarten? Wie hoch ist die Wahrscheinlichkeit, dass in mehr als 20 (von 100) Fällen Präeklampsie auftritt?

Bei Anteilen wird oft von Chancen (*odds*) gesprochen, definiert als $\omega = p/(1-p)$. Die entsprechende intutive Schätzfunktion (und Schätzwert) ist $\widehat{\omega} = \widehat{p}/(1-\widehat{p})$.

5.2 Konfidenzintervalle

Wir brauchen die Normalapproximation, um Konfidenzintervalle zu konstruieren:

$$1-\alpha \approx \mathrm{P}\Big(z_{\alpha/2} \leq \frac{X-np}{\sqrt{np(1-p)}} \leq z_{1-\alpha/2}\Big). \tag{5.4}$$

Durch Umformung erhält man

$$1-\alpha \approx \mathrm{P}\Big(z_{\alpha/2}\sqrt{np(1-p)} \leq X-np \leq z_{1-\alpha/2}\sqrt{np(1-p)}\Big) \tag{5.5}$$

$$= \mathrm{P}\Big(-\frac{X}{n} + z_{\alpha/2}\frac{1}{n}\sqrt{np(1-p)} \leq -p \leq -\frac{X}{n} + z_{1-\alpha/2}\frac{1}{n}\sqrt{np(1-p)}\Big) \tag{5.6}$$

Als eine weitere Annäherung brauchen wir

$$1-\alpha \approx \mathrm{P}\Big(-\frac{X}{n} + z_{\alpha/2}\frac{1}{n}\sqrt{n\widehat{p}(1-\widehat{p})} \leq -p \leq -\frac{X}{n} + z_{1-\alpha/2}\frac{1}{n}\sqrt{n\widehat{p}(1-\widehat{p})}\Big). \tag{5.7}$$

Da $\widehat{p} = x/n$ und $q := z_{1-\alpha/2} = -z_{\alpha/2}$ ist

$$b_{o,u} = \widehat{p} \pm q \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \widehat{p} \pm q \cdot \mathrm{SE}(\widehat{p}). \tag{5.8}$$

Bemerkung 5.1. Likelihoodtheorie besagt, dass für (reguläre) Modelle mit Parameter θ, wenn $n \to \infty$, $\widehat{\theta}_{\text{ML}}$ normalverteilt ist mit Erwartungswert θ und Varianz $\text{Var}(\widehat{\theta}_{\text{ML}})$.

Da $\text{Var}(X/n) = p(1-p)/n$, darf man (intuitiv) annehmen, dass $\text{SE}(\widehat{p}) = \sqrt{\widehat{p}(1-\widehat{p})/n}$. Das sogenannte Wald-Konfidenzintervall basiert auf dieser Annahme (welche natürlich formell gezeigt werden kann) und ist mit (5.8) identisch. □

Wird die Ungleichung in (5.4) durch Lösung einer quadratischen Gleichtung umgeformt, erhalten wir das Wilson-Konfidenzintervall

$$b_{o,u} = \frac{1}{1+q^2/n} \cdot \left(\widehat{p} + \frac{q^2}{2n} \pm q \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{q^2}{4n^2}} \right) \tag{5.9}$$

Konfidenzintervall eines Anteils

Ein approximatives $(1-\alpha)$-Wald-Konfidenzintervall für einen Anteil ist

$$b_{o,u} = \widehat{p} \pm q \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \tag{5.10}$$

Ein approximatives $(1-\alpha)$-Wilson-Konfidenzintervall für einen Anteil ist

$$b_{o,u} = \frac{1}{1+q^2/n} \cdot \left(\widehat{p} + \frac{q^2}{2n} \pm q \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{q^2}{4n^2}} \right) \tag{5.11}$$

In R wird ein exaktes $(1-\alpha)$-Konfidenzintervall für einen Anteil mit `binom.test( x, n)$conf.int` berechnet.

Das Wilson-Konfidenzintervall ist "komplizierter" als das Wald-Konfidenzintervall. Ist es auch "besser", weil eine Annäherung weniger gebraucht wurde?

Idealerweise sollte die Überdeckungswahrscheinlichkeit (*coverage*) eines $(1-\alpha)$-Konfidenzintervall $1-\alpha$ sein. Für diskrete Zufallsvariablen ist die Überdeckungswahrscheinlichkeit

$$\text{P}(p \in \text{KI}) = \sum_{x=0}^{n} \text{P}(X = x) I_{\{p \in \text{KI}\}}. \tag{5.12}$$

Der R-Code 5.1 berechnet die Überdeckungswahrscheinlichkeit der 95%-Konfidenzintervalle für $X \sim \mathcal{Bin}(n = 50, p = 0.4)$ und zeigt, dass das Wilson-Konfidenzintervall eine bessere nominale Dekung aufweisst (96% im Vergleich zu 94%).

Die Überdeckungswahrscheinlichkeit hängt von p ab, in Abbildung 5.2 (mit R-Code 5.2) gezeigt. Im Mittel hat das Wilson-Konfidenzintervall eine bessere nominale Dekung. Diese Beobachtung gilt auch, wenn n variert wir,d, wie in Abbildung 5.3 gezeigt wird.

R-Code 5.1 Überdeckungswahrscheinlichkeit der 95%-Konfidenzintervalle für $X \sim \mathcal{B}in(n = 50, p = 0.4)$.

```
p <- .4
n <- 20
x <- 0:n
   #
WaldKI <- function(x, n){
  mid <- x/n
  se <- sqrt(x*(n-x)/n^3)
  cbind( pmax(0, mid - 1.96*se),  pmin(1, mid + 1.96*se))
}
WaldKIs <- WaldKI(x,n)
Waldind <- (WaldKIs[,1] < p) & (WaldKIs[,2] > p)
Waldcoverage <- sum( dbinom(x, n, p)*Waldind)    # eq (5.12)
   #
WilsonKI <- function(x, n){
  mid <- (x + 1.96^2/2)/(n + 1.96^2)
  se <- sqrt(n)/(n+1.96^2)*sqrt(x/n*(1-x/n)+1.96^2/(4*n))
  cbind( pmax(0, mid - 1.96*se),  pmin(1, mid + 1.96*se))
}
WilsonKIs <- WilsonKI(x,n)
Wilsonind <- (WilsonKIs[,1] < p) & (WilsonKIs[,2] > p)
Wilsoncoverage <- sum( dbinom(x, n, p)*Wilsonind)
   #
print( c(true=0.95, Wald=Waldcoverage, Wilson=Wilsoncoverage))
##      true      Wald    Wilson
## 0.9500000 0.9280191 0.9630099
```

Die empirische Breite eines Konfidenzintervalls ist $b_o - b_u$. Für das Wald-Konfidenzintervall

$$2q \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \tag{5.13}$$

und für das Wilson-Konfidenzintervall

$$\frac{2q}{1+q^2/n}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{q^2}{4n^2}}. \tag{5.14}$$

Die Breiten sind in Abbildung 5.4 dargestellt. Ausser für $5 < x < 36$ hat das Wilson-Konfidenzintervall eine kleinere Breite.

R-Code 5.2 Hinweise zur Konstruktion der Abbildung 5.2.

```
pi <- seq(0.001, .999, .001)
  #
  # either a loop over all elements in p or a few applies
  # over the functions 'Wilsonind' and 'Wilsoncoverage'
  #
  # Wilsoncoverage is thus a vector!
Waldsmooth <- loess(Waldcoverage ~ p, span=.1)
Wilsonsmooth <- loess(Wilsoncoverage ~ p, span=.1)
plot(p, Waldcoverage, type='l', ylim=c(.8,1))
lines(c(-1, 2), c(.95, .95), col=2, lty=2)
lines(Waldsmooth$x, Waldsmooth$fitted, col=3, lw=2)
```

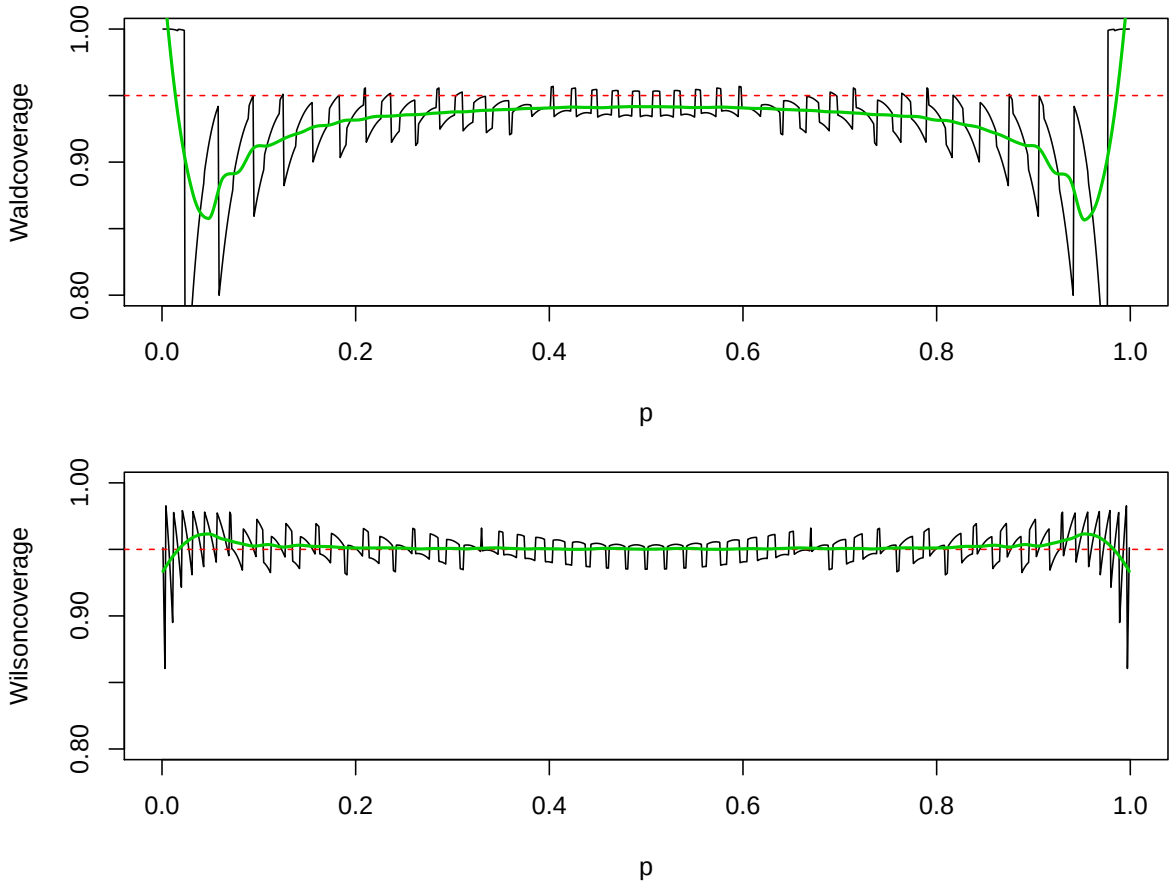


Abbildung 5.2: Überdeckungswahrscheinlichkeiten der 95%-Konfidenzintervalle für $X \sim \mathcal{B}in(n = 50, p)$. Rot gestrichelt ist das Ziel $1 - \alpha$, grün eine geglättete Kurve.

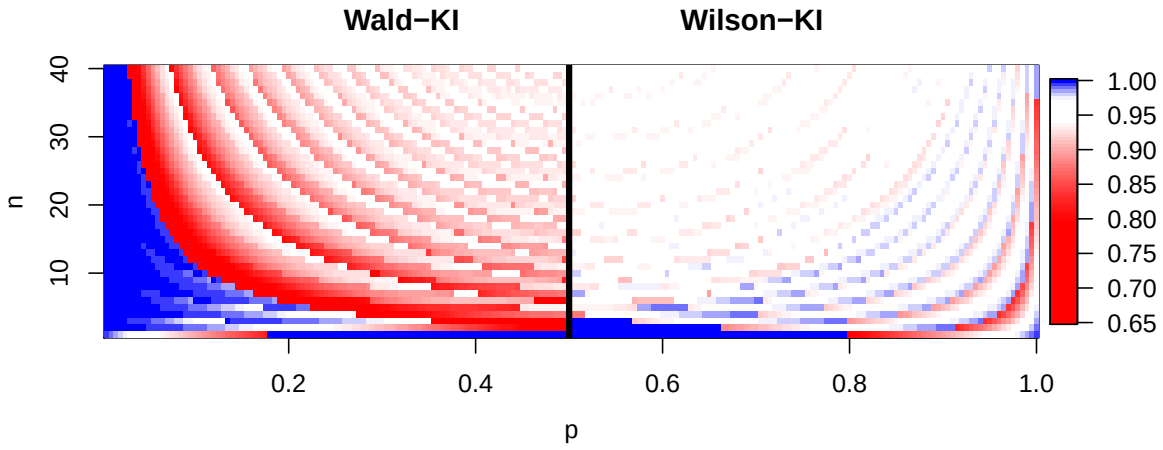


Abbildung 5.3: Überdeckungswahrscheinlichkeiten der 95%-Konfidenzintervalle für $X \sim \mathcal{B}in(n, p)$ als Funktion von p und n. Die Wahrscheinlichkeiten sind symmetrisch um $p = 1/2$.

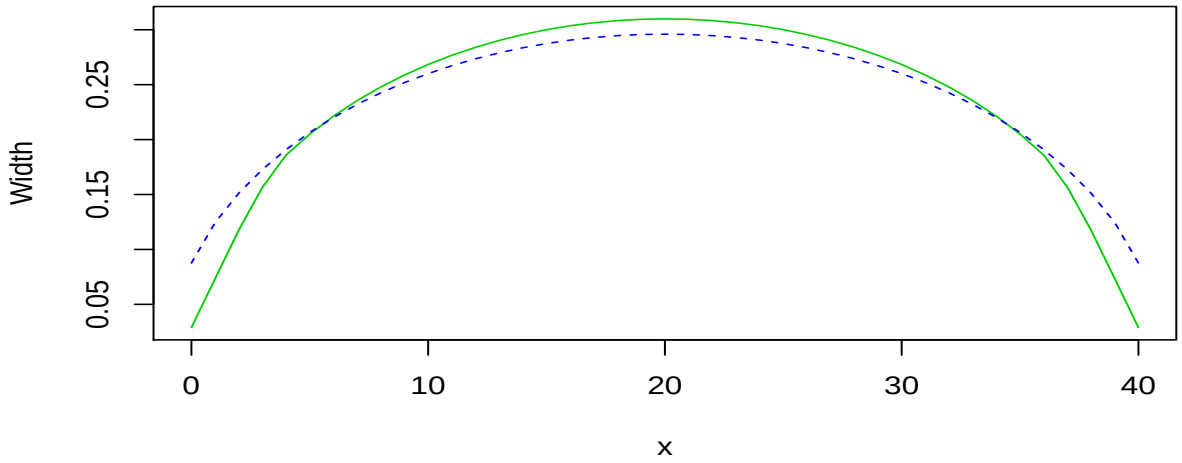


Abbildung 5.4: Breite der 95%-Konfidenzintervalle für $X \sim \mathcal{B}in(n = 40, p)$ (grün durchgezogen für Wald, blau gestrichelt für Wilson).

5.3 Testen

Die Präeklampsiedaten werden oft in einer sogenannten Vierfeldertafel (zweidimensionale Kontingenztafel, Kreuztabelle, *two-way table*) präsentiert, wie beispielsweise in Tabelle 5.1.

Wenn man wissen möchte, ob die Anteile in beiden Gruppen gleich sind, also ob sie aus der gleichen Verteilung stammen, kann man einen Test für Gleichheit der Proportionen durchführen, d.h. $H_0 : p_1 = p_2$. . Dieser Test wird auch Pearson's χ^2-Test genannt und ist in Test 6 gegeben und kann als Spezialfall vom Test 5 betrachtet werden.

Beispiel 5.1. Der R-Code 5.4 zeigt Resultate der Präeklampsiedaten. ♣

Tabelle 5.1: Beispiel für eine zweidimensionale Kontingenztafel, Häufigkeiten. Der erste Index bezieht sich auf das Risiko, der zweite auf die Diagnose.

		Diagnose		
		positiv	negativ	
Risiko	mit Faktor	h_{11}	h_{12}	n_1
	ohne Faktor	h_{21}	h_{22}	n_2

R-Code 5.3 Kontingenztafel für Präeklampsiedatenbeispiel.

```
xB <- 138
nB <- 1370
xK <- 175
nK <- 1336
tab <- rbind(B=c(xB, nB-xB), K=c(xK, nK-xK))
colnames( tab) <- c('pos','neg')
tab
```

```
##   pos  neg
## B 138 1232
## K 175 1161
```

5.4 Vergleich von Anteilen

Das Ziel ist zwei Anteile p_1 und p_2 zu vergleichen. Dies kann durch (i) eine Differenz $p_1 - p_2$ (ii) einen Quotienten p_1/p_2 oder (iii) ein Chancenverhältnis $p_1/(1-p_1) \Big/ (p_2/(1-p_2))$ erfolgen.

5.4.1 Differenz von Anteile

Basierend auf den Binomialverteilungen ist die Differenz $\frac{h_{11}}{h_{11}+h_{12}} - \frac{h_{21}}{h_{21}+h_{22}}$ approximativ normalverteilt

$$\mathcal{N}\Big(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\Big) \tag{5.15}$$

und ein entsprechendes Konfidenzintervall kann hergeleitet werden.

Die Risikoreduktion (*risk difference*) bezeichnet das (absolute) Ändern eines Ereignisses durch eine Behandlung.

Test 6: Test von Proportionen

Fragestellung: Sind die Anteile in zwei Gruppen gleich?

Voraussetzungen: Zwei ausreichend grosse, unabhängige Stichproben binärer Daten.

Berechnung: Wir verwenden die Bezeichnungen der Felder der Kreuztabelle wie in Tabelle 5.1. Die Teststatistik z berechnet sich dann als

$$z = \frac{(h_{11}h_{22} - h_{12}h_{21})^2(h_{11} + h_{12} + h_{21} + h_{22})}{(h_{11} + h_{12})(h_{21} + h_{22})(h_{12} + h_{22})(h_{11} + h_{21})}$$

und ist unter der Nullhypothese, dass beide Anteile gleich sind, χ^2-verteilt mit einem Freiheitsgrad.

Entscheidung: Vergleiche z und das $1-\alpha$ Quantil der $\chi^2(1)$-Verteilung Z_{Tab}. H_0 wird abgelehnt, falls $z > Z_{\text{Tab}}$.

Berechnung in R: `prop.test( tab)` oder `chisq.test(tab)`

5.4.2 Relatives Risiko

Das relative Risiko (*relative risk*) drückt aus, um welchen Faktor sich ein Risiko in zwei Gruppen unterscheidet:

$$\text{RR} = \frac{\text{P(Positive Diagnose mit Risikofaktor)}}{\text{P(Positive Diagnose ohne Risikofaktor)}}. \tag{5.16}$$

Die Gruppen mit oder ohne Risiko können natürlich auch als Behandlungs- und Kontrollgruppe betrachtet werden.

Das relative Risiko nimmt positive Werte an. Ein Wert von 1 bedeutet, dass das Risiko in beiden Gruppen gleich ist, und es besteht dementsprechend kein Anhaltspunkt für einen Zusammenhang zwischen der/m Diagnose/Erkrankung/Ereignis und dem Risikofaktor. Ein Wert grösser als eins ist ein Hinweis auf einen möglichen positiven Zusammenhang zwischen einem Risikofaktor und einer Diagnose/Erkrankung. Liegt das relative Risiko unter eins, hat die Exposition eine schützende (protektive) Wirkung, wie es beispielsweise bei Impfungen der Fall ist.

Ein Schätzwert vom relative Risiko ist (siehe Tabelle 5.1)

$$\widehat{\text{RR}} = \frac{\widehat{p}_1}{\widehat{p}_2} = \frac{\dfrac{h_{11}}{h_{11} + h_{12}}}{\dfrac{h_{21}}{h_{21} + h_{22}}}. \tag{5.17}$$

R-Code 5.4 Test von Proportionen

```
(RD <- tab[2,1]/ sum(tab[2,]) - tab[1,1]/ sum(tab[1,]) )
## [1] 0.0302581
prop.test(tab)
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  tab
## X-squared = 5.7619, df = 1, p-value = 0.01638
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.055107378 -0.005408816
## sample estimates:
##    prop 1    prop 2
## 0.1007299 0.1309880
chisq.test(tab)
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 5.7619, df = 1, p-value = 0.01638
```

Um Konfidenzintervalle zu konstruieren, wird zuerst $\widehat{\theta} = \log(\widehat{\mathrm{RR}})$ betrachted. Der Standardfehler von $\widehat{\theta}$ wird durch die Deltamethode bestimmt und basiert sich auf (2.34), angewandt auf Binomial anstelle von Bernoulli:

$$\mathrm{Var}(\widehat{\theta}) = \mathrm{Var}\left(\log(\widehat{\mathrm{RR}})\right) = \mathrm{Var}\left(\log(\frac{\widehat{p}_1}{\widehat{p}_2})\right) = \mathrm{Var}\left(\log(\widehat{p}_1) - \log(\widehat{p}_2)\right) \tag{5.18}$$

$$= \mathrm{Var}\left(\log(\widehat{p}_1)\right) + \mathrm{Var}((\log(\widehat{p}_2)\right) \approx \frac{1-p_1}{n_1 \cdot p_1} + \frac{1-p_2}{n_2 \cdot p_2} \tag{5.19}$$

$$\approx \frac{1 - \dfrac{h_{11}}{h_{11}+h_{12}}}{(h_{11}+h_{12}) \cdot \dfrac{h_{11}}{h_{11}+h_{12}}} + \frac{1 - \dfrac{h_{22}}{h_{21}+h_{22}}}{(h_{21}+h_{22}) \cdot \dfrac{h_{22}}{h_{21}+h_{22}}} \tag{5.20}$$

$$= \frac{1}{h_{11}} - \frac{1}{h_{11}+h_{12}} + \frac{1}{h_{21}} - \frac{1}{h_{21}+h_{22}}. \tag{5.21}$$

Eine Rücktransformation

$$\left[\ \exp\left(\widehat{\theta} \pm z_{1-\alpha/2}\,\mathrm{SE}(\widehat{\theta})\right)\ \right] \tag{5.22}$$

Konfidenzintervall für relative Risiko (RR)

Ein approximatives $(1-\alpha)$-Konfidenzintervall für RR basierend auf einer zweidimensionalen Kontingenztafel (Tabelle 5.1), ist

$$\left[\exp(\log(\widehat{\mathrm{RR}}) \pm z_{1-\alpha/2}\,\mathrm{SE}(\log(\widehat{\mathrm{RR}}))\right] \tag{5.23}$$

wobei $\widehat{\mathrm{RR}} = \dfrac{h_{11}(h_{21}+h_{22})}{(h_{11}+h_{12})h_{21}}$ und

$\mathrm{SE}(\log(\widehat{\mathrm{RR}})) = \sqrt{\dfrac{1}{h_{11}} - \dfrac{1}{h_{11}+h_{12}} + \dfrac{1}{h_{21}} - \dfrac{1}{h_{21}+h_{22}}}$ sind.

impliziert positive Konfidenzgrenzen.

Beispiel 5.2. Das relative Risiko und entsprechende Konfidenzintervall für die Präeklampsiedaten sind im R-Code 5.5 gegeben. Das relative Risiko ist kleiner als eins (Diuretika verringern das Risiko). Ein approximatives 95%-Konfidenzintervall liegt ausserhalb von eins. ♣

R-Code 5.5 Relative Risiko mit Konfidenzintervall.

```
RR <- ( tab[1,1]/ sum(tab[1,])) /  ( tab[2,1]/ sum(tab[2,]))
RR
## [1] 0.769001
s <- sqrt(
       1/tab[1,1] + 1/tab[2,1] - 1/sum(tab[1,]) - 1/sum(tab[2,]) )
exp( log(RR) + qnorm(c(.025,.975))*s)
## [1] 0.6233275 0.9487190
```

5.4.3 Odds Ratio

Das relative Risiko ist verwandt mit dem *Odds Ratio* (Quoten-, Chancenverhältnis, *odds ratio*)

$$\mathrm{OR} = \frac{\dfrac{\mathrm{P}(\text{Positive Diagnose mit Risikofaktor})}{\mathrm{P}(\text{Negative Diagnose mit Risikofaktor})}}{\dfrac{\mathrm{P}(\text{Positive Diagnose ohne Risikofaktor})}{\mathrm{P}(\text{Negative Diagnose ohne Risikofaktor})}} \tag{5.24}$$

$$= \frac{\dfrac{\mathrm{P}(A)}{1-\mathrm{P}(A)}}{\dfrac{\mathrm{P}(B)}{1-\mathrm{P}(B)}} = \frac{\mathrm{P}(A)(1-\mathrm{P}(B))}{\mathrm{P}(B)(1-\mathrm{P}(A))} \tag{5.25}$$

das etwas über die Stärke eines Zusammenhangs von zwei Merkmalen aussagt (Assoziationsmass). Die Berechnung des Odds Ratio macht auch Sinn, wenn die Anzahl der Erkrankten durch das Studiendesign vorgegeben ist wie in Fall-Kontroll-Studien.

Wenn die Wahrscheinlichkeit zu erkranken gering ist, sind Odds Ratio und relatives Risiko ungefähr gleich.

Ein Schätzwert des Odds Ratio ist

$$\widehat{\mathrm{OR}} = \frac{\dfrac{h_{11}}{h_{12}}}{\dfrac{h_{21}}{h_{22}}} = \frac{h_{11}\,h_{22}}{h_{12}\,h_{21}} \tag{5.26}$$

Die Konstruktion des Konfidenzintervalles für das Chancenverhältnis basiert sich auf (2.33) und (2.34), analog zu dem des relativen Risikos.

Konfidenzintervall für das Chancenverhältnis (OR)

Ein approximatives $(1-\alpha)$-Konfidenzintervall für OR basiered auf einer zweidimensionalen Kontingenztafel (Tabelle 5.1), ist

$$\left[\exp\Big(\log(\widehat{\mathrm{OR}}) \pm z_{1-\alpha/2}\,\mathrm{SE}(\log(\widehat{\mathrm{OR}}))\Big)\right] \tag{5.27}$$

wobei $\widehat{\mathrm{OR}} = \dfrac{h_{11}h_{22}}{h_{12}h_{21}}$ und $\mathrm{SE}(\log(\widehat{\mathrm{OR}})) = \sqrt{\dfrac{1}{h_{11}} + \dfrac{1}{h_{21}} + \dfrac{1}{h_{12}} + \dfrac{1}{h_{22}}}$ sind.

Beispiel 5.3. Die Odds Ratio mit Konfidenzintervall für die Präeklampsiedaten sind im R-Code 5.6 gegeben. ♣

R-Code 5.6 Odds Ratio mit Konfidenzintervall, approximativ und exakt nach Fisher.

```
OR <- tab[1]*tab[4]/(tab[2]*tab[3])
OR
## [1] 0.7431262
s <- sqrt( sum( 1/tab) )
exp( log(OR) + qnorm(c(.025,.975))*s)
## [1] 0.5862636 0.9419594
# Exakter Test nach Fisher:
fisher.test(tab)
##
##  Fisher's Exact Test for Count Data
##
## data:  tab
## p-value = 0.01609
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.581660 0.948257
## sample estimates:
## odds ratio
##  0.7432123
```

Kapitel 6

Rangbasierte Methoden

In diesem Kapitel diskutieren wir Ansätze zum Schätzen und Testen für Fälle, in denen die Daten nicht normal verteilt sind. Das heisst, es können Ausreisser vorhanden sein, die Daten können nicht intervallskaliert sein, etc.

6.1 Robuste Schätzung von Kennzahlen

"Klassische" Schätzfunktionen des Mittelwerts und der Varianz sind

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Wenn wir einen beliebigen Wert x_i hypothetisch auf einen unendlich grossen Wert setzen (Ausreisser), "explodieren" auch die beiden obigen Schätzer.

Robuste Schätzfunktionen sind nicht sensibel auf Ausreisser. Ein robuster Schätzer des Mittelwerts ist das gestutzte Mittel (*trimmed mean*), bei dem die grössten und kleinsten Werte nicht berücksichtigt werden.

Der Median (*median*) (der mittelste bei einer ungereaden Anzahl oder das Mittel der beiden mittelsten bei einer geraden Anzahl) ist ein weiterer robuster Schätzer des Mittelwerts.

Robuste Schätzer der Streuung sind (1) der Interquartilsabstand (*interquartile range*), abgekürzt IQR und berechnet als Differenz des dritten und ersten Quartiles und (2) die mittlere absolute Abweichung (*median absolute deviation*), abgekürzt MAD, berechnet als

$$\text{MAD} = c \cdot \text{median}|x_i - \text{median}\{x_1, \ldots, x_n\}| \tag{6.1}$$

wobei die meisten Softwareprogramme $c = 1.4826$ wählen. Die Wahl von c ist so, dass für normalverteilte Zufallsvariablen $\text{E(MAD)} = \sigma$. Da für normalverteilte Zufallsvariablen $\text{IQR} = 2\Phi^{-1}(3/4)\sigma$, ist IQR/1.349 eine Schätzfunction von σ.

Beispiel 6.1. Gegeben seien die Werte $1.1, 3.2, 2.2, 1.8, 1.9, 2.1, 17$. Der R-Code 6.1 vergleicht einige Kennzahlen. ♣

R-Code 6.1 Klassische und robuste Schätzwerte vom Beispiel 6.1.

```
sam <- c(1.1, 3.2, 2.2, 1.8, 1.9, 2.1, 17)
print( c(mean(sam), mean(sam, trim=.2), median(sam)))
## [1] 4.185714 2.240000 2.100000
print( c(sd(sam), IQR(sam), mad(sam)))
## [1] 5.684901 0.850000 0.444780
```

Die Schätzfunktionen des gestutzten Mittels oder des Medians besitzen keine einfache Verteilungsfunktion. Deshalb können die entsprechenden Konfidenzintervalle nicht auf einfache Art berechnet werden. Basierend auf

$$\widehat{\text{Mittelwert}} \pm z_{\alpha/2}\sqrt{\frac{\widehat{\text{Varianz}}}{n}} \tag{6.2}$$

können einfache approximative Konfidenzintervalle konstruiert werden. Zum Beispiel ist `median( x)+c(-2,2)*mad( x)/sqrt(length( x))` ein approximatives 95%-Konfidenzintervall für den Mittelwert.

Ein weiterer Nachteil von robusten Schätzern ist die etwas geringere Effizienz (*efficiency*), das heisst, die Schätzer haben eine etwas grössere Varianz. In einigen Fällen kann die exakte Varianz bestimmt werden. Asymptotisch ist der Median aber auch wieder normalverteilt um den Median η mit Varianz $(4nf(\eta)^2)^{-1}$, wobei $f(x)$ die Dichtefunktion ist.

Beispiel 6.2. Seien $X_1, \ldots, X_{10} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Abbildung 6.1 zeigt die (geglättete) empirische Dichte von $\overline{X}$ basierend auf $N = 1000$ Realisationen. Aufgrund der symmetrischen Dichte ist $\eta = \mu = 0$ und somit ist die asymptotische Effizienz

$$\frac{\sigma^2/n}{1/(4nf(0)^2)} = \frac{\sigma^2}{n} \cdot 4n\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^2 = \frac{2}{\pi} \approx 64\%. \tag{6.3}$$

♣

Die Entscheidung, ob eine Realisation einer Zufallsstichprobe Ausreisser enthält, ist nicht immer einfach. Bei allen Verteilungen mit Werten aus $\mathbb{R}$ werden bei genügend grossem n im Boxplot Ausreisser markiert. Bei eindeutigen Ausreissern ist es oft einfach, diese zu identifizieren und eliminieren. Im Grenzfall sind robuste Schätzmethoden zu bevorzugen.

Bei Vektorstichproben ist es oft sehr schwer Ausreisser zu erkennen, da diese bezüglich den Randverteilungen oft nicht auffallen. Robuste Methoden für Zufallsvektoren existieren, sind aber oft rechenintensiv und nicht so intuitiv wie für skalare Werte.

R-Code 6.2 Verteilung vom empirischen Mittel und Median, siehe Beispiel 6.2. (Siehe Abbildung 6.1.)

```
n <- 10
N <- 1000
samples <- array( rnorm(n*N), c(N,n))
means <- apply( samples, 1, mean)
medians <- apply( samples, 1, median)
print( c( var(means), var(medians),  var(means)/var(medians)))
## [1] 0.1030284 0.1403173 0.7342531
hist( medians, border=7, col=7, prob=T, main='', ylim=c(0,1.2))
hist( means, add=T, prob=T)
lines( density( medians), col=2)
lines( density( means), col=1)
# with a t_4 the situation is different!
samples <- array( rt(n*N, df=4), c(N,n))
means <- apply( samples, 1, mean)
medians <- apply( samples, 1, median)
print( c( var(means), var(medians),  var(means)/var(medians)))
## [1] 0.2010834 0.1587618 1.2665732
```

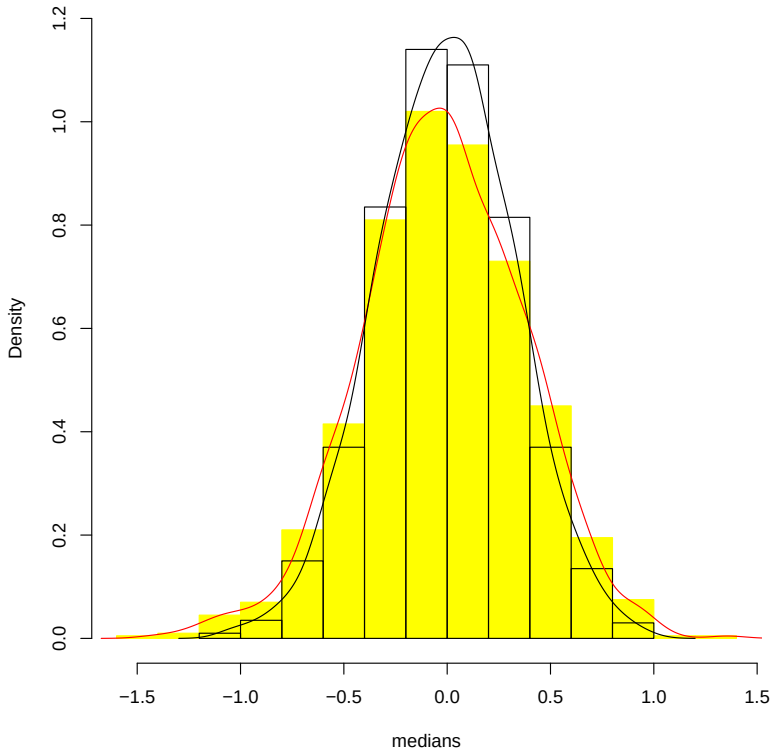


Abbildung 6.1: Effizienz von Schätzfunktionen. Mediane in gelb mit roter Dichte, Mittelwerte in schwarz. (Siehe R-Code 6.2.)

6.2 Rangbasierte Tests

Im Kapitel 4 haben wir Tests zum Vergleichen von Mittelwerten betrachtet. Diese Tests setzen normalverteilte Daten voraus. Falls Ausreisser vorhanden sind oder die Daten ordinalskaliert sind, ist es empfehlenswert, rangbasierte Tests zu verwenden. Rangbasierte Tests werden oft auch nicht-parametrische Tests genannt.

Der Rang eines Wertes in einer Sequenz ist die Position (*order*) der geordneten vom kleinsten zum grössten Sequenz. Insbesondere hat die kleinste Zahl den Rang 1 und die grösste den Rang n. Bei Mehrfachauftretungen werden die Arithmetischen Mittel der Ränge gegeben.

Beispiel 6.3. Die Werte 1.1, -0.6, 0.3, 0.1, 0.6, 2.1 haben die Ränge 5, 1, 3, 2, 4 und 6. Hingegen sind die Ränge der Absolutwerte 5, (3+4)/2, 2, 1, (3+4)/2 und 6. ♣

Bei Rangtests werden nur die Ränge der Beobachtungen berücksichtigt und nicht deren (absoluten) Werte. Zum Beispiel hat der grösste Wert immer denselben Rang und somit immer den gleichen Einfluss die Teststatistik.

Wir betrachten zuerst den U-Test (Wilcoxon-Mann-Whitney-Test) und den Wilcoxon-Test, entsprechend zu den Tests 2 und Test 3.

6.2.1 Wilcoxon-Mann-Whitney-Test

Wir nehmen hier an, dass die Zufallsvariablen X und Y stetige Verteilungsfunktionen F_X und F_Y haben, die sich um eine Verschiebung δ unterscheiden

$$F_Y(x) = F_X(x - \delta). \tag{6.4}$$

Der Wilcoxon-Mann-Whitney-Test (*Mann–Whitney U test*, *Wilcoxon rank-sum test* oder *Wilcoxon–Mann–Whitney test*) testet die Nullhypothese $H_0 : \delta = 0$. Natürlich impliziert ein verwerfen der Hypothese eine signifikante Differenz in den Mittelwerten oder den Medians.

Die Quantil-, Dichte- und Verteilungsfunktion der Teststatistik des U-Test ist in R mit `[q,d,p]wilcox` implementiert.

Liegt statt der U-Tabelle nur eine z-Tabelle vor, so transformiert man wie folgt

$$z_{\text{Vers}} = \frac{\left| U_{\text{Vers}} - \frac{n_x \cdot n_y}{2} \right|}{\sqrt{\frac{n_x \cdot n_y \cdot (n_x + n_y + 1)}{12}}} \tag{6.5}$$

und vergleicht mit dem entsprechenden Quantil der Standardnormalverteilung. Hierbei sollten die Stichproben nicht zu klein sein, $n_x \geq 8$ und $n_y \geq 2$.

Um Konfidenzintervalle zu konstruieren, muss im Aufruf von `wilcox.test` das Argument `conf.int=TRUE` gesetzt werden.

Test 7: Lagevergleich zweier unabhängiger Stichproben

Fragestellung: Sind die Mediane zweier unabhängiger Stichproben X und Y signifikant verschieden?

Voraussetzungen: Die beiden Grundgesamtheiten folgen stetigen Verteilungen von gleicher Form, die Stichproben sind unabhängig und die Daten mindestens ordinalskaliert.

Berechnung: Sei $n_x \leq n_y$, ansonsten tausche die Stichproben. Bringe die (n_x+n_y) Stichprobenwerte in eine *gemeinsame* Rangfolge und berechne die Summe R_x und R_y der Ränge der Stichproben Berechne U_x und U_y

- $U_x = n_x \cdot n_y + \frac{n_x \cdot (n_x+1)}{2} - R_x$
- $U_y = n_x \cdot n_y + \frac{n_y \cdot (n_y+1)}{2} - R_y$ (Probe: $U_x + U_y = n_x \cdot n_y$)
- $U_{\text{Vers}} = \min(U_x, U_y)$

Entscheidung: Vergleiche U_{Vers} und $U_{\text{Tab}}(n_x, n_y; \alpha)$: verwerfe H_0: "Mediane gleich" wenn $U_{\text{Vers}} > U_{\text{Tab}}$.

Berechnung in R: `wilcox.test(x, y, conf.level=1-alpha)`

6.2.2 Wilcoxon-Vorzeichen-Rang-Test

Analog zum gepaarten t-test, ist der Wilcoxon-Vorzeichen-Rang-Test (*Wilcoxon signed-rank test*) für verbundene Stichproben und kann entsprechend auch benuzt werden um den Median mit einem theoretischen Wert zu vergleichen. Die Stichproben müssen nicht notwendigerweise aus einer symmetrischen Verteilung stammen.

Die Quantil-, Dichte- und Verteilungsfunktion der Teststatistik des U-Test ist in R mit `[q,d,p]signrank` implementiert.

Liegt statt der W-Tabelle nur eine z-Tabelle vor, so transformiert man wie folgt

$$z_{\text{Vers}} = \frac{\left| W_{\text{Vers}} - \frac{n \cdot (n+1)}{4} \right|}{\sqrt{\frac{n \cdot (n+1) \cdot (2n+1)}{24}}}. \tag{6.6}$$

und vergleicht mit dem entsprechenden Quantil der Standardnormalverteilung. Hierbei sollten die Stichproben nicht zu klein sein, $n_\star \geq 20$.

Test 8: Lagevergleich zweier verbundener Stichproben

Fragestellung: Sind die Mediane zweier verbundener Stichproben X und Y signifikant verschieden?

Voraussetzungen: Die beiden Grundgesamtheiten folgen stetigen Verteilungen von gleicher Form, die Stichproben sind verbunden und die Daten mindestens intervallskaliert.

Berechnung: (1) Berechne die Messwertdifferenzen $d_i = x_i - y_i$. Im weiteren bleiben alle Differenzen $d_i = 0$ unberücksichtigt. Seien also noch $n_\star$ Differenzen $d_i \neq 0$ zu betrachten.

(2) Bringe diese $n_\star$ Messwertdifferenzen d_i entsprechend ihrer Absolutbeträge $|d_i|$ in eine Rangfolge.

(3) Berechne die Summe W^+ über die Rangzahlen aller positiven Messwertdifferenzen $d_i > 0$ und entsprechend die Summe W^- aller negativen Differenzen $d_i < 0$

$$W_{\text{Vers}} = \min(W^+, W^-) \qquad (\text{Probe: } W^+ + W^- = \tfrac{n_\star \cdot (n_\star + 1)}{2})$$

Entscheidung: Vergleiche W_{Vers} mit $W_{\text{Tab}}(n_\star; \alpha)$: verwerfe H_0: "Mediane gleich" wenn $W_{\text{Vers}} > W_{\text{Tab}}$.

Berechnung in R: `wilcox.test(x-y, conf.level=1-alpha)` oder

`wilcox.test(x, y, paired=TRUE, conf.level=1-alpha)`

Beispiel 6.4. Wir betrachten nochmals die `essmenge` Daten. Der R-Code 6.3 führt den Vorzeichentest und die Wilcoxon-Tests durch. Wie erwartet sind die p-Werte ähnlich zu denjenigen im Kapitel 4. Weil bei den Tests gleiche Werte auftreten, können p-Werte nicht exakt berechnet werden. Um `warnings` zu vermeiden, wird das Argument `exact=FALSE` gesetzt.

Wird hingegen der erste Wert von 5.17 zu 51.7 vertauscht, ist der Effekt von robusten Methoden klar ersichtlich, wie in R-Code 6.4 aufgegezeigt. ♣

R-Code 6.3 Rangtests

```
d <- Essmenge.normal - Essmenge.waehrend
binom.test( sum(d>0), length(d))
##
##  Exact binomial test
##
## data:  sum(d > 0) and length(d)
## number of successes = 4, number of trials = 17, p-value =
## 0.04904
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.06810774 0.49899327
## sample estimates:
## probability of success
##              0.2352941
wilcox.test( Essmenge.normal, mu=9, exact=FALSE)
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Essmenge.normal
## V = 25, p-value = 0.02792
## alternative hypothesis: true location is not equal to 9
wilcox.test( Essmenge.normal, Essmenge.waehrend, exact=FALSE)
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Essmenge.normal and Essmenge.waehrend
## W = 118, p-value = 0.3703
## alternative hypothesis: true location shift is not equal to 0
wilcox.test( Essmenge.normal, Essmenge.waehrend, paired=TRUE)
## Warning in wilcox.test.default(Essmenge.normal, Essmenge.waehrend,
paired = TRUE): cannot compute exact p-value with ties
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Essmenge.normal and Essmenge.waehrend
## V = 31.5, p-value = 0.03513
## alternative hypothesis: true location shift is not equal to 0
```

R-Code 6.4 Gepaarte Tests mit einer korrupten Beobachtung.

```
Essmenge.waehrend[1] <- Essmenge.waehrend[1] * 10
t.test( Essmenge.normal, Essmenge.waehrend, paired=TRUE)
##
##  Paired t-test
##
## data:  Essmenge.normal and Essmenge.waehrend
## t = -1.1308, df = 16, p-value = 0.2748
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.900265   3.316736
## sample estimates:
## mean of the differences
##               -3.791765
wilcox.test( Essmenge.normal, Essmenge.waehrend, paired=TRUE,
            conf.int=TRUE, exact=FALSE)
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Essmenge.normal and Essmenge.waehrend
## V = 30.5, p-value = 0.03123
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -1.17499609 -0.05001666
## sample estimates:
## (pseudo)median
##     -0.4999147
Essmenge.waehrend[1] <- Essmenge.waehrend[1] / 10
```

6.3 Weitere Tests

6.3.1 Vorzeichentest

Bei gepaarten Daten $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$ aus symmetrischen Verteilungen betrachten wir das Vorzeichen der Differenz $D_i = Y_i - X_i$. Unter der Nullhypothese sind die Vorzeichen binomialverteilt mit Wahrscheinlichkeit $p = 0.5$. Bei sehr wenigen oder sehr vielen negativen Vorzeichen wird die Hypothese verworfen. Der Test basiert auf sehr schwachen Voraussetzungen und hat dementsprechend eine kleinere Power.

Test 9: Vorzeichentest

Fragestellung: Sind die Mediane zweier verbundener Stichproben X und Y signifikant verschieden?

Voraussetzungen: Die beiden Grundgesamtheiten folgen stetigen Verteilungen von gleicher Form, die Stichproben sind unabhängig und die Daten mindestens ordinalskaliert.

Berechnung: (1) Berechne die Messwertdifferenzen $d_i = x_i - y_i$. Im weiteren bleiben alle Differenzen $d_i = 0$ unberücksichtigt. Seien also noch $n_\star$ Differenzen $d_i \neq 0$ zu betrachten.

(2) Ordne jeder Differenz d_i ihr Vorzeichen zu (also + oder −) und zähle die Anzahl der positiven Vorzeichen: $k = \sum d_+$.

(3) Falls die Mediane der beiden Stichproben gleich sind, folgt k einer Binomialverteilung $B(n, p_0)$ mit $p_0 = 0.5$.

Entscheidung: Vergleiche k mit $b_{\text{Tab}}(n, p_0, 1-\frac{\alpha}{2})$ und $b_{\text{Tab}}(n, p_0, \frac{\alpha}{2})$. Verwerfe $H_0 : p = 0.5$, falls $k > b_{\text{Tab}}(n, p_0, 1-\frac{\alpha}{2})$ oder $k < b_{\text{Tab}}(n, p_0, \frac{\alpha}{2})$.

Berechnung in R:
```
binom.test( sum( d>theta0), length(d),
conf.level=1-alpha)
```

6.3.2 Permutationstest

Permutationstests können zur Beantwortung vieler verschiedener Fragestellungen verwendet werden. Sie basieren auf der Idee, dass unter der Nullhypothese die zu vergleichenden Stichproben gleich sind. In diesem Fall würde sich auch bei einer zufälligen Verteilung der Werte auf die Gruppen nichts am Ergebnis des Tests ändern. Wir zeigen hier exemplarisch einen Permutationstest zum Vergleich der Mittelwerte zweier unabhängiger Stichproben.

Beachten Sie, dass das Paket `exactRankTests` nicht mehr weiterentwickelt wird. Daher wird empfohlen, statt der Funktion `perm.test` aus diesem Paket die Funktionen des Pakets `coin` zu verwenden, welches auch Erweiterungen zu den anderen Rangtests enthält.

Test 10: Permutationstest

Fragestellung: Sind die Mittelwerte zweier unabhängiger Stichproben X und Y signifikant verschieden?

Voraussetzungen: Nullhypothese ist so formuliert, dass die Gruppen unter H_0 austauschbar sind.

Berechnung: (1) Berechne die Differenz der Mittelwerte D_{obs} der beiden zu vergleichenden Gruppen (m Beobachtungen in Gruppe 1, n Beobachtungen in Gruppe 2).

(2) Bilde eine zufällige Permutation der Werte der beiden Gruppen, das heisst, teile die beobachteten Werte zufällig auf die beiden Gruppen auf (m Beobachtungen in Gruppe 1, n Beobachtungen in Gruppe 2). Hierzu gibt es $\binom{m+n}{n}$ Möglichkeiten.

(3) Berechne die Differenz der Mittelwerteder beiden neu gebildeten Gruppen.

(4) Wiederhole dieses Vorgehen viele Male.

Entscheidung: Der p-Wert kann berechnet werden als

$$\frac{\text{Anzahl der permutierten Differenzen, die extremer sind als } D_{\text{obs}}}{\binom{m+n}{n}}$$

und wird dann mit dem gewählten Signifikanzniveau verglichen.

Berechnung in R:

```
require(coin)
oneway_test( formula, data)
```

R-Code 6.5 Permutationstest.

```
require(coin)
Essen <- c(Essmenge.normal, Essmenge.waehrend)
Group <-factor(c(rep(0, length(Essmenge.normal)),
                 rep(1, length(Essmenge.waehrend))))

oneway_test(Essen ~ Group)
##
##  Asymptotic Two-Sample Fisher-Pitman Permutation Test
##
## data:  Essen by Group (0, 1)
## Z = -0.71479, p-value = 0.4747
## alternative hypothesis: true mu is not equal to 0
```

Kapitel 7

Multivariate Normalverteilung

7.1 Zufallsvektoren

Ein Zufallsvektor ist ein (Spalten-)Vektor $\mathbf{X} = (X_1, \ldots, X_p)^\mathsf{T}$ mit p Zufallsvariablen als Komponenten. Die mehrdimensionale, d.h. multivariate, Verteilungsfunktion von $\mathbf{X}$ ist definiert als

$$F_{\mathbf{X}}(\boldsymbol{x}) = \mathrm{P}(\mathbf{X} \leq \boldsymbol{x}) = \mathrm{P}(X_1 \leq x_1, \ldots, X_p \leq x_p), \tag{7.1}$$

wobei Liste ist als Schnittmenge ($\cap$) zu verstehen ist. Die multivariate Verteilungsfunktion enthält in der Regel mehr Information als die Menge der Marginalen Verteilungen, da sich (7.1) (nur) unter Unabhängigkeit zu $F_{\mathbf{X}}(\boldsymbol{x}) = \prod_{i=1}^{p} \mathrm{P}(X_i \leq x_i)$ vereinfachen lässt (vergleiche Gleichung (2.3)).

Ähnlich wie bei Zufallsvariablen sind die Dichtefunktionen und Wahrscheinlichkeitsfunktionen definiert. Einfachheitshalber fassen wir hier für einen Zufallsvektoren mit zwei stetigen Komponenten $F_{X,Y}(x, y)$ einige Punkte zusammen:

- $f_{X,Y}(x, y) = \dfrac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$;
- $\mathrm{P}(\mathbf{X} \in A) = \displaystyle\int_A f_{X,Y}(x, y) dx dy$;
- $f_X(x) = \displaystyle\int_{\mathbb{R}} f_{X,Y}(x, y) dy$;
- $F_X(x) = F_{X,Y}(x, \text{`}\infty\text{'}) dy$.

Die Kovarianz (*covariance*) zwischen zwei Zufallsvariablen X_1 und X_2 ist definiert als

$$\mathrm{Cov}(X_1, X_2) = \mathrm{E}\big((X_1 - \mathrm{E}(X_1))(X_2 - \mathrm{E}(X_2))\big). \tag{7.2}$$

Die Kovarianz beschreibt den *linearen* Zusammenhang zwischen den Zufallsvariablen. Es gilt:

$$\operatorname{Cov}(X_1, X_2) = \operatorname{Cov}(X_2, X_1) \tag{7.3}$$

$$\operatorname{Cov}(X_1, X_1) = \operatorname{Var}(X_1) \tag{7.4}$$

$$\operatorname{Cov}(a + bX_1, c + dX_2) = bd \operatorname{Cov}(X_1, X_2) \tag{7.5}$$

Die Korrelation (*correlation*) zwischen zwei Zufallsvariablen X_1 und X_2 ist definiert als

$$\operatorname{Corr}(X_1, X_2) = \frac{\operatorname{Cov}(X_1, X_2)}{\sqrt{\operatorname{Var}(X_1) \operatorname{Var}(X_2)}} \tag{7.6}$$

und entspricht der *normierten* Kovarianz. Es gilt $-1 \leq \operatorname{Corr}(X_1, X_2) \leq 1$.

Definition 7.1. Der Erwartungswert eines Zufallsvektors $\mathbf{X}$ ist definiert als

$$\operatorname{E}(\mathbf{X}) = \operatorname{E}\left(\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}\right) = \begin{pmatrix} \operatorname{E}(X_1) \\ \vdots \\ \operatorname{E}(X_p) \end{pmatrix} \tag{7.7}$$

und die Varianz eines Zufallsvektors $\mathbf{X}$ ist definiert als

$$\operatorname{Var}(\mathbf{X}) = \operatorname{E}\big((\mathbf{X} - \operatorname{E}(\mathbf{X}))(\mathbf{X} - \operatorname{E}(\mathbf{X}))^\mathsf{T}\big) \tag{7.8}$$

$$= \operatorname{Var}\left(\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}\right) = \begin{pmatrix} \operatorname{Var}(X_1) & \dots & \operatorname{Cov}(X_i, X_j) \\ & \ddots & \\ \operatorname{Cov}(X_j, X_i) & \dots & \operatorname{Var}(X_p) \end{pmatrix}. \tag{7.9}$$

Ähnlich zu Eigenschaften 2.4 haben wir für Zufallsvektoren folgende Eigenschaften.

Eigenschaften 7.1. *Für beliebige multivariate Zufallsvektoren* $\mathbf{X}$*, Vektoren* $\boldsymbol{a} \in \mathbb{R}^q$ *und Matrizen* $\mathbf{B} \in \mathbb{R}^{q \times p}$ *gilt:*

- $\operatorname{Var}(\mathbf{X}) = \operatorname{E}(\mathbf{X}\mathbf{X}^\mathsf{T}) - \operatorname{E}(\mathbf{X}) \operatorname{E}(\mathbf{X})^\mathsf{T}$
- $\operatorname{E}(\boldsymbol{a} + \mathbf{B}\mathbf{X}) = \boldsymbol{a} + \mathbf{B} \operatorname{E}(\mathbf{X})$,
- $\operatorname{Var}(\boldsymbol{a} + \mathbf{B}\mathbf{X}) = \mathbf{B} \operatorname{Var}(\mathbf{X}) \mathbf{B}^\mathsf{T}$

Wir betrachten hier nur eine spezielle multivariate Verteilung: die multivariate Normalverteilung.

7.2 Bivariate Normalverteilung

Definition 7.2. Das Zufallspaar (X, Y) ist bivariat normalverteilt, falls gilt

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(x, y) dx dy \tag{7.10}$$

mit Dichte

$$\begin{aligned} f(x, y) &= f_{X,Y}(x, y) \\ &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right), \end{aligned} \tag{7.11}$$

für alle x und y ($\mu_x \in \mathbb{R}$, $\mu_y \in \mathbb{R}$, $\sigma_x > 0$, $\sigma_y > 0$ und $-1 < \rho < 1$). Die Randverteilungen (Marginalverteilung, *marginal distribution*) sind $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ und $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$.

Beispiel 7.1. R Code 7.2 und Abbildung 7.2 zeigen die Dichte einer bivariat normalverteilten Zufallszahle mit $\mu_x = \mu_y = 0$ und $\sigma_x = 1$, $\sigma_y = \sqrt{5}$ und $\rho = 2/\sqrt{5} \approx 0.9$. Durch die quadratische Form in (7.11) sind die Höhenlinien (*contour lines*) (Isolinien, *isolines*) Ellipsen.

Einige R Pakete implementieren die bivariate/multivariate Normalverteilung. Wir empfehlen `mvtnorm`. ♣

R-Code 7.1 Dichte einer bivariaten Normalverteilung. (Siehe Abbildung 7.1.)

```
require( mvtnorm)
require( fields) # providing tim.colors()
Sigma <- array( c(1,2,2,5), c(2,2))
x1 <- x2 <- seq( -3, to=3, length=100)
grid <- expand.grid( x1, x2)
densgrid <- dmvnorm( grid, mean=c(0,0), sigma=Sigma)
dens <- array( densgrid, c(100,100))
image(x1, x2, dens, col=tim.colors())          # left panel
faccol <- tim.colors()[cut(dens[-1,-1],64)]
persp(x1, x2, dens, col=faccol, border = NA,  # right panel
      tick='detailed', theta=120, phi=30, r=100)
```

Eigenschaften 7.2. *Für die bivariate Normalverteilung gilt:*

$$\mathrm{E}\left[\begin{pmatrix} X \\ Y \end{pmatrix}\right] = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \qquad \mathrm{Var}\left[\begin{pmatrix} X \\ Y \end{pmatrix}\right] = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}. \tag{7.12}$$

Somit

$$\mathrm{Cov}(X, Y) = \rho\sigma_x\sigma_y, \qquad \mathrm{Corr}(X, Y) = \rho. \tag{7.13}$$

Falls $\rho = 0$, sind X und Y unabhängig und umgekehrt.

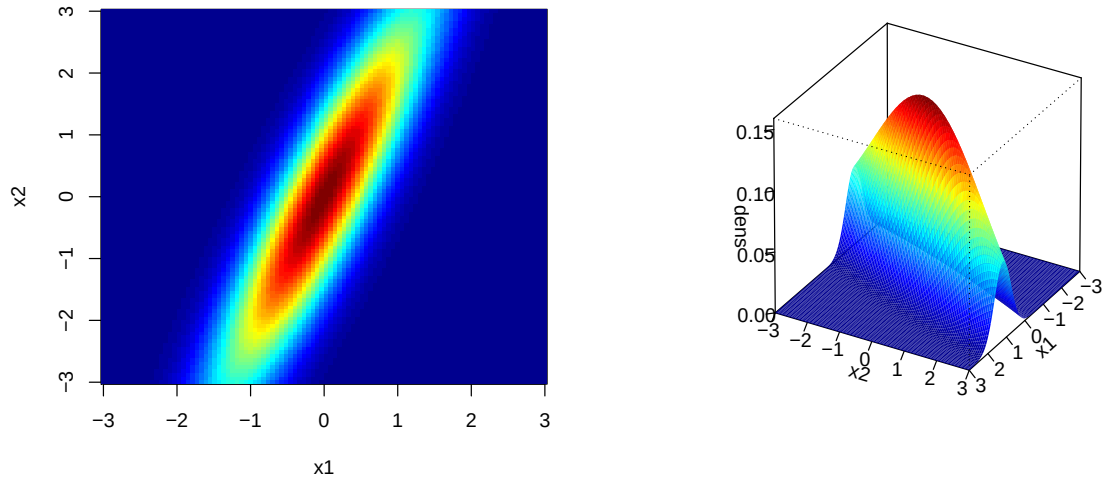


Abbildung 7.1: Dichte einer bivariaten Normalverteilung. (Siehe R-Code 7.1.)

Beispiel 7.2. R Code 7.2 und Abbildung 7.2 zeigen Realisationen von bivariat normalverteilte Zufallszahlen für verschiedene Korrelationen ρ. Selbst für grosse Stichproben (hier $n = 500$) sind Korrelationen zwischen -0.25 und 0.25 nur erahnbar. ♣

R-Code 7.2 Realisationen 500 bivariat normalverteilten Zufallszahlen für verschiedene Werte von ρ (Siehe Abbildung 7.2.)

```
set.seed(12)
rho <- c(-.25, 0, .1, .25, .75, .9)
for (i in 1:6) {
  Sigma <- array( c(1, rho[i], rho[i], 1), c(2,2))
  sample <- rmvnorm( 500, sigma=Sigma)
  plot(sample, pch='.', xlab='', ylab='')
  legend( "topleft", legend=bquote(rho==.(rho[i])), bty='n')
}
```

7.3 Multivariate Normalverteilung

Definition 7.3. Der Zufallsvektor $\mathbf{X} = (X_1, \ldots, X_p)^\mathsf{T}$ is multivariat normalverteilt, falls gilt

$$F_{\mathbf{X}}(\boldsymbol{x}) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_p} f_{\mathbf{X}}(x_1, \ldots, x_p) dx_1 \ldots dx_p \tag{7.14}$$

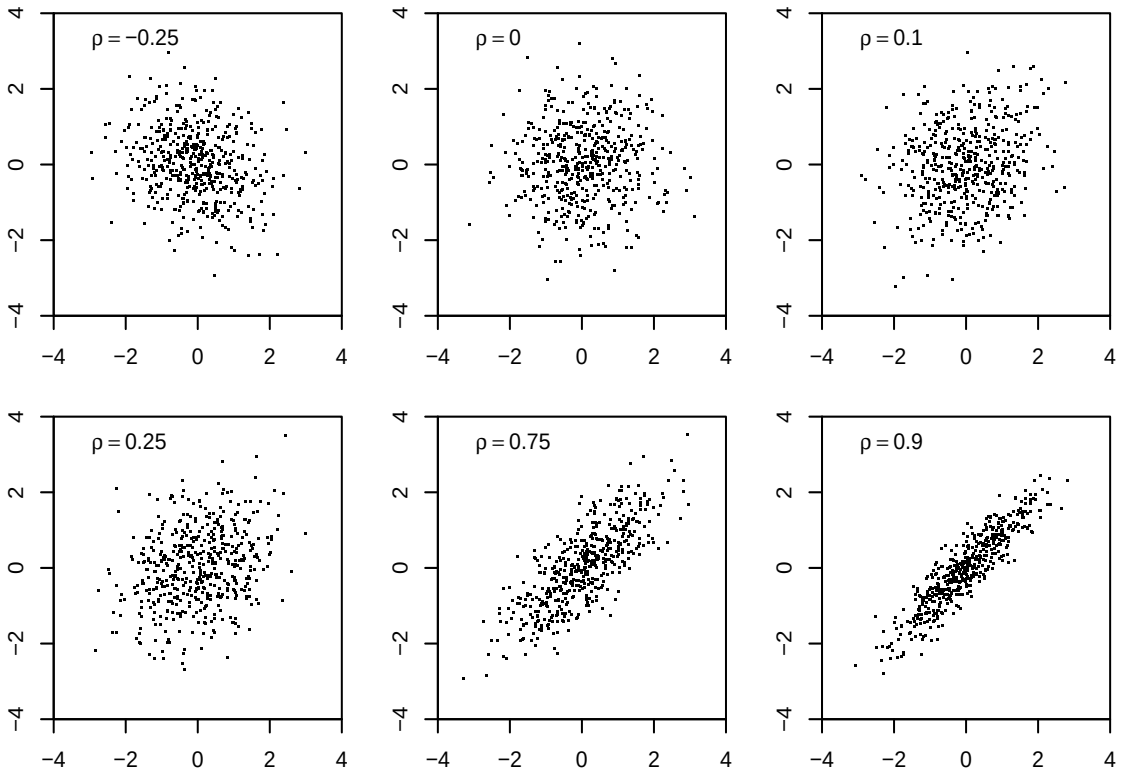


Abbildung 7.2: Realisationen von bivariat normalverteiten Zufallszahlen basierend auf verschiedenen Korrelationen ρ. Jede Punktwolke besteht aus 500 Wertpaare. (Siehe R-Code 7.2.)

mit Dichte

$$f_{\mathbf{X}}(x_1,\ldots,x_p) = f_{\mathbf{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big) \qquad (7.15)$$

für alle $\boldsymbol{x} \in \mathbb{R}^p$ ($\boldsymbol{\mu} \in \mathbb{R}^p$ und $\boldsymbol{\Sigma}$ symmetrisch und positiv definit). Wir bezeichnen diese Verteilung mit $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Die eindimensionalen Randverteilungen sind $X_i \sim \mathcal{N}\big((\boldsymbol{\mu})_i, (\boldsymbol{\Sigma})_{ii}\big)$, $i = 1,\ldots,p$.

Eigenschaften 7.3. *Für die multivariate Normalverteilung gilt:*

$$\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}\,, \qquad \mathrm{Var}(\mathbf{X}) = \boldsymbol{\Sigma}\,. \qquad (7.16)$$

Eigenschaften 7.4. *Sei* $\boldsymbol{a} \in \mathbb{R}^q$, $\mathbf{B} \in \mathbb{R}^{q\times p}$, $q \le p$, $\mathrm{rang}(\mathbf{B}) = q$ *und* $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *dann*

$$\boldsymbol{a} + \mathbf{B}\mathbf{X} \sim \mathcal{N}_q\big(\boldsymbol{a} + \mathbf{B}\mu, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^{\mathsf{T}}\big). \qquad (7.17)$$

Sei $\mathbf{I} \in \mathbb{R}^{p\times p}$ die *Einheitsmatrix* (Identitätsmatrix, *identity matrix*), d.h. eine quadratische Matrix, deren Hauptdiagonale nur aus Einsen besteht und die sonst nur Nullen enthählt, und sei $\mathbf{L} \in \mathbb{R}^{p\times p}$ so dass $\mathbf{L}\mathbf{L}^{\mathsf{T}} = \boldsymbol{\Sigma}$. Um p-variate Zufallszahlen mit $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ zu ziehen, beginnt man mit $Z_1,\ldots,Z_p \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, und setzt $\boldsymbol{z} = (z_1,\ldots,z_p)^{\mathsf{T}}$. Die Werte

werden nun mit $\boldsymbol{\mu} + \mathbf{L}\boldsymbol{z}$ transformiert. Da $\mathbf{Z} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ ist $\mathbf{X} = \boldsymbol{\mu} + \mathbf{L}\mathbf{Z} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{L}\mathbf{L}^\mathsf{T})$, vergleiche Eigenschaften 2.6.

In der Praxis wird oft die Cholesky-Zerlegung von $\boldsymbol{\Sigma}$ gebraucht. Die Cholesky-Zerlegung ist für symmetrische positiv definite Matrizen eindeutig und $\mathbf{L}$ ist eine untere Dreiecksmatrix. Zudem gilt $\det(\boldsymbol{\Sigma}) = \det(\mathbf{L})^2 = \prod_i (\mathbf{L})_{ii}^2$.

7.4 Bedingte Verteilungen

Wir betrachten Eigenschaften von "Teilen" aus dem Zufallsvektor $\mathbf{X}$. Der Einfachheithalber ordnen wir diese und schreiben

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \mathbf{X}_1 \in \mathbb{R}^q, \quad \mathbf{X}_2 \in \mathbb{R}^{p-q}. \tag{7.18}$$

Entsprechend unterteilen wir die Matrix $\boldsymbol{\Sigma}$ in 2×2 Blöcke:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim \mathcal{N}_p \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right) \tag{7.19}$$

Die beiden (multivariaten) Randverteilungen $\mathbf{X}_1$ und $\mathbf{X}_2$ sind wieder normalverteilt mit $\mathbf{X}_1 \sim \mathcal{N}_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ und $\mathbf{X}_2 \sim \mathcal{N}_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

$\mathbf{X}_1$ und $\mathbf{X}_2$ sind unabhängig wenn $\boldsymbol{\Sigma}_{21} = \mathbf{0}$ und umgekehrt.

Bedingt man einen multivariat normalverteilten Zufallsvektor auf einen Teilvektor, so ist das Ergebnis selbst wieder multivariat normalverteilt mit

$$\mathbf{X}_2 \mid \mathbf{X}_1 = \boldsymbol{x}_1 \sim \mathcal{N}_{p-q} \left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \right). \tag{7.20}$$

Der Erwartungswert hängt linear vom Wert von $\boldsymbol{x}_1$ ab, aber die Varianz ist unabhängig vom Wert von $\boldsymbol{x}_1$. Der bedingte Erwartungswert stellt eine Aktualisierung ("Prognose") von $\mathbf{X}_2$ durch $\mathbf{X}_1 = \boldsymbol{x}_1$ dar: die Differenz $\boldsymbol{x}_1 - \boldsymbol{\mu}_1$ wird durch die Varianz normalisiert und durch die Kovarianz skaliert. Für $p = 2$ ist $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} = \rho\sigma_y/\sigma_x$.

7.5 Schätzen

Die Schätzfunktionen werden ähnlich wie im univariaten Fall konstruiert. Sei $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ eine Realisation der Zufallsstichprobe $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Wir haben die folgenden Schätzfunktionen

$$\widehat{\boldsymbol{\mu}} = \overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \qquad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\mathsf{T} \tag{7.21}$$

und Schätzwerte

$$\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \qquad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\mathsf{T}. \tag{7.22}$$

Eine normalverteilte Zufallsvariable ist durch 2 Parameter bestimmt. Eine multivariate normalverteilte Zufallsvariable ist durch p (für μ) und $p(p+1)/2$ (für $\mathbf{\Sigma}$) Parameter bestimmt. Die restlichen $p(p-1)/2$ Parameter in $\mathbf{\Sigma}$ sind durch die Symmetrie bestimmt. Die $p(p+1)/2$ Werte können aber nicht beliebig gewählt werden, da $\mathbf{\Sigma}$ positiv definit sein muss (im univariaten Fall muss $\sigma > 0$ erfüllt sein). Solange $n \geq p$, erfüllt der Schätzwert in (7.22) diese Bedingungen. Wenn $p < n$ braucht es zusätzliche Annahmen üeber die Struktur der Matrix $\mathbf{\Sigma}$ (wie wir im Kapitel 9 sehen werden).

Beispiel 7.3. Der R-Code 7.3 schätzt den Erwartungswert und die Varianzmatrix/Kovarianz der Realisation mit $\rho = 0.9$ in der Abbildung 7.2. Da $n = 500$ sind die Werte nahe am "wahren" Wert.

Die Funktion `ellipse` aus dem Paket `ellipse` zeichnet Konfidenzregionen aus den geschätzten Parameter. Der R-Code 7.4 und die Abbildung 7.3 zeigen die geschätzten 95% und 50% Konfidenzregionen. Mit grösser werdenden n verbessert sich die Schätzung. ♣

R-Code 7.3 Schätzwerte für die Realisation $\rho = 0.9$ der Abbildung 7.2.

```
colMeans(sample)
## [1]  0.002734424 -0.013690699
# apply( sample, 2, mean) # is identical
var(sample)
##             [,1]      [,2]
## [1,] 1.0177321 0.9170572
## [2,] 0.9170572 0.9952524
# cov( sample)  # is identical
```

Anhang: Positiv definite Matrizen

Eine sehr gute Zusammenfassung von wichtigen Formeln rund um Vektoren und Matrizen ist das Onlinebuch "Matrix Cookbook", (Petersen and Pedersen, 2008).

Eine $n \times n$ Matrix $\mathbf{A}$ ist positiv definit (pd) wenn

$$\boldsymbol{x}^\mathsf{T}\mathbf{A}\boldsymbol{x} > 0, \qquad \forall \boldsymbol{x} \neq \mathbf{0}.$$

Falls $\mathbf{A} = \mathbf{A}^\mathsf{T}$, ist die Matrix symmetrisch positiv definit (spd). Einige wichtige Eigenschaften von spd Matrizen $\mathbf{A} = (a_{ij})$ sind:

i) $\text{rang}(\mathbf{A}) = n$

ii) $\det(\mathbf{A}) > 0$

R-Code 7.4 Bivariat normalverteilte Zufallszahlen für verschiedene Stichprobengrössen mit Höhenlinien der Dichte und geschätzte Momente. (Siehe Abbildung 7.3.)

```
require( ellipse)
n <- c(10, 100, 500, 1000)
mu <- c(2,1)
Sigma <- matrix( c(4,2,2,2), 2)

for (i in 1:4) {
  plot(ellipse( Sigma, cent=mu, level=.95), col='gray',
       xaxs='i', yaxs='i', xlim=c(-4,8),ylim=c(-4,6), type='l')
  lines(ellipse( Sigma, cent=mu, level=.5), col='gray')
  sample <- rmvnorm( n[i], mean=mu, sigma=Sigma)
  points(sample, pch='.', cex=2)
  Sigmahat <- cov( sample)
  muhat <- apply( sample, 2, mean)
  lines(ellipse( Sigmahat, cent=muhat, level=.95), col=2, lwd=2)
  lines(ellipse( Sigmahat, cent=muhat, level=.5), col=4, lwd=2)
  points( rbind(muhat), col=3, cex=2)
  text(-2,4, paste('n =',n[i]))
}
```

iii) Alle Eigenwerte sind positiv, $\lambda_i > 0$

iv) $a_{ii} > 0$

v) $a_{ii}a_{jj} - a_{ij}^2 > 0,\ i \neq j$

vi) $a_{ii} + a_{jj} - 2|a_{ij}| > 0,\ i \neq j$

vii) $\mathbf{A}^{-1}$ ist spd

viii) Alle Hauptuntermatrizen von $\mathbf{A}$ sind spd.

ix) Es existiert eine nicht singuläre untere Dreiecksmatrix $\mathbf{L}$, so dass $\mathbf{A} = \mathbf{L}\mathbf{L}^\mathsf{T}$.

Wenn wir eine nicht singuläre Matrix $\mathbf{A}$ als 2×2 Blockmatrix schreiben, gilt

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{C}_1^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{C}_2^{-1} \\ -\mathbf{C}_2\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{C}_2^{-1} \end{pmatrix}$$

mit

$$\mathbf{C}_1 = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}, \qquad \mathbf{C}_2 = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$$

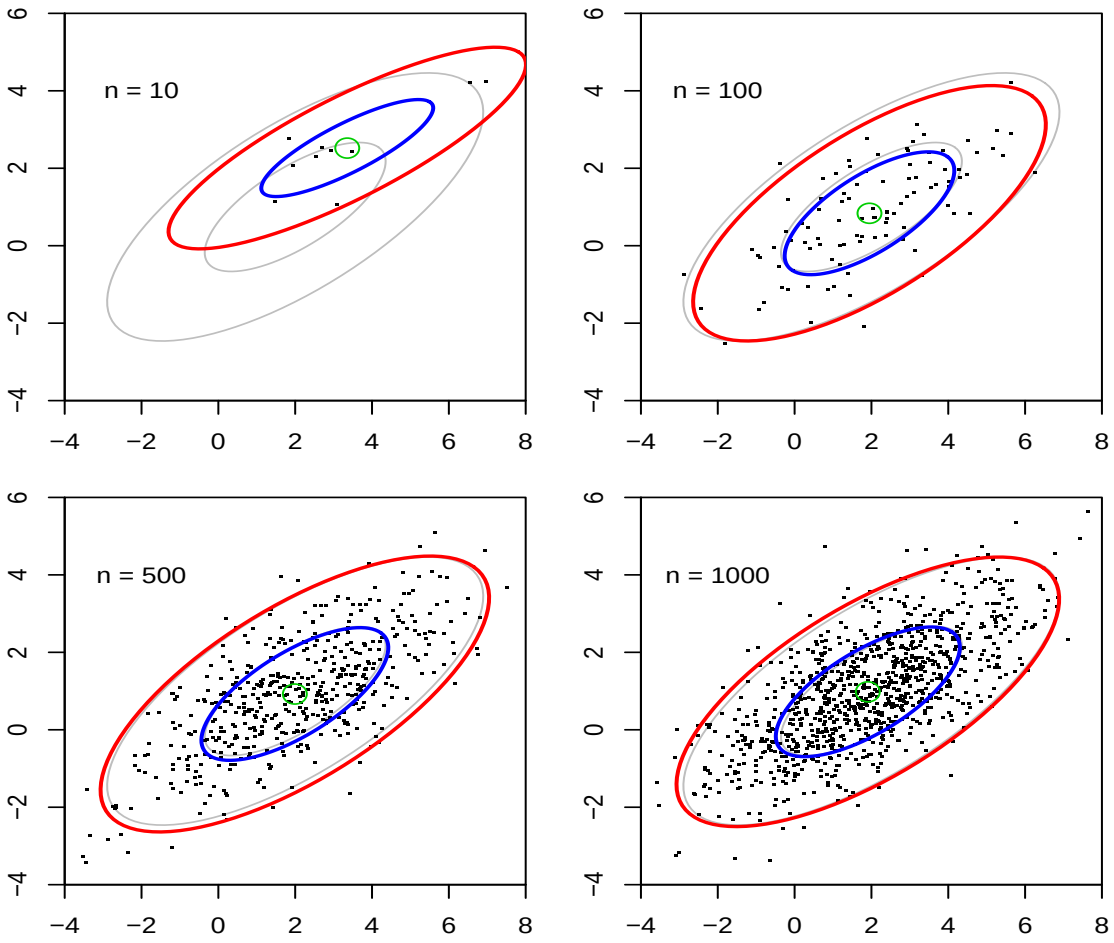


Abbildung 7.3: Bivariat normalverteite Zufallszahlen. Die Höhenlinien der Dichte sind in grau, die entsprechenden geschätzten 95% (50%) Konfidenzregion in rot (blau) und die arithmetischen Mittel in grün (geschätzte Momente). (Siehe R-Code 7.4.)

und

$$\mathbf{C}_1^{-1} = \mathbf{A}_{11}^{-1} - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}, \qquad \mathbf{C}_2^{-1} = \mathbf{A}_{22}^{-1} - \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}.$$

Es gilt auch $\det(\mathbf{A}) = \det(\mathbf{A}_{11})\det(\mathbf{C}_2) = \det(\mathbf{A}_{22})\det(\mathbf{C}_1)$.

Kapitel 8

Regression

In diesem und den nächsten drei Kapitel werden wir die sogenannten "linearen Modellen" betrachten. Eine ausführliche Diskussion von linearen Modellen würde ein kompletes Vorlesungsmodul füllen und wir können hier nur die allerwichtigsten Elemente betrachten. Das Buch von Fahrmeir *et al.* (2009) ist empfehlenswert, da ausführlich und zugänglich.

Die (einfache) Regression wird häufig als archetypische Aufgabe der Statistik betrachtet, und wird oft schon auf dem Mittelschulniveau eingeführt.

In diesem Kapitel werden (i) der (lineare) Zusammenhang zwischen zwei Variablen quantifiziert, (ii) eine Variable durch eine andere Variable mit Hilfe eines "Modells" erklärt und (iii) einige weiterführende Hinweise gegeben.

8.1 Korrelation

Unser Ziel ist mit Hilfe von n Wertepaaren $(x_1, y_n), \ldots, (x_n, y_n)$, betrachtet als Realisationen zweier Zufallsvariablen (X, Y), den linearen Zusammenhang zwischen X und Y zu quantifizieren.

Ein intuitiver Schätzer der Korrelation zwischen X und Y (siehe Gleichung (7.6)) ist

$$r = \widehat{\mathrm{Corr}(X,Y)} = \frac{\widehat{\mathrm{Cov}(X,Y)}}{\sqrt{\widehat{\mathrm{Var}(X)}\widehat{\mathrm{Var}(Y)}}} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_j (y_j - \overline{y})^2}}, \tag{8.1}$$

welcher auch Pearson-Korrelationskoeffizient (*Pearson correlation coefficient*) genannt wird. Wie die Korrelation liegt auch der Pearson-Korrelationskoeffizient im Intervall $[-1, 1]$.

Beispiel 8.1. R-Code 8.1 schätzt die Korrelationen der Punktwolken der Abbildung 7.2. Obwohl viele Datenpaare vorliegen ($n = 500$) ist im Falle von $\rho = 0.1$ der Schätzwert viel zu gross.

R-Code 8.1 *binorm* Daten: Pearson-Korrelationskoeffizient der Punktwolken der Abbildung 7.2.

```
require( mvtnorm)
set.seed(12)
rho <- c(-.25, 0, .1, .25, .75, .9)
rhohat <- numeric(6)
for (i in 1:6) {
  Sigma <- array( c(1, rho[i], rho[i], 1), c(2,2))
  sample <- rmvnorm( 500, sigma=Sigma)
  rhohat[i] <- cor( sample)[2]
}
print(rbind(rho, rhohat), digits=3)
```

```
##          [,1]   [,2]  [,3] [,4]  [,5]  [,6]
## rho    -0.250 0.0000 0.100 0.25 0.750 0.900
## rhohat -0.217 0.0482 0.221 0.28 0.778 0.911
```

```
##           [,1]  [,2] [,3] [,4] [,5] [,6]
## rho      -0.25 0.000 0.10 0.25 0.75 0.90
## Pearson  -0.22 0.048 0.22 0.28 0.78 0.91
## Spearman -0.22 0.066 0.18 0.26 0.77 0.91
## Kendall  -0.15 0.045 0.12 0.18 0.57 0.74
```

Die Korrelation ist ein Mass für den linearen Zusammenhang; zudem ist der Korrelationskoeffizient nicht robust, siehe R-Code 8.2 und Abbildung 8.1.

Rangkorrelationskoeffizienten sind "robuste" oder nicht-parametrische Korrelationsschätzer, Beispiele sind Spearmans ρ oder Kendalls τ. Spearmans ρ wird ähnlich zu (8.1) berechnet, wobei die Werte durch deren Ränge ersetzt werden. Kendalls τ vergleicht die Zahl der konkordanten (falls $x_i < x_j$, dann $y_i < y_j$) und der diskonkordanten Paare (falls $x_i < x_j$, dann $y_i > y_j$). R-Code 8.2 vergleicht die verschiedenen Korrelationsschätzer der *anscombe* Daten der Abbildung 8.1.

R-Code 8.2: *anscombe* Daten: Visualisierung und Korrelationsschätzwerte. (See Figure 8.1.)

```
library( faraway)
data( anscombe)
head( anscombe, 3)
```

```
##   x1 x2 x3 x4   y1   y2    y3   y4
```

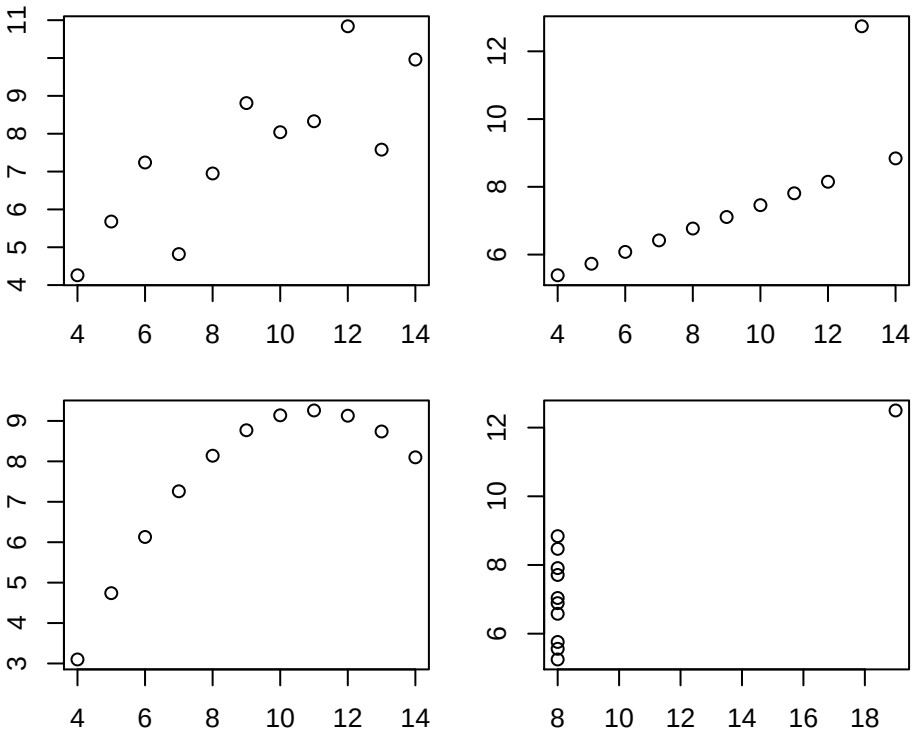


Abbildung 8.1: *anscombe* Daten, die vier Fälle haben einen Pearson-Korrelationskoeffizient von 0.82. (Siehe R-Code 8.2.)

```
## 1 10 10 10  8 8.04 9.14  7.46 6.58
## 2  8  8  8  8 6.95 8.14  6.77 5.76
## 3 13 13 13  8 7.58 8.74 12.74 7.71
with( anscombe,
     c(cor(x1, y1), cor(x2, y2), cor(x3, y3), cor(x4, y4)))
## [1] 0.8164205 0.8162365 0.8162867 0.8165214
with( anscombe, { plot(x1, y1); plot(x2, y2); plot(x3, y3);
                  plot(x4, y4) })
sel <- c(0:3*9+5)  # extract diagonal entries of sub-block
print(rbind( pearson=cor(anscombe)[sel],
         spearman=cor(anscombe, method='spearman')[sel],
         kendall=cor(anscombe, method='kendall')[sel]), digits=2)
##          [,1] [,2] [,3] [,4]
## pearson  0.82 0.82 0.82 0.82
## spearman 0.82 0.69 0.99 0.50
## kendall  0.64 0.56 0.96 0.43
```

Bei bivarat normalverteilten Zufallsvariablen ist r ein Schätzer des Korrelationsparameters ρ. Sei R die entsprechende Schätzfunktion von ρ basierend auf (8.1). Die Zufalls-

variable

$$T = R\frac{\sqrt{n-2}}{\sqrt{1-R^2}} \tag{8.2}$$

ist unter $H_0 : \rho = 0$ t-verteilt mit $n-2$ Freiheitsgraden. Der entsprechende Test ist unter Test 11 beschrieben.

Test 11: Test von Korrelationen

Fragestellung: Ist der aus n Wertepaaren ermittelte Korrelationskoeffizient r signifikant?

Voraussetzungen: Die Wertepaare stammen aus einer bivariaten Normalverteilung.

Berechnung: $t_{\text{Vers}} = |r|\frac{\sqrt{n-2}}{\sqrt{1-r^2}}$
wobei r der Pearson-Korrelationskoeffizient (8.1) ist.

Entscheidung: Vergleiche t_{Vers} mit $t_{\text{Tab}}(n-2;\alpha/2)$: verwerfe H_0: $\rho = 0$ wenn $t_{\text{Vers}} > t_{\text{Tab}}$.

Berechnung in R: `cor.test( x, y, conf.level=1-alpha)`

Um Konfidenzintervalle von Korrelationsschätzwerten zu konstruieren, brauchen wir typischerweise die sogenannte Fisher-Transformation (*Fisher transformation*)

$$W(r) = \frac{1}{2}\log\frac{1+r}{1-r} = \operatorname{arctanh}(r) \tag{8.3}$$

und die Tatsache, dass für bivarat normalverteilte Zufallsvariablen die Verteilung von $W(R)$ ungefähr $\mathcal{N}(W(\rho), 1/(n-3))$ ist.

Es braucht erstaunlich grosse Stichproben damit Korrelationsschätzwerte um .25 signifikant sind.

8.2 Einfache Regression

In der einfachen Regression wird eine (abhängige) Variable durch eine einzige unabhängige/freie Variable linear erklärt:

$$Y_i = \mu_i + \varepsilon_i \tag{8.4}$$

$$= \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n \tag{8.5}$$

mit

Konfidenzintervall für den Pearson-Korrelationskoeffizient

Ein approximatives $(1-\alpha)$-Konfidenzintervall für r ist

$$\left[\tanh\left(\operatorname{arctanh}(r) - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right), \tanh\left(\operatorname{arctanh}(r) + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right)\right]$$

wobei tanh und arctanh der Tangens Hyperbolicus und Areatangens Hyperbolicus sind.

- Y_i: abhängige Variable, Messwert, Beobachtung
- x_i: unabhängige/freie Variable, Prädiktor
- β_0, β_1: Parameter (unbekannt)
- ε_i: Messfehler, Fehler, mit symmetrischer Verteilung und $\mathrm{E}(\varepsilon_i) = 0$ (unbekannt).

Oft wird auch angenommen, dass $\mathrm{Var}(\varepsilon_i) = \sigma^2$ und/oder dass die Fehler unabhängig sind. Einfachheitshalber nehmen wir $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ mit σ^2 unbekannt an. Somit ist $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1, \ldots, n$ und Y_i und Y_j unabhängig $i \neq j$.

Beispiel 8.2. (*hardness* Daten) In der Herstellung von Metall Federn ist einer der Produktionsschritte ein Abschreckbad. Die Temperatur des Bades hat einen Einfluss auf die Härte der Federn. Abbildung 8.2 zeigt die Rockwell Härte von Spiralfedern als Funktion der Temperatur des Abschreckbades sowie die "beste" Gerade. Rockwell ist eine Masseinheit für die Härte technischer Werkstoffe und wird mit HR (*Hardness Rockwell*) bezeichnet. Der R-Code 8.3 zeigt, wie eine einfache Regression für die entsprechenden Daten durchgeführt werden. ♣

R-Code 8.3 *hardness* Daten vom Beispiel 8.2, siehe Abbilding 8.2.

```
Temp <- rep( 10*3:6, c(4,3,3,4))
Hard <- c(55.8, 59.1, 54.8, 54.6, 43.1, 42.2, 45.2,
          31.6, 30.9, 30.8, 17.5, 20.5, 17.2, 16.9)
plot( Temp, Hard,         # scatter plot of the data
     xlab="Temperature [C]", ylab="Hardness [HR]")
lm1 <- lm( Hard~Temp)     #  fit a linear model
abline( lm1)
```

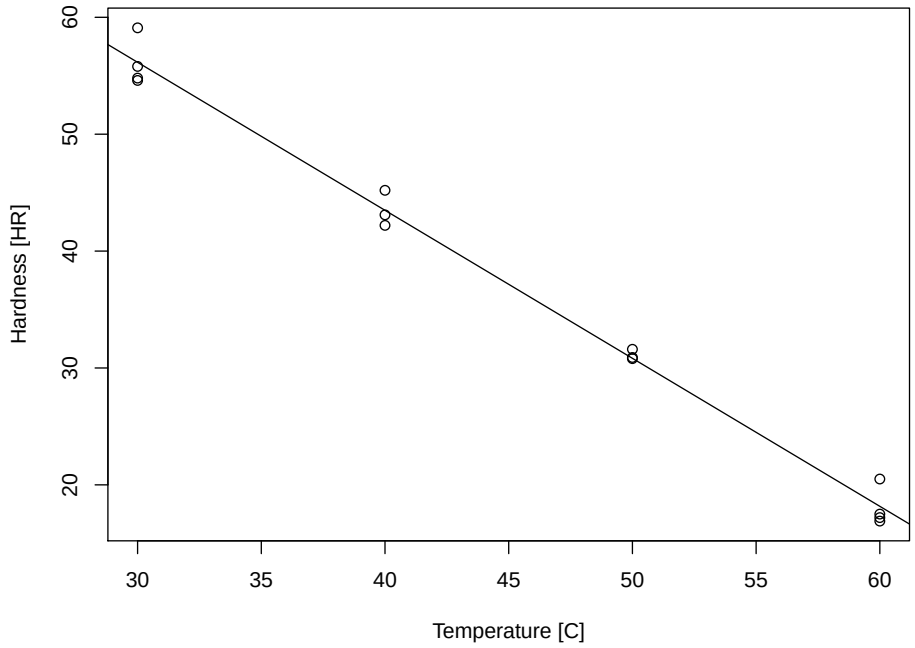


Abbildung 8.2: *hardness* Daten: Härte als Funktion der Temperatur (siehe R-Codes 8.3 und 8.4).

Die Idee der Regression beruht auf Schätzern $\widehat{\beta}_i$, die den Quadratsummenfehler minimieren. Das heisst, $\widehat{\beta}_0$ und $\widehat{\beta}_1$ werden so bestimmt, dass

$$\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \tag{8.6}$$

minimal ist. Dieses Prinzip wird auch *Kleinste-Quadrate Methode* (*least squares method*) genannt. Die Lösung ist

$$\widehat{\beta}_1 = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{8.7}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \tag{8.8}$$

Somit ist die geschätzte Regressionsgerade

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x \tag{8.9}$$

In der einfachen Regression können von (8.5) aus die entsprechenden Schätzer direkt hergeleitet werden. Für die multiple Regression, dass heisst für Modelle mit mehr als einer freien Variable, ist es jedoch besser direkt mit Matrizenschreibweise zu beginnen und die Schätzer für eine beliebige Anzahl Prädiktoren zu bestimmen (siehe Kapitel 10).

R-Code 8.4 *hardness* Daten vom Beispiel 8.2, siehe Abbilding 8.2.

```
summary( lm1)            # summary of the fit
##
## Call:
## lm(formula = Hard ~ Temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5494 -1.1898 -0.3687  0.5986  2.9505
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 94.13407    1.57501   59.77 3.18e-16 ***
## Temp        -1.26615    0.03386  -37.40 8.58e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.495 on 12 degrees of freedom
## Multiple R-squared:  0.9915,Adjusted R-squared:  0.9908
## F-statistic:  1399 on 1 and 12 DF,  p-value: 8.578e-14
coef( lm1)
## (Intercept)        Temp
##   94.134066   -1.266154
head( fitted( lm1))
##        1        2        3        4        5        6
## 56.14945 56.14945 56.14945 56.14945 43.48791 43.48791
head( resid( lm1))
##          1          2          3          4          5          6
## -0.3494505  2.9505495 -1.3494505 -1.5494505 -0.3879121 -1.2879121
head( Hard - (fitted( lm1) + resid( lm1)))
## 1 2 3 4 5 6
## 0 0 0 0 0 0
```

In der einfachen Regression ist es oft zentral zu wissen, ob ein linearer Zusammenhang zwischen den festen und freien Variable besteht. Dies kann mit der Hypothese $H_0 : \beta_1 = 0$ getestet werden (Test 12).

Test 12: Test eines linearen Zusammenhangs

Fragestellung: Gibt es einen linearen Zusammenhang zwischen zwei Stichproben?

Voraussetzungen: Basierend auf der Stichprobe $x_1, \ldots, x_n$, die zweite Stichprobe ist normalverteilt mit Erwartungswert $\beta_0 + \beta_1 x_i$ und Varianz σ^2.

Berechnung: $t_{\text{Vers}} = |r| \frac{s_y}{s_x}$

Entscheidung: Vergleiche t_{Vers} mit $t_{\text{Tab}}(n-2;\alpha)$: verwerfe H_0: $\beta_1 = 0$ wenn $t_{\text{Vers}} > t_{\text{Tab}}$.

Berechnung in R: `summary( lm( y ~ x ))`

Zur Prognose der abhängigen Variablen bei gegebener freien Variable x_0 wird der Wert x_0 in (8.9) benützt. In R kann die "Prognose"-Methode `predict` verwendet werde. Die Prognoseunsicherheit hängt von den Unsicherheiten der geschätzten Parameter (und von der Varianz des Fehlers) ab. Spezifisch:

$$\text{Var}(\widehat{\mu}_0) = \text{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) \tag{8.10}$$

Jedoch sind im Allgemeinen $\widehat{\beta}_0$ und $\widehat{\beta}_1$ nicht unabhängig und die Varianz ist nicht ganz einfach zu berechnen. Wir kommen im Kapitel 10 darauf zurück. Um Konfidenzintervalle einer Prognose zu konstruieren, muss unterschieden werden, ob dieses für $\widehat{y}_0$ oder für $\widehat{\mu}_0$ ist.

Die Prognose kann auch als

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_0 = \overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 \overline{x}_0 = \overline{y} + s_{xy}(s_x^2)^{-1}(x_0 - \overline{x}) \tag{8.11}$$

geschrieben werden. Diese Gleichung entspricht (7.20) mit den geschätzten Werten anstelle den (unbekannten) Parametern.

Die lineare Regression setzt Voraussetzungen bezüglich der Fehler ε_i voraus. Nach einer Berechnung muss überprüft werden, wie realistisch diese sind. Dazu werden die Residuen und deren Eigenschaften betrachtet. Wir kommen im Kapitel 10 darauf zurück.

R-Code 8.5 *hardness* Daten: Prognose (Prädiktion, *prediction*) und punktweise Konfidenzintervalle. (Siehe Abbildung 8.3.)

```
new <- data.frame( Temp = seq(15, 75, by=.5))
pred.w.clim <- predict( lm1, new, interval="confidence")
 # for hat( mu )
pred.w.plim <- predict( lm1, new, interval="prediction")
 # for hat( y ), hence wider!
plot( Temp, Hard,
     xlab="Temperature [C]", ylab="Hardness [HR]")

matlines( new$Temp, cbind(pred.w.clim, pred.w.plim[,-1]),
         col=c(1,2,2,3,3),lty=c(1,1,1,2,2))
```

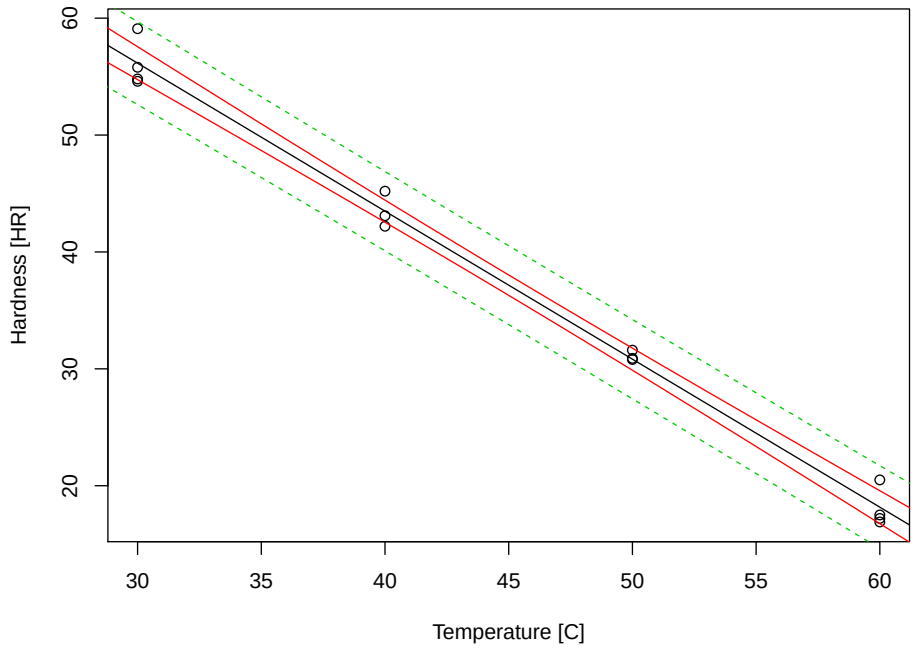


Abbildung 8.3: *hardness* Daten: Härte als Funktion der Temperatur. (Siehe R-Code 8.5.)

8.3 Logistische Regression

Beispiel 8.3. ***orings* Daten** Im Januar 1986 explodierte das Space Shuttle Challenger kurz nach dem Start. Teil des Problems waren die Gummidichtungen, die sogenannten *O-rings*, der Boosterraketen. Der Datensatz `data( orings, package=faraway")` enthält die Anzahl Defekte der sechs Dichtungen in den 23 vorangegangenen Starts (Abbildung 8.4). Die Frage, die wir uns hier stellen, ist, ob für eine Lufttemperatur von 31°F (wie im Januar 1986) die Wahrscheinlichkeit eines Defektes einer beliebigen Dichtungen prognostiziert werden kann. Siehe Dalal *et al.* (1989) für eine detaillierte Darstellung. ♣

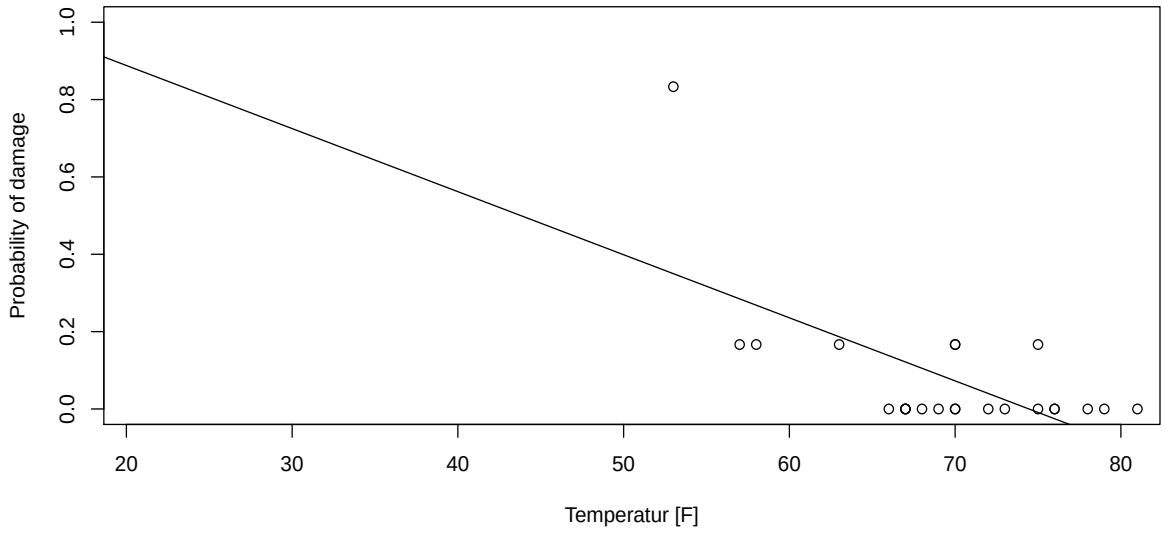


Abbildung 8.4: *orings* Daten und linearer Fit mit einfacher Regression.

Das Beispiel zeigt, dass lineare Modelle oft nicht adequat sind, da hier die Zielvariable eine Wahrscheinlichkeit ist, das heisst $y_i \in [0,1]$, ein lineares Modell kann aber $\widehat{y}_i \in [0,1]$ nicht garantieren. In diesem und ähnlichen Fällen ist eine logistische Regression angebracht, die die Wahrscheinlichkeit eines Defekts mit

$$p = \mathrm{P}(\text{Defekt}) = \frac{1}{1+\exp(-\beta_0 - \beta_1 x)},$$

modelliert, wobei x die Lufttemperatur ist. Durch Umkehrung erhält man ein lineares Modell für die log-odds

$$g(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

wobei $g(\cdot)$ im allgemeinen die Linkfunktion (*link function*) genannt wird. Im speziellen Fall hier heisst die Funktion $g^{-1}(\cdot)$ logistische Funktion.

Die logistische Regression ist ein Spezialfall eines generalisierten linearen Modells (*generalized linear model*).

R-Code 8.6 *orings* Daten und geschätzte Bruchwahrscheinlichkeiten in Abhängigkeit der Lufttemperatur (siehe Abbildung 8.5).

```
data( orings, package="faraway")
glm1 <- glm( cbind(damage,6-damage)~temp, family=binomial, data=orings)
plot( damage/6~temp, xlim=c(21,80), ylim=c(0,1), data=orings,
     xlab="Temperatur [F]", ylab='Probability of damage')
points( orings$temp, glm1$fitted, col=2)

ct <- seq(20, to=85, length=100)
p.out <- predict( glm1, new=data.frame(temp=ct), type="response")
lines(ct, p.out)
abline( v=31, col='gray', lty=2)
predict( glm1, new=data.frame(temp=31), type="response")
```

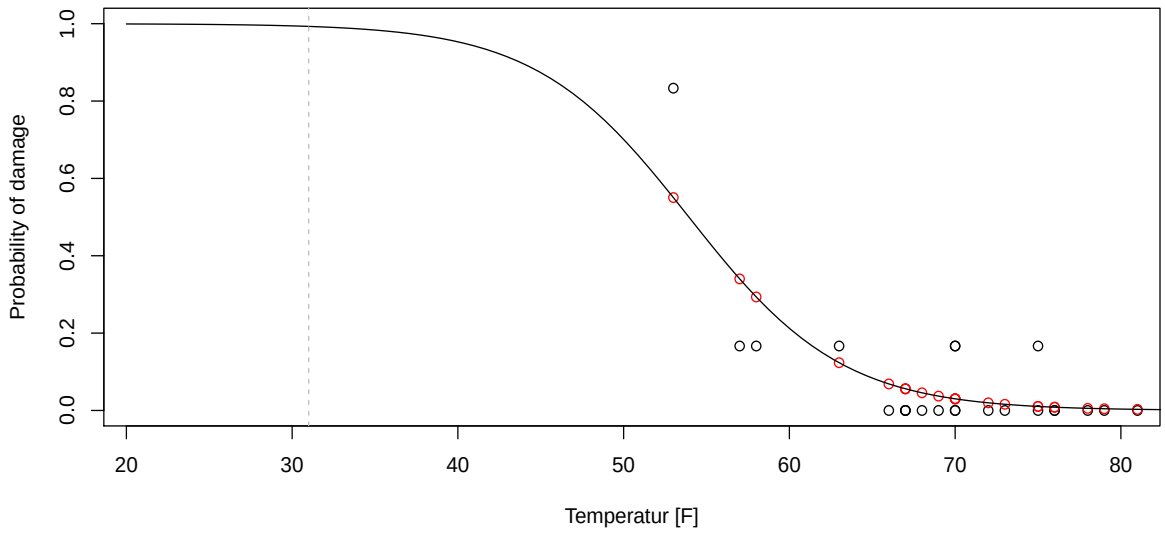


Abbildung 8.5: *orings* Daten und geschätzte Bruchwahrscheinlichkeiten in Abhängigkeit der Lufttemperatur (siehe R-Code 8.6).

Kapitel 9

Multiple Regression

Das (einfache) Regressionsmodell wird erweitert, indem wir mehrere freie Variablen zulassen. Wir werden zuerst Modell und Schätzer einführen. Anschliessend werden wir die wichtigsten Schritte der Modellvalidierung kennen lernen. Einige typische Beispiele von multipler Regression sind am Ende aufgezeigt.

9.1 Modell und Schätzer

Gleichung (8.5) wird auf p unabhängige Variablen verallgemeinert

$$\begin{aligned} Y_i &= \mu_i + \varepsilon_i, && (9.1)\\ &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, && (9.2)\\ &= \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i \qquad i = 1, \ldots, n, \quad n > p && (9.3) \end{aligned}$$

mit

- Y_i: abhängige Variable, Messwert, Beobachtung
- $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\mathsf{T}$: unabhängige/freie Variablen, Prädiktoren
- $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\mathsf{T}$: Parametervektor (unbekannt)
- ε_i: Messfehler, Fehler, mit symmetrischer Verteilung und $\mathrm{E}(\varepsilon_i) = 0$ (unbekannt).

Oft wird auch angenommen, dass $\mathrm{Var}(\varepsilon_i) = \sigma^2$ und/oder dass die Fehler unabhängig sind. Der Einfachheithalber nehmen wir an, $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, mit σ^2 unbekannt. In Matrizenschreibweise schreibt sich (9.3)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{9.4}$$

mit $\mathbf{X}$ einer $n \times (p+1)$ Matrix mit Zeilen $\boldsymbol{x}_i$. Wir nehmen an, dass der Rang von $\mathbf{X}$ gleich $p+1$ ist ($\mathrm{rang}(\mathbf{X}) = p+1$, Spaltenrang).

Das Kleinste Quadrate Prinzip wird somit folgendermassen angewandt:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \tag{9.5}$$

$$\Rightarrow \quad \frac{d}{d\boldsymbol{\beta}}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \tag{9.6}$$

$$= \frac{d}{d\boldsymbol{\beta}}(\boldsymbol{y}^\mathsf{T}\boldsymbol{y} - 2\boldsymbol{\beta}^\mathsf{T}\mathbf{X}^\mathsf{T}\boldsymbol{y} + \boldsymbol{\beta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}) = -2\mathbf{X}^\mathsf{T}\boldsymbol{y} + 2\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} \tag{9.7}$$

$$\Rightarrow \quad \mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\mathsf{T}\boldsymbol{y} \tag{9.8}$$

$$\Rightarrow \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y} \tag{9.9}$$

Gleichung (9.8) wird auch Normalengleichung genannt (*normal equation*).

Es kann gezeigt werden, dass

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \tag{9.10}$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y} \qquad \widehat{\boldsymbol{\beta}} \sim \mathcal{N}_{p+1}\big(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\big) \tag{9.11}$$

$$\widehat{\boldsymbol{y}} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y} = \mathbf{H}\boldsymbol{y} \qquad \widehat{\mathbf{Y}} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H}) \tag{9.12}$$

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = (\mathbf{I} - \mathbf{H})\boldsymbol{y} \qquad \boldsymbol{e} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})) \tag{9.13}$$

In der linken Spalte befinden sich die Schätzwerte, in der rechten die Stichprobenfunktionen. Aus der Verteilung (9.11) ist die Randverteilung der einzelnen Komponenten bestimmt:

$$\widehat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2 v_{ii}) \quad \text{mit } v_{ii} = \big((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\big)_{ii}, \quad i = 0, \ldots, p \tag{9.14}$$

Da σ^2 in der Regel unbekannt ist, brauchen wir

$$\frac{\widehat{\beta}_i - \beta_i}{\sqrt{\widehat{\sigma}^2 v_{ii}}} \sim t_{n-p-1} \quad \text{mit } \widehat{\sigma}^2 = \frac{1}{n-p-1}\boldsymbol{e}^\mathsf{T}\boldsymbol{e} \tag{9.15}$$

als Statistik zum Testen und zum Herleiten von Konfidenzintervallen.

Konfidenzintervall für Regressionskoeffizienten

Ein $(1-\alpha)$-Konfidenzintervall für β_i ist

$$\left[\, \widehat{\beta}_i \pm t_{n-p-1,1-\alpha/2}\sqrt{\frac{1}{n-p-1}\boldsymbol{e}^\mathsf{T}\boldsymbol{e}\; v_{ii}} \,\right]$$

mit $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ und $v_{ii} = \big((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\big)_{ii}$.

9.2 Modellvalidierung

Modellvalidierung überprüft (im wesentlichen), ob (9.3) ein adäquates Modell für die Daten zur Bearbeitung der vorliegenden Hypothese ist. Die Frage ist nicht, ob das Modell korrekt ist, sondern ob es nützlich ist ("Essentially, all models are wrong, but some are useful", Box and Draper, 1987 Seite 424).

Die Validierung überprüft (i) die feste Komponente μ_i und (ii) den stochastischen Teil ε_i und ist gewöhnlich ein iterativer Prozess.

Bezüglich der festen Komponente muss überprüft werden, ob zu viele oder alle notwendigen Prädiktoren im Modell enthalten sind. Nicht notwendige Prädiktoren werden oft durch nicht signifikante Koeffizienten identifiziert. Bei fehlenden Prädiktoren zeigen die Residuen (im Idealfall) Strukturen und lassen so auf Modellverbesserungen schliessen. In anderen Fallen ist die Qualität der Regression klein (F-Test, (zu) kleines R^2).

Es ist wichtig zu verstehen, dass der Fehler ε nicht notwendigerweise nur den Messfehler darstellt. Im allgemeinen ist der Fehler die restliche "Variabilität" (auch *noise*), was nicht durch die Prädiktoren "Signal" (*signal*) erklärt wird. Daher ist es möglich, dass die Fehler nicht iid normalverteilt sind.

Bezüglich der Fehler ε_i sollten folgende Punkte überprüft werden:

i) konstante Varianz: wenn die (absoluten) Residuen als Funktion von den Prädiktoren, den geschätzten Werten der Messsequenz aufgezeichnet werden, soll keine Struktur erkennbar sein. Oft können die Beobachtungen transformiert werden und dadurch eine konstante Varianz erreicht werden. Bei konstanter Varianz spricht man auch von Homoskedastizität (*homoscedasticity*), ansonsten von Heteroskedastizität (*heteroscedasticity*) oder Varianzheterogenität.

ii) unabhängig: Korrelationen zwischen den Residuen sollen vernachlässigbar sein. Diese Art von "Fehlerstruktur" wird im Kapitel 10 diskutiert.

iii) symmetrische Verteilung: es ist nicht einfach, Evidenz gegen diese Annahme zu finden. Wenn die Verteilung stark rechts- oder linksschief ist, sind Punktwolken der Residuen strukturiert. Eventuell helfen Transformationen oder generalisierte lineare Modelle.

Beispiel 9.1. Wir konstruieren wieder synthetische Daten um die Phänomene klarer darzustellen. In der Tabelle 9.1 sind die Grundmodelle und die verwenden (gefitteten) Modelle gegeben. In allen Fällen ist $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1/2)$. Die R-Codes 9.2 bis 9.5 und die entsprechenden Abbildungen illustrieren die Modellmängel. Da wir als Pädiktor `x1` eine steigende Sequenz wählen, sind Residuenplots in Abhängigkeit des Messindex überflüssig.

Da der Output von `summary(.)` recht lang ist, zeigen wir nur einzelne Elemente davon an. Das Darstellen erfolgt mit den Funktionen `print` und `cat`.

Tabelle 9.1: Grundmodelle und verwenden (gefitteten) Modelle

R-Code	Grundmodel	gefittetes Model
9.2	$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon_i$	$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$
9.3	$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$	$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$
9.4	$Y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$	$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$
9.4	$Y_i = (\beta_0 + \beta_1 x_1 + \varepsilon_i)^2$	$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon_i$

R-Code 9.1 Vordefinierte Werte für alle Beispiele.

```
n <- 60
eps <- rnorm(n,sd=1/16)
x1 <- seq(0, to=1, length=n)
x2 <- runif(n)
```

R-Code 9.2 Beispiel 1: Fehlender Prädiktor (Siehe Abbildung 9.1.)

```
y <- 2*x1+2*x1^2+eps
sres <- summary( res <- lm(y~x1))
print( sres$coef, digits=2)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.28      0.044    -6.4  3.1e-08
## x1              3.93      0.077    51.3  5.3e-50
cat("Adjusted R-squared: ", formatC(sres$adj.r.squared),
  "\nF-Test: ", pf(sres$fstatistic[1], 1, n-2, lower.tail = FALSE))
## Adjusted R-squared:  0.978
## F-Test:  5.28261e-50
plot(res$fitted, res$resid)
plot(x1, res$resid)
```

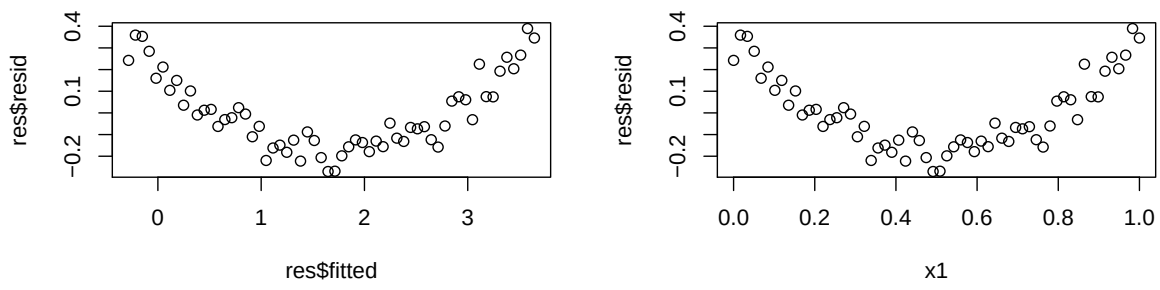


Abbildung 9.1: Residual-Plots (Siehe R-Code 9.2.)

R-Code 9.3 Beispiel 2: Fehlender Prädiktor (Siehe Abbildung 9.2.)

```
y <- 2*x1+2*x2+eps
sres <- summary( res <- lm(y~x1))
print( sres$coef, digits=2)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.97       0.15     6.6  1.5e-08
## x1              2.05       0.25     8.1  4.5e-11
cat("Adjusted R-squared: ", formatC(sres$adj.r.squared),
  "\nF-Test: ", pf(sres$fstatistic[1], 1, n-2, lower.tail = FALSE))
## Adjusted R-squared:  0.5213
## F-Test:  4.531258e-11
plot(res$fitted, res$resid)
plot(x1, sres$resid)
```

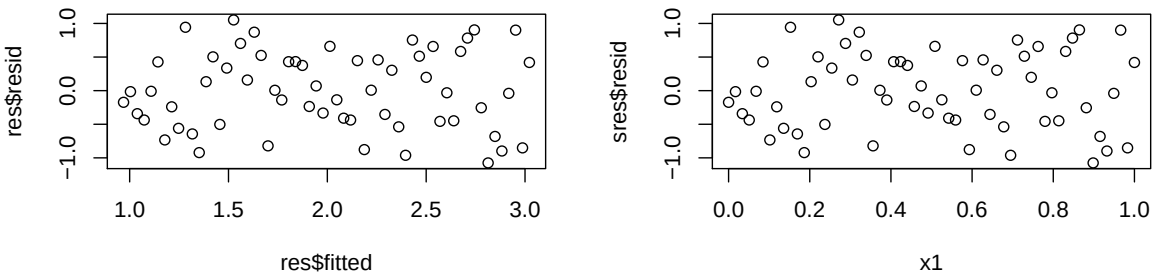


Abbildung 9.2: Residual-Plots (Siehe R-Code 9.3.)

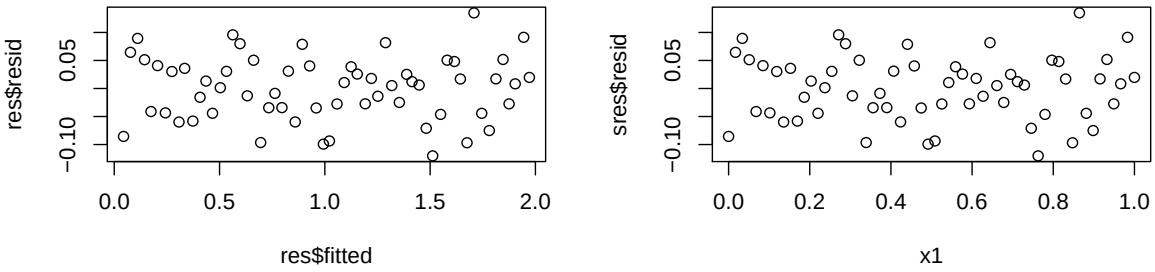


Abbildung 9.3: Residual-Plots (Siehe R-Code 9.4.)

R-Code 9.4 Beispiel 3: Zu viele Prädiktoren (Siehe Abbildung 9.3.)

```
y <- 2*x1+eps
sres <- summary( res <- lm(y~x1+x2))
print( sres$coef, digits=2)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0476      0.019    2.50  1.5e-02
## x1            1.9298      0.025   76.06  5.2e-59
## x2           -0.0082      0.026   -0.31  7.5e-01
cat("Adjusted R-squared: ", formatC(sres$adj.r.squared),
  "\nF-Test: ", pf(sres$fstatistic[1], 2, n-3, lower.tail = FALSE))
## Adjusted R-squared:  0.9899
## F-Test:  4.466334e-58
plot(res$fitted, res$resid)
plot(x1, sres$resid)
```

R-Code 9.5 Beispiel 4: Transformation erforderlich (Siehe Abbildung 9.4.)

```
y <- (2*x1+eps)^2
sres <- summary( res <- lm(y~x1+I(x1^2)))
print( sres$coef, digits=2)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.035      0.052    0.68  5.0e-01
## x1            -0.073      0.239   -0.30  7.6e-01
## I(x1^2)        3.992      0.231   17.25  2.7e-24
cat("Adjusted R-squared: ", formatC(sres$adj.r.squared),
  "\nF-Test: ", pf(sres$fstatistic[1], 2, n-3, lower.tail = FALSE))
## Adjusted R-squared:  0.9869
## F-Test:  7.526943e-55
plot(res$fitted, res$resid)
plot(x1, sres$resid)
```

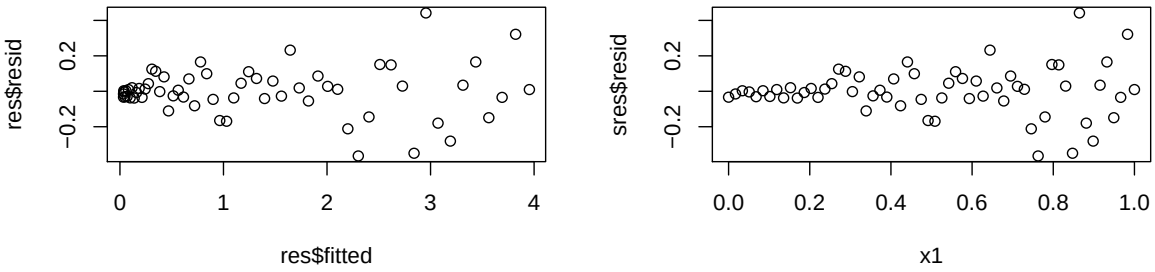


Abbildung 9.4: Residual-Plots (Siehe R-Code 9.5.)

9.3 Informationskriterien

Ein Informationskriterium ist ein Kriterium zur Auswahl eines Modells. Man folgt dabei der Idee von Ockhams Rasiermesser (*Occam's razor*), dass ein Modell nicht unnötig komplex sein soll, und gleicht die Anpassungsgüte des geschätzten Modells und dessen Komplexität aus. Die Komplexität wird mit der Anzahl der Parameter "strafend" berücksichtigt, da sonst komplexe Modelle mit vielen Parametern bevorzugt würden.

Um die Anpassungsgüte zu quantifizieren, nehmen wir an, dass die Verteilung der abhängigen Variable einer bekannten Verteilung mit einem unbekannten Parameter $\boldsymbol{\theta}$ folgt. Bei der Maximum-Likelihood-Schätzung ist die negative Loglikelihood-Funktion $-\ell(\widehat{\boldsymbol{\theta}})$ umso kleiner, je besser das Modell ist.

Das historisch älteste Kriterium wurde 1973 von Hirotugu Akaike als "an information criterion" vorgeschlagen und ist heute als Akaikes Informationskriterium (AIC *Akaike information criterion*) bekannt:

$$\text{AIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + 2p. \tag{9.16}$$

In Regressionsmodellen mit Gaussverteilten Fehlern ist der maximierte log-Likelihood linear zu $\log \widehat{\sigma}^2$ und somit beschreibt der erste Term die Anpassungsgüte.

Der Nachteil des AIC ist, dass der Strafterm von der Stichprobengrösse unabhängig ist. Bei grossen Stichproben sind Verbesserungen der Log-Likelihood bzw. der Residualvarianz "leichter" möglich, weshalb das Kriterium bei grossen Stichproben tendenziell Modelle mit verhältnismässig vielen Parametern vorteilhaft erscheinen lässt. Deshalb empfiehlt sich die Verwendung des Bayesschen Informationskriteriums (BIC, *Bayesian information criterion*):

$$\text{BIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + \log(n)p. \tag{9.17}$$

9.4 Beispiele

Beispiel 9.2. (*abrasion* Daten) Der Beschrieb des Datesatzes ist wie folgt:

> The data come from an experiment to investigate how the resistance of rubber to abrasion is affected by the hardness of the rubber and its tensile strength (Cleveland, 1993). Each of 30 samples of rubber was tested for hardness and for tensile strength, and then subjected to steady abrasion for a fixed time.

Die Konfidenzintervalle `confint( res)` enthalten die Null nicht. Enstprechend sind die p-Werte der drei t-Teste klein.

Ein quadratischer Term für `strength` ist nicht nötig. Einzig die Residuen scheinen korreliert zu sein. ♣

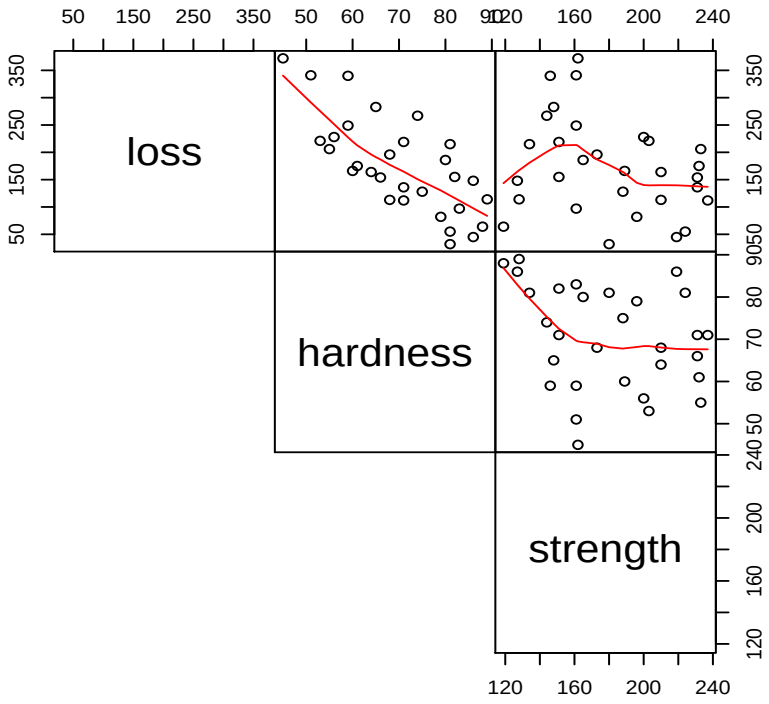


Abbildung 9.5: *abrasion* Daten: EDA und lineares Modell. (Siehe R-Code 9.6.)

R-Code 9.6: *abrasion* Daten: EDA und lineares Modell und Modellvalidierung.

```
abrasion <- read.csv('data/abrasion.csv')
str(abrasion)
## 'data.frame': 30 obs. of  3 variables:
##  $ loss     : int  372 206 175 154 136 112 55 45 221 166 ...
```

```
##  $ hardness: int  45 55 61 66 71 71 81 86 53 60 ...
##  $ strength: int  162 233 232 231 231 237 224 219 203 189 ...
pairs(abrasion, upper.panel=panel.smooth, lower.panel=NULL, gap=0)
res <- lm(loss~hardness+strength, data=abrasion)
summary( res)
##
## Call:
## lm(formula = loss ~ hardness + strength, data = abrasion)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.385 -14.608   3.816  19.755  65.981
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 885.1611    61.7516  14.334 3.84e-14 ***
## hardness     -6.5708     0.5832 -11.267 1.03e-11 ***
## strength     -1.3743     0.1943  -7.073 1.32e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.49 on 27 degrees of freedom
## Multiple R-squared:  0.8402,Adjusted R-squared:  0.8284
## F-statistic:    71 on 2 and 27 DF,  p-value: 1.767e-11
confint( res)
##                 2.5 %       97.5 %
## (Intercept) 758.457323 1011.8648954
## hardness     -7.767432   -5.3742274
## strength     -1.773001   -0.9756228

# Fitted values
plot( loss~hardness, ylim=c(0,400), yaxs='i', data=abrasion)
points( res$fitted~hardness, col=4, data=abrasion)
plot( loss~strength, ylim=c(0,400), yaxs='i', data=abrasion)
points( res$fitted~strength, col=4, data=abrasion)
# Residual vs ...
plot( res$resid~res$fitted)
lines( lowess( res$fitted, res$resid), col=2)
```

```
abline( h=0, col='gray')
plot( res$resid~hardness, data=abrasion)
lines( lowess( abrasion$hardness, res$resid), col=2)
abline( h=0, col='gray')
plot( res$resid~strength, data=abrasion)
lines( lowess( abrasion$strength, res$resid), col=2)
abline( h=0, col='gray')
plot( res$resid[-1]~res$resid[-30])
abline( h=0, col='gray')
```

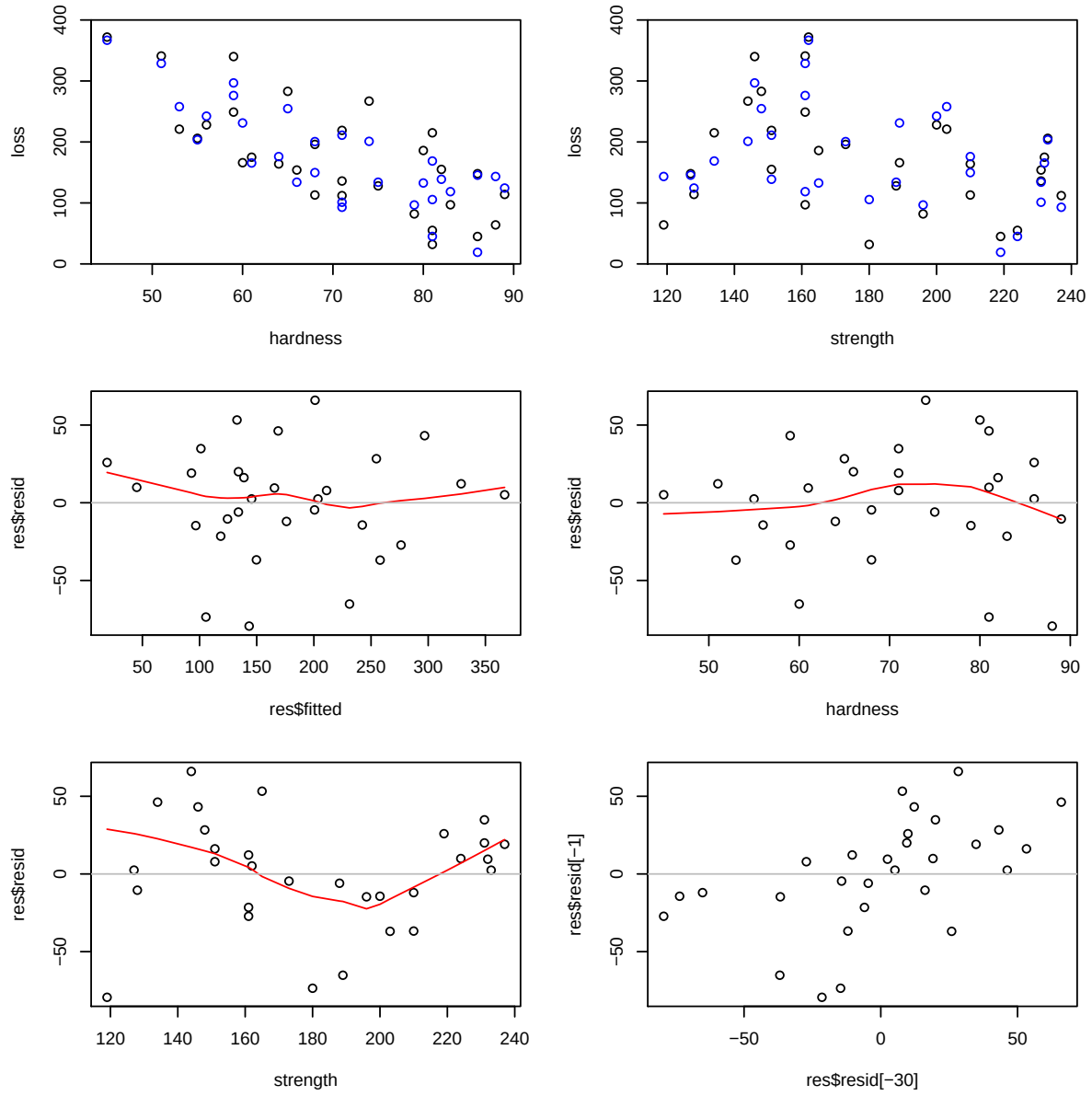


Abbildung 9.6: *abrasion* Daten: Modellvalidierung. (Siehe R-Code 9.6.)

Beispiel 9.3. (*LifeCycleSavings* Daten) Der Beschrieb des Datesatzes ist wie folgt:

> Under the life-cycle savings hypothesis as developed by Franco Modigliani, the savings ratio (aggregate personal saving divided by disposable income) is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the percentage of population less than 15 years old and the percentage of the population over 75 years old. The data are averaged over the decade 1960–1970 to remove the business cycle or other short-term fluctuations.

Der Datensatz besteht für 50 Länder aus den fünf Variablen: `sr` aggregate personal savings, `pop15` % of population under 15, `pop75` % of population over 75, `dpi` real per-capita disposable income, `ddpi` % growth rate of dpi.
Die Punktwolken sind in Abbildung 9.7 dargestellt. Der R-Code 9.7 fittet ein multiples lineares Modell, vereinfacht das Modell durch den Vergleich verschiedener Anpassungsgütekriterien (AIC, BIC) und zeigt eine grafische Modellvalidation auf.
Je nach Kriterium resultieren andere Modelle: bei Modellvereinfachung mit BIC fällt `pop75` aus dem Modell. ♣

R-Code 9.7: *LifeCycleSavings* Daten: EDA und erstes Modell (Abbildung 9.7).

```
data( LifeCycleSavings)
head( LifeCycleSavings)
##              sr pop15 pop75     dpi ddpi
## Australia 11.43 29.35  2.87 2329.68 2.87
## Austria   12.07 23.32  4.41 1507.99 3.93
## Belgium   13.17 23.80  4.43 2108.47 3.82
## Bolivia    5.75 41.89  1.67  189.13 0.22
## Brazil    12.88 42.19  0.83  728.47 4.56
## Canada     8.79 31.72  2.85 2982.88 2.43
pairs(LifeCycleSavings, upper.panel=panel.smooth, lower.panel=NULL,
          gap=0)
lcs.all <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)
summary( lcs.all)
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

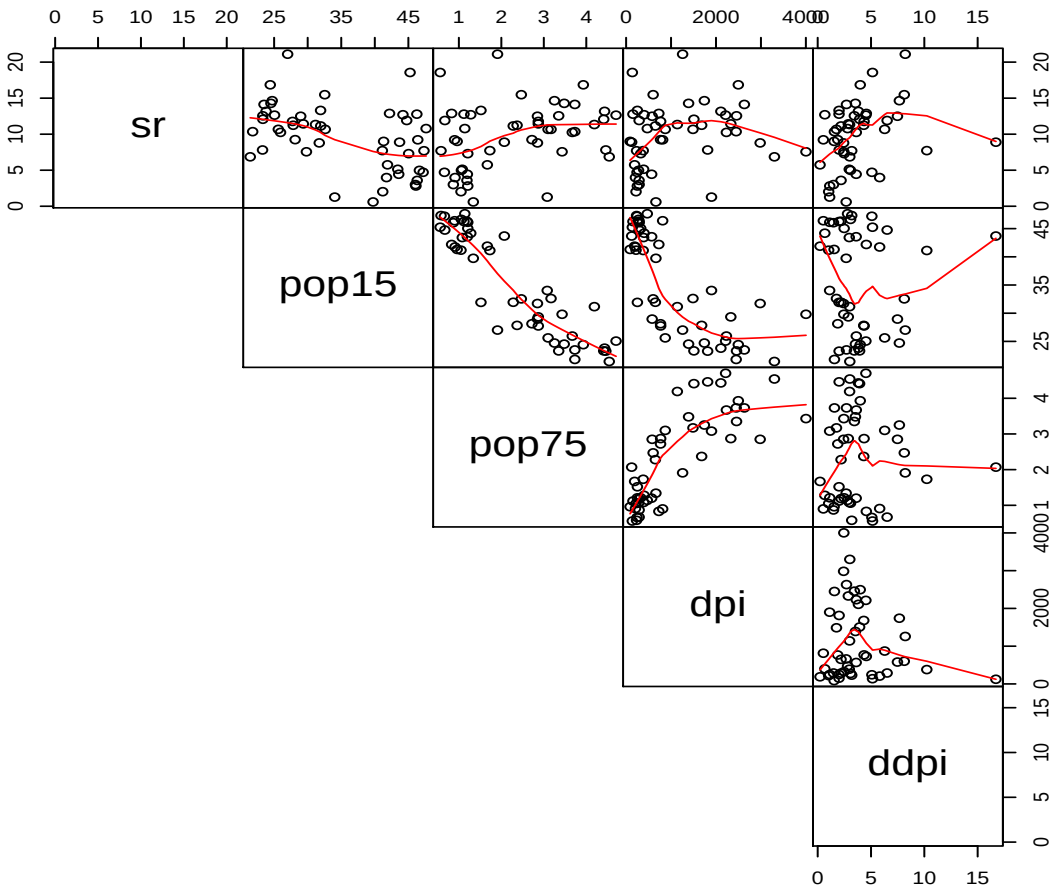


Abbildung 9.7: *LifeCycleSavings* Daten: Punktwolken der Daten.

```
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385,Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

```
lcs.aic <- step( lcs.all)       # AIC is default choice
## Start:  AIC=138.3
## sr ~ pop15 + pop75 + dpi + ddpi
##
##         Df Sum of Sq    RSS    AIC
## - dpi    1     1.893 652.61 136.45
## <none>                650.71 138.30
## - pop75  1    35.236 685.95 138.94
## - ddpi   1    63.054 713.77 140.93
## - pop15  1   147.012 797.72 146.49
##
## Step:  AIC=136.45
## sr ~ pop15 + pop75 + ddpi
##
##         Df Sum of Sq    RSS    AIC
## <none>                652.61 136.45
## - pop75  1    47.946 700.55 137.99
## - ddpi   1    73.562 726.17 139.79
## - pop15  1   145.789 798.40 144.53
summary( lcs.aic)$coefficients
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 28.1246633  7.1837859  3.915020 0.0002969796
## pop15       -0.4517775  0.1409317 -3.205649 0.0024515384
## pop75       -1.8354083  0.9983996 -1.838350 0.0724726983
## ddpi         0.4278317  0.1878856  2.277087 0.0274781754
par( mfrow=c(2,2))
plot( lcs.aic)                  # 4 plots to assess the model@
# BIC:
summary( step( lcs.all, k=log(50), trace=0))$coefficients
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 15.5995758 2.33439442  6.682494 2.479591e-08
## pop15       -0.2163762 0.06033473 -3.586263 7.959672e-04
## ddpi         0.4428302 0.19240135  2.301596 2.583739e-02
```

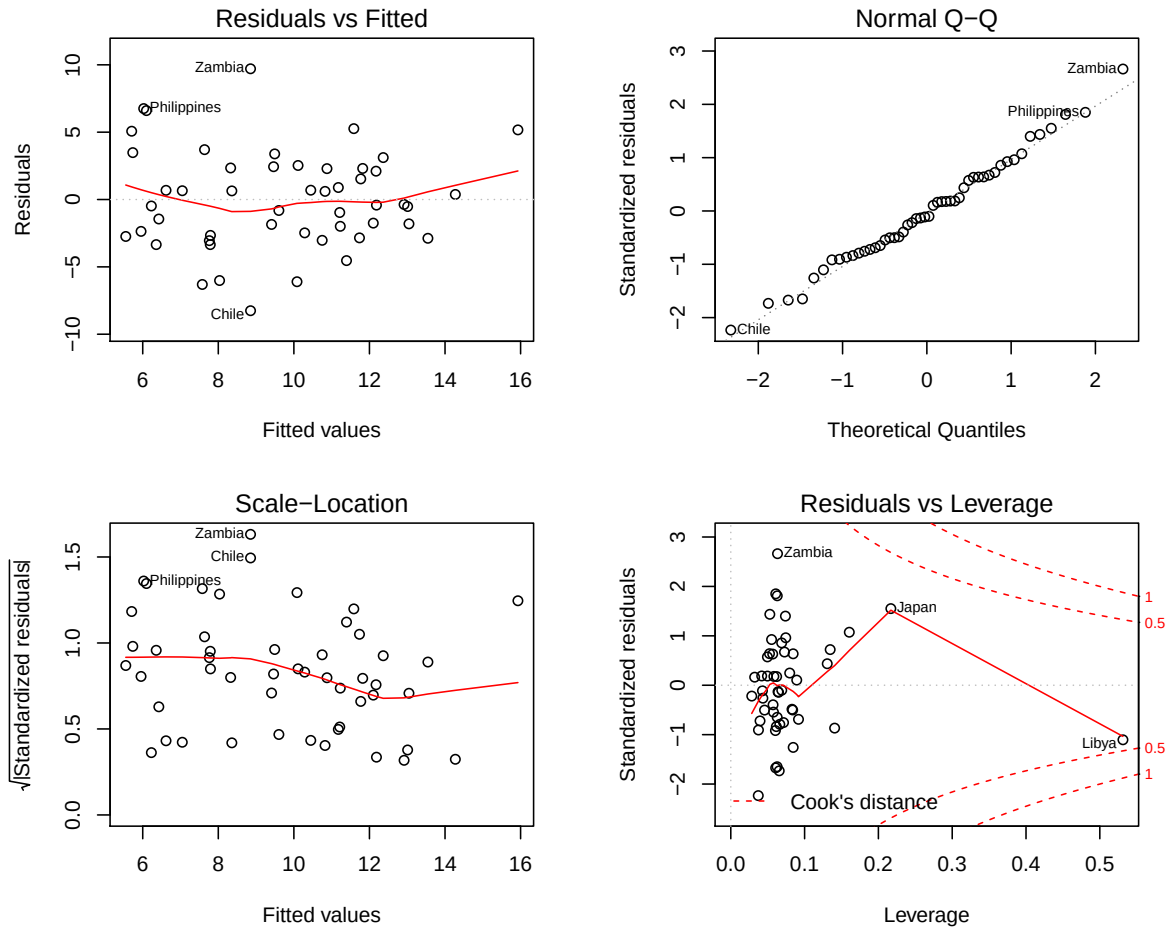


Abbildung 9.8: *LifeCycleSavings* Daten: Modellvalidation

Kapitel 10

Vertiefung: Räumliche Statistik

Im letzten Kapitel haben wir multiple Regression betrachtet. Eine Annahme des Modells waren unkorrelierte Daten. Sehr oft sind jedoch die Daten auf Grund einer räumlichen Nahe korreliert. Wir betrachten verschiedene "Prägungen" von räumlichen Daten und betrachten ein "Geostatistik" Beispiel.

10.1 Regression mit korrelierten Daten

Die Gleichung (9.3) wird verallgemeinert

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma}). \tag{10.1}$$

Das heisst, die Fehler sind nicht mehr iid. Der Kleinste Quadrate Schätzer ist trotzdem $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}$. Dieser Schätzer ist jedoch nicht optimal, da er nicht die gesamte Information des Modells berücksichtigt (hier $\boldsymbol{\Sigma}$). Der "Beste" Schätzer wäre

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{Y}, \tag{10.2}$$

welcher die Korrelationen zwischen den Beobachtungen und die eventuell verschiedenen Varianzen korrekt berücksichtigt.
In der Praxis werden oft folgende spezial Fälle betrachtet:

i) $\boldsymbol{\Sigma} = \sigma^2\mathbf{V}$, wobei $\mathbf{V}$ eine (bekannte) Diagonalmatrix ist. In R wird diese Diagonalmatrix mit dem Argument `weights` berücksichtigt, `weights=1/diag(V)` (gewichtete kleinste Quadrate; *weighted least squares*, WLS).

ii) $\boldsymbol{\Sigma} = \sigma^2\mathbf{R}$, wobei $\mathbf{R}$ eine (bekannte) Korrelationsmatrix ist. In diesem Fall werden die Daten und Prädiktoren mit dem Inversen einer "Quadratwurzel" von $\mathbf{R}$ vormultipliziert (generalisierte kleinste Quadrate; *generalized least squares*, GLS).

iii) $\boldsymbol{\Sigma}$ ist parametrisiert, deren Parameter aber unbekannt. In Rbeinhalted das Packet `nlme` (steht für *Nonlinear Mixed-Effects Models*) die Funktion `gls`.

Der dritte Fall tritt typischerweise bei räumlichen Daten auf, welche wir im folgenden genauer betrachten.

10.2 Räumlichen Daten

Die räumliche Statistik befasst sich mit Daten/Messungen, die ein zusätzliches "Ort"- oder "Raum"-Attribut besitzen. Beispiele sind:

i) Kadmiumgehalt (giftiges Schwermetall) im Sediment des Genfersees, siehe Beispiel 10.1.

ii) Niederschlagsmenge im Monat April 2014 in den über 130 SwissMetNet-Stationen (www.meteoschweiz.admin.ch/web/de/klima/messsysteme/boden/swissmetnet.html)

iii) Anzahl Blauzungen Fälle bei Rindern in den ca. 150 Bezirken der Schweiz (de.wikipedia.org/wiki/Blauzungenkrankheit)

iv) Geburten in den Gemeinden des Kantons Zürich in den Jahren 2000 bis 2010.

v) Detonationen von Unkonventionellen Spreng- und Brandvorrichtungen (*improvised explosive devices, IEDs*) entlang einer US Versorgungsstrasse in Bagdad zwischen dem 13. Februar und dem 10. März 2007.

vi) Viehhöfe, auf denen Blauzungenkrankheit diagnostiziert wurde.

Die sechs Beispiele widerspiegeln die drei (grundsätzlich) verschiedenen Arten von räumlichen Daten:

[A] Geostatistische Daten, *geostatistical data*

[B] "Gitter" Daten, *lattice data*

[C] Punktprozess Daten, *point process data*

In allen Fällen scheint es eine "Korrelation" zwischen den Daten zu geben: Werte in der Nähe scheinen ähnlicher zu sein als solche, die weiter entfernt sind. Diese Beobachtung ist im wesentlichen nach W. Tobler[1] das *first law of geography* "Everything is related to everything else, but near things are more related than distant things." (Tobler, 1970). Wir erfahren dieses "Gesetz" täglich, z.B. durch unser "Wetterempfinden": zwischen Zürich und Effretikon gibt es keine grossen Temperaturunterschiede, zwischen Visp (VS) und Zürich aber schon. Diese Abhänigkeit wird in der räumlichen Statistik ausgenutzt.
Oft können die Fälle [A] und [B] mit multivariater Normalverteilung modeliert werden. Der nachste Abschnitt betrachted den Fall [A].

[1] http://en.wikipedia.org/wiki/Waldo_Tobler

10.3 Geostatistik

Als Basis der Geostatistik benutzen wir räumliche Prozesse

$$\{Y(\boldsymbol{s}), \boldsymbol{x} \in \mathcal{D} \subset \mathbb{R}^d\} \tag{10.3}$$

Im Unterschied zu der "klassischen" multivariaten Statistik hat man in der Geostatistik oft nur eine einzige Beobachtung, an einigen Punkten (Orte, *locations*). Zum Beispiel $z_1, \ldots, z_n$ ist eine Realisation des Prozesses $Y(\boldsymbol{s})$ an den Orten $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$.
Ein klassisches Problem in der Geostatistik ist die Vorhersage einer Grösse an einem Ort oder mehreren Orten, d.h., eine Vorhersage von $Y(\boldsymbol{s}_0)$.
Um das Problem zu vereinfachen, nehmen wir an, dass der Prozess (10.3) normalverteilt ist, d.h., für alle n und alle $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ ist der Zufallsvektor $\big(Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)\big)^\mathsf{T}$ normalverteilt. Die ersten beiden Momente (Erwartungswert und Kovarianz) sind durch die "Funktionen"

$$\mathrm{E}(Y(\boldsymbol{s})) = \mu(\boldsymbol{s}) \tag{10.4}$$

$$\mathrm{Cov}(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)) = c(\boldsymbol{s}_i, \boldsymbol{s}_j) \tag{10.5}$$

gegeben. Da wir nur "eine" (oder wenige) Beobachtung(en) zum Schätzen besitzen, wählen wir parametrische Funktionen für die Erwartungswertfunktion und die Kovarianzfunktion. Als Beispiel:

$$\mathrm{E}(Y(\boldsymbol{s})) = \mu(\boldsymbol{s}) = (1, s_x, s_y)\boldsymbol{\beta} \tag{10.6}$$

$$\mathrm{Cov}(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)) = c(\boldsymbol{s}_i, \boldsymbol{s}_j) = \sigma^2 \exp(-\|\boldsymbol{s}_i - \boldsymbol{s}_j\|/\rho) \tag{10.7}$$

Somit kann das Problem als ein Regressionsproblem mit korreliertem Fehler betrachtet werden

$$Y(\boldsymbol{s}) = \mathbf{X}\boldsymbol{\beta} + Z(\boldsymbol{s}), \tag{10.8}$$

mit $Z(\boldsymbol{s})$ als einem normalverteilem Prozess mit Erwartungswert null und Kovarianzfunktion (10.7). Jedoch werden oft andere Methoden als (verallgemeinerte) Kleinste-Quadrate-Methode (*generalized least squares*, *GLS*) gebraucht um das Modell (10.8) zu fitten.
Vorhersage oder Prädiktion wird in der Geostatistik oft als Kriging *kriging* bezeichnet. Der Krigingprädiktor wird als linearer unverfälschter Schätzer mit kleinstem Quadratfehler hergeleitet. Im Wesentlichen

$$\widehat{Y}(\boldsymbol{s}_0) = \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{y} \tag{10.9}$$

wobei $\boldsymbol{y}^\mathsf{T} = (y_1, \ldots, y_n)^\mathsf{T}$. Die Gewichte $\boldsymbol{\lambda}$ hängen von $\boldsymbol{s}_0$ ab und sind so, dass Beobachtungen in der Nahe von $\boldsymbol{s}_0$ stärker berücksichtigt werden als Beobachtungen, die weiter entfernt sind. Zudem erhalten Beobachtungen mit grösserer Varianz ein kleineres Gewicht.

Oft wird zu (10.8) noch ein Messfehler angefügt. Im Fall von einem normalverteilen Prozess mit konstantem (unbekanntem) Mittelwert ist der Kriging Prädiktor immer noch der Plug-in-Prädiktor vom bedingten Erwartungswert

$$\widehat{\mu} + \widehat{\boldsymbol{\Sigma}}_{12}\widehat{\boldsymbol{\Sigma}}_{22}^{-1}(\boldsymbol{y} - \widehat{\mu}\mathbf{1}) \tag{10.10}$$

wobei $\widehat{\boldsymbol{\Sigma}}_{12}$ und $\widehat{\boldsymbol{\Sigma}}_{22}$ basierend auf (10.7) mit $\widehat{\sigma}^2$ und $\widehat{\rho}$ konstruiert sind.

Historisch wurde die Kovarianzfunktion (10.5) durch eine ähnliche Form ausgedrückt, welche das Variogram genannt wurde. Unter Varigramschätzung verstehen wir das Schätzen der Kovarianzmatrix in (10.1) oder Kovarianzfunktion 10.5.
Alternativ können die Parameter der Kovarianzfunktion auch mit einer Likelihood Methode geschätzt werden.
In R sind die Pakete `fields` oder `geoR` hilfreich für die Modellierung von räumlichen Daten.

10.4 Beispiel(e)

Beispiel 10.1. (*leman* Daten) Kadmiumgehalt (giftiges Schwermetall) im Sediment des Genfersees, 293 Messwerte aus dem Jahr 1983. Siehe Furrer and Genton (1999).
Die Abbildung 10.1 zeigt das Tiefenprofil des Genfersees. Die Messwerte sind in R-Code 10.1 und Abbildung 10.2 gezeigt. Offensichtlich gibt es einen sehr grossen Messwert (bei Vidy/Lausanne). Wir fahren mit der Analyse ohne diesen Wert fort.
Kriging ist in R-Code 10.2 gegeben (Abbildungen in 10.3). Wir zeigen insbesondere die einzelnen Teile des Modells (10.8) auf. ♣

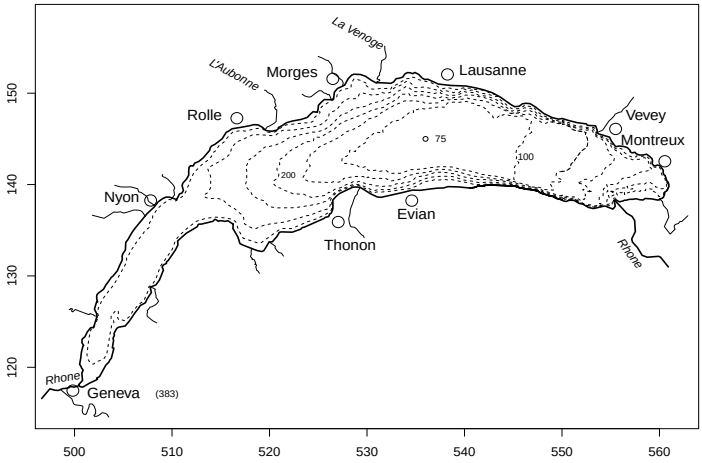


Abbildung 10.1: Genfersee (Schweizer Landeskoordinaten, CH1903).

R-Code 10.1 *leman* Daten: EDA. (Siehe Abbildung 10.2.)

```
load( "data/leman83_red.RData")
str( leman)
## 'data.frame': 293 obs. of  10 variables:
##  $ x : num  501 502 501 503 502 ...
##  $ y : num  119 120 121 121 122 ...
##  $ Zn: num  72.2 98.2 81.6 131 160 125 48 131 127 106 ...
##  $ Cr: num  49.3 45.5 72.8 63.6 66.8 55 11.1 59.4 67.1 43.6 ...
##  $ Cd: num  0.23 0.37 0.14 0.3 0.56 0.3 0.17 0.44 0.39 0.26 ...
##  $ Co: num  22.1 8.5 14 4.6 5.7 9.5 7.5 4.5 19.6 6.6 ...
##  $ Sr: num  345 308 240 383 356 349 350 374 345 419 ...
##  $ Hg: num  0.17 0.21 0.06 0.24 0.35 0.14 0.08 0.26 0.23 0.18 ...
##  $ Pb: num  11 15 15 18 24 22 7 19 23 7 ...
##  $ Ni: num  32.5 33.8 55.9 47.1 52.5 41.8 11.7 45.5 50.6 32.8 ...
library( fields)
plot( y~x, col=tim.colors()[cut(Cd,64)], pch=19, data=leman)
   # simpler alternative is  quilt.plot
lines( lake.data)
hist( leman$Cd, breaks=20, main='')
which.max( leman$Cd)   # index of the largest value
## [1] 220
```

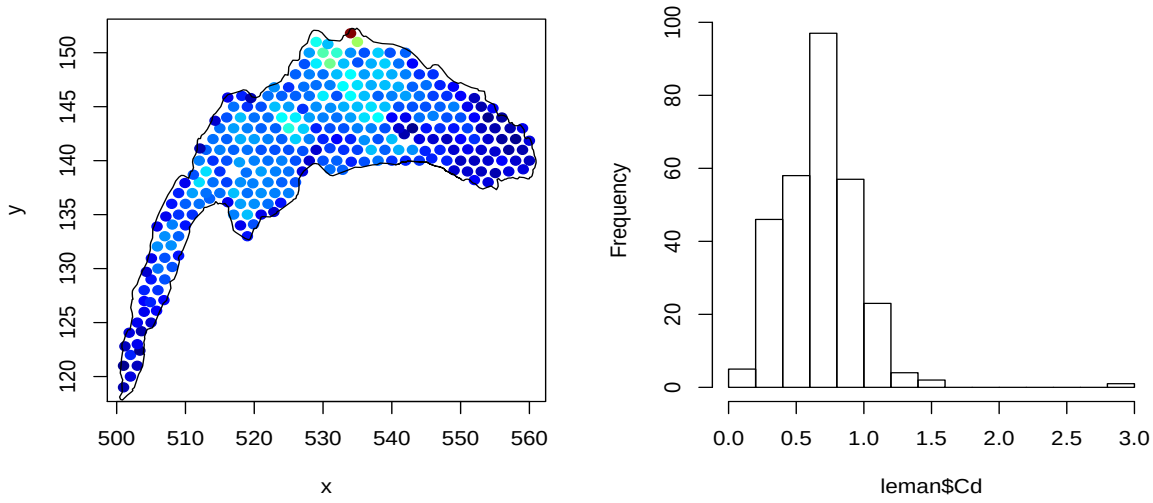


Abbildung 10.2: *leman* Daten: EDA. (Siehe R-Code 10.1.)

R-Code 10.2 *leman* Daten: Prädiktion. (Siehe Abbildung 10.3.)

```
# Construct a grid within the lake boundaries:
library(splancs)
xr <- seq(min(lake.data[,1]),to=max(lake.data[,1]),l=100)
yr <- seq(min(lake.data[,2]),to=max(lake.data[,2]),by=xr[2]-xr[1])
   # xr and yr are fine sequences of points

locs <- data.frame( x=lake.data[,1], y=lake.data[,2])
grid <- expand.grid( x=xr, y=yr)      # create a 2-dim grid
pts <- pip( grid, locs, bound=TRUE)  # pip points-in-polygon

# Kriging:
out <- Krig( cbind(leman$x,leman$y)[-220,], leman$Cd[-220],
            give.warnings=FALSE)
pout <- predict( out, pts)
quilt.plot(pts, pout, nx=length(xr)-2, ny=length(yr)-1)

fit0 <- predict( out, pts,  just.fixed=TRUE) # trend
quilt.plot(pts, fit0, nx=length(xr)-2, ny=length(yr)-1)

fit1 <- pout - fit0                    # smooth spatial
quilt.plot(pts, fit1, nx=length(xr)-2, ny=length(yr)-1)
```

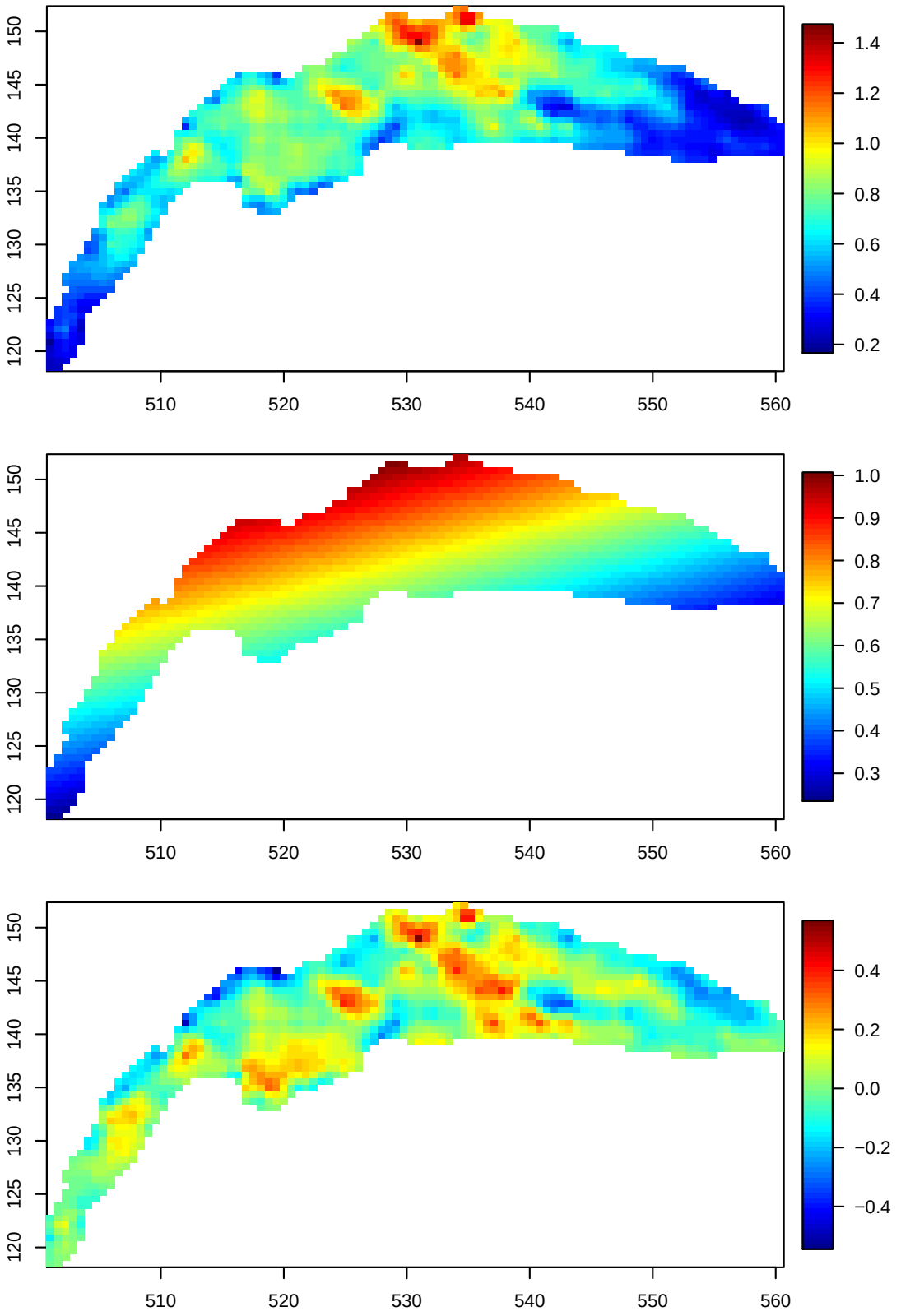


Abbildung 10.3: *leman* Daten: Prädiktion $\widehat{Y}(\cdot)$ (oben), linearer Trend $\mathbf{X}\widehat{\boldsymbol{\beta}}$ (mitte) und räumliche Komponente $\widehat{Z}(\cdot)$ (unten). Die Figuren besitzen verschiedene Skalen. (Siehe R-Code 10.2.)

Kapitel 11

ANOVA

Im Kapitel 4 haben wir zwei Mittelwerte miteinander verglichen. Natürlich kann der selbe Ansatz auf I Stichproben angewandt werden (durchführen von $\binom{I}{2}$ solcher Tests).
In diesem Abschnitt lernen wir eine "bessere" Methode kennen, basierend auf der Varianzanalyse (*analysis of variance*, *ANOVA*).

11.1 Einfaktorielle Varianzanalyse

Das Modell besteht aus I Gruppen (Faktorstufen), $i = 1, ..., I$, und jede Gruppe enthält n_i Stichprobenelemente, $j = 1, ..., n_i$ und $N = n_1 + \cdots + n_I$. Das heisst

$$Y_{ij} = \mu_i + \varepsilon_{ij} \tag{11.1}$$
$$= \mu + \beta_i + \varepsilon_{ij} \tag{11.2}$$

mit $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Jedoch ist das Modell (11.2) über-parametrisiert (I Gruppen und $I + 1$ Parameter) und eine zusätzliche Bedingung (*constraint*) ist notwendig. Oft wird

$$\sum_{i=1}^{I} \beta_i = 0 \quad \text{oder} \quad \beta_1 = 0 \tag{11.3}$$

verwendet (*sum-to-zero-contrast* und *treatment contrast*).
Unser Interesse "gibt es einen Unterschied in den Gruppen" induziert die statistische Hypothese $H_0 : \beta_1 = \beta_2 = \cdots = \beta_I = 0$. Wir werden zuerst die Schätzer herleiten und anschliessend die Teststatistik.
Das Modell (11.2) mit $I = 2$ und Bedingung $\beta_1 = 0$ kann als Regressionsproblem betrachtet werden

$$Y_i^* = \beta_0^* + \beta_1^* x_i + \varepsilon_i^*, \qquad i = 1, \ldots, N \tag{11.4}$$

mit Y_i^* den Komponenten des Vektors $(Y_{11}, Y_{12}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2})^\mathsf{T}$ und $x_i = 0$ wenn $i = 1, \ldots, n_1$ und $x_i = 1$ andernfalls. Um die Schreibweise zu vereinfachen ersparen wir

uns den oberen Index und somit

$$\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} + \boldsymbol{\varepsilon} = \begin{pmatrix}\mathbf{1} & \mathbf{0}\\ \mathbf{1} & \mathbf{1}\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} + \boldsymbol{\varepsilon} \tag{11.5}$$

$$\begin{aligned}\widehat{\boldsymbol{\beta}} = \begin{pmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\end{pmatrix} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{pmatrix} = \begin{pmatrix}N & n_2\\ n_2 & n_2\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{1}^\mathsf{T} & \mathbf{1}^\mathsf{T}\\ \mathbf{0}^\mathsf{T} & \mathbf{1}^\mathsf{T}\end{pmatrix}\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{pmatrix}\\ &= \frac{1}{n_1 n_2}\begin{pmatrix}n_2 & -n_2\\ -n_2 & N\end{pmatrix}\begin{pmatrix}\sum_{ij} y_{ij}\\ \sum_j y_{2j}\end{pmatrix} = \begin{pmatrix}\frac{1}{n_1}\sum_j y_{1j}\\ \frac{1}{n_2}\sum_j y_{2j} - \frac{1}{n_1}\sum_j y_{1j}\end{pmatrix}\end{aligned} \tag{11.6}$$

Das heisst, die Kleinste-Quadrate Schätzer von μ und β_2 in (11.2) sind die Mittelwerte der ersten Gruppe und die Differenz der Mittelwerte der beiden Gruppen.
Mit anderen Bedingungen können die Schätzer entsprechend ähnlich hergeleitet werden.

Historisch war oft $n_1 = \cdots = n_I = J$. In diesem Fall mit der Summenbedingung gibt es auch noch einen weiteren, einfachen Ansatz die Schätzer herzuleiten. Dazu benutzen wir die "Punkt"-Schreibweise, um zu zeigen für welchen Index das Mittel berechnet wurde, zum Beispiel,

$$\overline{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \quad \text{und} \quad \overline{y}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} \tag{11.7}$$

Basierend auf $Y_{ij} = \mu + \beta_i + \varepsilon_{ij}$ benutzen wir für die Schätzfunktionen den folgenden Ansatz

$$y_{ij} = \overline{y}_{\cdot\cdot} + (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}) + (y_{ij} - \overline{y}_{i\cdot}) \tag{11.8}$$

Mit der Kleinste-Quadrate Methode werden $\widehat{\mu}$ und $\widehat{\beta}_i$ so gewählt, dass

$$\sum_{i,j}(y_{ij} - \widehat{\mu} - \widehat{\beta}_i)^2 = \sum_{i,j}(\overline{y}_{\cdot\cdot} + \overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot} + y_{ij} - \overline{y}_{i\cdot} - \widehat{\mu} - \widehat{\beta}_i)^2 \tag{11.9}$$

$$= \sum_{i,j}((\overline{y}_{\cdot\cdot} - \widehat{\mu}) + (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot} - \widehat{\beta}_i) + y_{ij} - \overline{y}_{i\cdot})^2 \tag{11.10}$$

minimal ist. Dazu wird das Quadrat entwickelt. Die Kreuzterme sind Null, da

$$\sum_{j=1}^{J}(y_{ij} - \overline{y}_{i\cdot}) = 0 \quad \text{und} \quad \sum_{i=1}^{I}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot} - \widehat{\beta}_i) = 0 \tag{11.11}$$

und somit $\widehat{\mu} = \overline{y}_{\cdot\cdot}$ und $\widehat{\beta}_i = \overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}$.

Die Idee der Teststatistik zur statistischen Hypothese $H_0 : \beta_1 = \beta_2 = \cdots = \beta_I = 0$ beruht auf einer Zerlegung der Varianz bei der die Varianzen zwischen den Gruppen mit denjenigen innerhalb verglichen wird.
Wie in der Regression werden die Beobachtungen in den Raum der durch μ und β_i aufgespannt wird, orthogonal projiziert. Diese orthogonale Projektion erlaubt es die Quadratsumme der Beobachtungen in eine Quadratsumme Modell und Quadratsumme Fehler

aufzuteilen (die mittelwertkorrigierten Werte, um präzise zu sein). Diese Quadratsummen werden anschliessend gewichtet und verglichen. Eine Darstellung dieser in einer Tabelle mit anschliessender Interpretation wird oft der Varianzanalyse gleichgestellt.
Die Quadratsummenzerlegung kann auch mit Hilfe von (11.8) hergeleitet werden. Es werden keine Annahmen über die Bedingungen und über n_i gemacht:

$$\sum_{i,j}(y_{ij}-\overline{y}_{..})^2 = \sum_{ij}(\overline{y}_{i.}-\overline{y}_{..}+y_{ij}-\overline{y}_{i.})^2 \tag{11.12}$$

$$= \sum_{ij}(\overline{y}_{i.}-\overline{y}_{..})^2 + \sum_{i,j}(y_{ij}-\overline{y}_{i.})^2 + \sum_{i,j}2(\overline{y}_{i.}-\overline{y}_{..})(y_{ij}-\overline{y}_{i.}). \tag{11.13}$$

Der Kreuzterm ist auf Grund $\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.}) = 0$ wieder null. Daher haben wir die Zerlegung der Quadratsummen

$$\underbrace{\sum_{i,j}(y_{ij}-\overline{y}_{..})^2}_{\text{Total}} = \underbrace{\sum_{i,j}(\overline{y}_{i.}-\overline{y}_{..})^2}_{\text{Modell}} + \underbrace{\sum_{i,j}(y_{ij}-\overline{y}_{i.})^2}_{\text{Fehler}}. \tag{11.14}$$

Diese Gleichung mit den Kleinste-Quadrate Schätzwerten $\widehat{\mu} = \overline{y}_{..}$ und $\widehat{\beta}_i = \overline{y}_{i.} - \overline{y}_{..}$ kann als

$$\frac{1}{n}\sum_{i,j}(y_{ij}-\widehat{\mu})^2 = \frac{1}{n}\sum_i n_i(\widehat{\mu+\beta_i}-\widehat{\mu})^2 + \frac{1}{n}\sum_{i,j}(y_{ij}-\widehat{\mu+\beta_i})^2 \tag{11.15}$$

$$\widehat{\mathrm{Var}(y_{ij})} = \widehat{\sigma^2} + \frac{1}{n}\sum_i n_i\widehat{\beta}_i^2 \tag{11.16}$$

gelesen werden. Formell muss jedoch ein bisschen präziser gearbeitet werden. Ein gutes Modell hat einen kleinen Schätzwert für $\widehat{\sigma^2}$ im Vergleich zum zweiten Summand. Wir arbeiten nun einen quantitativen Vergleich der Summanden aus.

Die Anzahl Beobachtungen muss berücksichtigt werden. Um die einzelnen Quadratsummen zu gewichten, werden diese durch die Freiheitsgrade (*degrees of freedom*) dividiert. Unter der Nullhypothese sind diese mittleren Quadratsummen Chi-Quadrat verteilt und somit deren Quotient F-verteilt. Daher wird wieder ein F-Test gebraucht (siehe Test 4). Die ANOVA-Tafel 11.1 besteht aus den Spalten Streuungsursache (Streuung zwischen den Gruppen, innerhalb der Gruppen oder total), Quadratsummen, Freiheitsgrade, Mittlere Quadratsummen und Testgrössen. Die Zeilen basieren auf den Faktor(en), Fehler und Total/Summe.
Der Erwartungswert der Mittleren Quadratsummen ist

$$\mathrm{E}(\mathrm{MQ}_A) = \sigma^2 + \frac{\sum n_i\beta_i^2}{I-1} \qquad \mathrm{E}(\mathrm{MQ}_E) = \sigma^2 \tag{11.17}$$

Somit ist unter H_0 das Verhältnis $\mathrm{MQ}_A/\mathrm{MQ}_E$ klein ("$\approx$" 1, siehe Abschnitt 2.7.4) und wir verwerfen H_0 für grosse Werte.

Tabelle 11.1: Tafel der einfachen Varianzanalyse.

Ursache der Streuung	Quadratsumme (QS)	Freiheitsgrade (FG)	mittlere Quadratsumme (MQ)	Testgrösse
Faktor A (Gruppen, Stufen, ...)	$\mathrm{QS}_A = \sum_{i,j}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2$	$I-1$	$\mathrm{MQ}_A = \frac{\mathrm{QS}_A}{I-1}$	$F_{\mathrm{Vers}} = \frac{\mathrm{MQ}_A}{\mathrm{MQ}_E}$
Fehler (Rest-Streuung)	$\mathrm{QS}_E = \sum_{i,j}(y_{ij} - \overline{y}_{i\cdot})^2$	$N-I$	$\mathrm{MQ}_E = \frac{\mathrm{QS}_E}{N-I}$	
Total	$\mathrm{QS}_T = \sum_{i,j}(y_{ij} - \overline{y}_{\cdot\cdot})^2$	$N-1$		

Der Test 13 zeit den schrittweisen Ablauf auf. Die Testgrösse der Tabelle 11.1 stimmt natürlich mit derjenigen des Tests 13 überein. Zu beachten ist, dass wenn $\mathrm{MQ}_A \leq \mathrm{MQ}_E$, d.h. $F_{\mathrm{Vers}} \leq 1$, H_0 nie verworfen wird. Details zu F-verteilten Zufallsvariablen sind im Abschnitt 2.7.4 gegeben.
Das Beispiel 11.1 diskutiert eine einfache Varianzanalyse.

Test 13: Durchführung der einfaktoriellen Varianzanalyse

Fragestellung: Gibt es unter den Mittelwerten $\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_I$ mindestens zwei, die voneinander signifikant verschieden sind?

Voraussetzungen: Die I Grundgesamtheiten sind normalverteilt mit homogenen Varianzen. Die entnommenen Stichproben sind unabhängig.

Berechnung: Erstellen einer Tafel der einfachen Varianzanalyse. Gebraucht wird der Quotient der Mittleren Quadratsummen des Faktors und des Fehlers:

$$F_{\mathrm{Vers}} = \frac{\mathrm{QS}_A/(I-1)}{\mathrm{QS}_E/(N-I)} = \frac{\mathrm{MQ}_A}{\mathrm{MQ}_E}$$

Entscheidung: Vergleiche F_{Vers} mit $F_{\mathrm{Tab}}(I-1, N-I; \alpha)$, wobei $I-1$ die Freiheitsgrade "zwischen" und $N-k$ die Freiheitsgrade "innerhalb": verwerfe $H_0 : \beta_1 = \beta_2 = \cdots = \beta_I$ wenn $F_{\mathrm{Vers}} > F_{\mathrm{Tab}}$

Berechnung in R: `summary( lm(...))` für den Wert der Teststatistik oder `summary( aov(...))` für die explizite Darstellung.

Beispiel 11.1. ***retardant*** **Daten** Der Beschrieb des Datensatzes ist wie folgt:

> Many substances related to human activities end up in wastewater and accumulate in sewage sludge. The present study focuses on hexabromocyclododecane (HBCD) detected in sewage sludge collected from a monitoring network in Switzerland. HBCD's main use is in expanded and extruded polystyrene for thermal insulation foams, in building and construction. HBCD is also applied in the backcoating of textiles, mainly for upholstery furniture. A very small application of HBCD is in high impact polystyrene, which is used for electrical and electronic appliances, for example in audio visual equipment. Data and more detailed background information are given in Kupper *et al.* (2008) where it is also argued that loads from different types of monitoring sites showed that brominated flame retardants ending up in sewage sludge originate mainly from surface runoff, industrial and domestic wastewater.

HBCD ist gesundheitsschädlich, beeinträchtigt möglicherweise die Fortpflanzungsfähigkeit und kann das Kind im Mutterleib möglicherweise schädigen.
Im R-Code 11.1 werden die Daten geladen und auf Hexabromcylkododekan reduziert.
Zuerst benutzen wir die Bedingung $\beta_1 = 0$, das heisst, Modell (11.5). Die Schätzwerte stimmen natürlich mit denjenigen aus (11.6) überein.
Anschliessend brauchen wir die Summenbedingung und vergleichen die Resultate. Die geschätzten Werte ändern sich, ebenso die Standardfehler (und somit die p-Werte des t-Tests. Die p-Werte der F-Tests sind aber identisch, da es sich um den gleichen Test handelt.

Der R Befehl `aov` ist eine Alternative, eine ANOVA zu berechnen und ist im R-Code 11.2 illustriert. Wir bevorzugen aber den allgemeineren `lm` Ansatz. Jedoch benötigen einige Funktionen deren Resultate, wie zum Beispiel die Funktion `TukeyHSD`, welche *Tukey's HSD* (honest significant difference) Test ausführt. Die Differenzen können auch aus den Koeffizienten in R-Code 11.1 berechnet werden. Die p-Werte sind kleiner, da multiples Testen berücksichtigt wird. ♣

R-Code 11.1: *retardant* Daten: ANOVA mit `lm` Befehl und Illustration von verschiedenen Kontrasten.

```
tmp <- read.csv('./data/retardant.csv')
retardant <- read.csv('./data/retardant.csv', skip=1)
names( retardant) <- names(tmp)
HBCD <- retardant$cHBCD
str( retardant$StationLocation)
##  Factor w/ 16 levels "A11","A12","A15",..: 1 2 3 4 5 6 7 8 11 12 ...
```

```r
type <- as.factor( rep(c("A","B","C"), c(4,4,8)))

lmout <- lm( HBCD ~ type )
summary(lmout)
##
## Call:
## lm(formula = HBCD ~ type)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -87.62 -44.43 -26.31  22.02 193.07
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    75.67      42.40   1.785   0.0977 .
## typeB          77.25      59.97   1.288   0.2201
## typeC         107.79      51.93   2.076   0.0583 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.8 on 13 degrees of freedom
## Multiple R-squared:  0.2494,Adjusted R-squared:  0.134
## F-statistic:  2.16 on 2 and 13 DF,  p-value: 0.1549
options( "contrasts")
## $contrasts
##         unordered           ordered
## "contr.treatment"      "contr.poly"
# manually construct the estimates:
c( mean(HBCD[1:4]), mean(HBCD[5:8])-mean(HBCD[1:4]),
   mean(HBCD[9:16])-mean(HBCD[1:4]))
## [1]  75.6750  77.2500 107.7875
# change the constrasts to sum-to-zero
options(contrasts=c("contr.sum","contr.sum"))
lmout1 <- lm( HBCD ~ type )
summary(lmout1)
##
## Call:
## lm(formula = HBCD ~ type)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -87.62 -44.43 -26.31  22.02 193.07
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   137.35      22.35   6.146 3.51e-05 ***
## type1         -61.68      33.15  -1.861   0.0855 .
## type2          15.57      33.15   0.470   0.6463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.8 on 13 degrees of freedom
## Multiple R-squared:  0.2494,Adjusted R-squared:  0.134
## F-statistic:  2.16 on 2 and 13 DF,  p-value: 0.1549
```

```
beta <- as.numeric(coef(lmout1))
# Construct 'contr.treat' coefficients:
c( beta[1]+beta[2], beta[3]-beta[2], -2*beta[2]-beta[3])
```

```
## [1]  75.6750  77.2500 107.7875
```

11.2 Mehrfaktoren Varianzanalyse

Das Modell (11.2) kann mit zusätzlichen Faktoren erweitert werden. Ein Zweifaktoren-Modell ist

$$Y_{ijk} = \mu + \beta_i + \gamma_j + \varepsilon_{ijk}, \tag{11.18}$$

mit $k = 1, \ldots, n_{ij}$ und $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$. Es müssen auch wieder zusätzliche Bedingungen formuliert werden
Auch können "klassische" Prädiktoren in (11.18) berücksichtigt werden.
Wenn nicht alle n_{ij} gleich sind, ist die Quadratsummenzerlegung nicht mehr notwendigerweise eindeutig. In der Praxis spielt das oft eine untergeordnete Rolle da vor allem die geschätzten Koeffizienten miteinander verglichen werden ("Kontraste"). Wir empfehlen daher immer den Befehl `lm(...)` zu benutzen.

R-Code 11.2 *retardant* Daten: ANOVA mit `aov` und multiples Testen der Mittelwerte.

```
aovout <- aov( HBCD ~ type )
options("contrasts")
## $contrasts
## [1] "contr.sum" "contr.sum"
coefficients( aovout)  # coef( aovout) is suffient as well.
## (Intercept)       type1       type2
##   137.35417   -61.67917    15.57083
summary(aovout)
##             Df Sum Sq Mean Sq F value Pr(>F)
## type         2  31069   15534    2.16  0.155
## Residuals   13  93492    7192
TukeyHSD( aovout)
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = HBCD ~ type)
##
## $type
##         diff        lwr      upr     p adj
## B-A  77.2500  -81.08515 235.5852 0.4260245
## C-A 107.7875  -29.33476 244.9098 0.1337630
## C-B  30.5375 -106.58476 167.6598 0.8288246
```

11.3 Vollständige Zweifaktoren Varianzanalyse

Das Modell besteht aus $I \cdot J$ Gruppen und jede Gruppe enthält K Stichprobenelemente und $N = I \cdot J \cdot K$. Das heisst

$$Y_{ijk} = \mu + \beta_i + \gamma_j + \varepsilon_{ijk} \tag{11.19}$$

mit $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$. Als zusätzliche Bedingungen werden oft

$$\sum_{i=1}^{I} \beta_i = 0 \qquad \sum_{j=1}^{J} \gamma_j = 0 \quad \text{oder} \quad \beta_1 = 0,\ \gamma_1 = 0 \tag{11.20}$$

verwendet.

Da für alle i und j Zellen je K Beobachtungen vorhanden sind, nennt man diesen Fall vollständig (*balanced*). Die Berechnungen der Schätzwerte sind um einiges einfacher als im unbalancierten (*unbalanced*) Fall.

Wie im Einfaktoren-Fall können wir die Quadratsummen auftrennen.

$$y_{ijk} = \underbrace{\overline{y}_{...}}_{\widehat{\mu}} + \underbrace{\overline{y}_{i..} - \overline{y}_{...}}_{\widehat{\beta}_i} + \underbrace{\overline{y}_{.j.} - \overline{y}_{...}}_{\widehat{\gamma}_j} + \underbrace{y_{ijk} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...}}_{r_{ijk}} \tag{11.21}$$

Unser Interesse basiert auf den statistischen Hypothesen $H_0 : \beta_1 = \cdots = \beta_I = 0$ und $H_0 : \gamma_1 = \cdots = \gamma_J = 0$. Die Testgrössen dieser beiden Tests sind in der letzten Spalte der Tabelle 11.3 gegeben. Die Testgrössen $F_{\text{Vers},A}$ ($F_{\text{Vers},B}$) werden mit den Quantilen der F-Verteilung mit $I-1$ ($J-1$) und $N-I-J+1$ Freiheitsgraden verglichen. Das heisst, die Tests sind entsprechen demjenigen des Tests 13.

Tabelle 11.2: Tafel der vollständigen zweifachen Varianzanalyse.

Ursache	QS	FG	MQ	Testgrösse
Faktor A	$\text{QS}_A = \sum_{i,j,k}(\overline{y}_{i..} - \overline{y}_{...})^2$	$I-1$	$\text{MQ}_A = \frac{\text{QS}_A}{I-1}$	$F_{\text{Vers},A} = \frac{\text{MQ}_A}{\text{MQ}_E}$
Faktor B	$\text{QS}_B = \sum_{i,j,k}(\overline{y}_{.j.} - \overline{y}_{...})^2$	$J-1$	$\text{MQ}_B = \frac{\text{QS}_B}{J-1}$	$F_{\text{Vers},B} = \frac{\text{MQ}_B}{\text{MQ}_E}$
Fehler	$\text{QS}_E = \sum_{i,j,k}(y_{ijk} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2$	$N-I$	$\text{MQ}_E = \frac{\text{QS}_E}{N-I}$	
Total	$\text{QS}_T = \sum_{i,j,k}(y_{ijk} - \overline{y}_{...})^2$	$N-1$		

Im Modell kann auch noch eine Interaktion $(\beta\gamma)_{ij}$ eingefügt werden, falls die Effekte nicht linear sind:

$$Y_{ijk} = \mu + \beta_i + \gamma_j + (\beta\gamma)_{ij} + \varepsilon_{ijk} \tag{11.22}$$

mit $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$. Als zusätzliche Bedingungen zu (11.20) werden oft

$$\sum_{i=1}^{I}(\beta\gamma)_{ij} = 0 \quad \text{und} \quad \sum_{j=1}^{J}(\beta\gamma)_{ij} = 0 \quad \text{für alle } i \text{ und } j \tag{11.23}$$

oder entsprechende Treatment-Bedingungen verwendet. Im Falle $K = 1$, ist die Interaktion der Fehler. Wie im Einfaktoren-Fall können wir die Quadratsummen auftrennen.

$$y_{ijk} = \underbrace{\overline{y}_{...}}_{\widehat{\mu}} + \underbrace{\overline{y}_{i..} - \overline{y}_{...}}_{\widehat{\beta}_i} + \underbrace{\overline{y}_{.j.} - \overline{y}_{...}}_{\widehat{\gamma}_j} + \underbrace{\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...}}_{\widehat{(\beta\gamma)}_{ij}} + \underbrace{y_{ijk} - \overline{y}_{ij.}}_{r_{ijk}} \tag{11.24}$$

Die Tabelle 11.3 zeigt die Tafel der vollständigen zweifachen Varianzanalyse. Die Testgrössen $F_{\text{Vers},A}$, $F_{\text{Vers},B}$ und $F_{\text{Vers},AB}$ werden mit den Quantilen der F-Verteilung mit $I-1$ ($J-1$, $(I-1)(J-1)$) und $N-IJ$ Freiheitsgraden verglichen.

Tabelle 11.3: Tafel der vollständigen zweifachen Varianzanalyse mit Wechselwirkung.

Ursache	QS	FG	MQ	Testgrösse
Faktor A	$\text{QS}_A = \sum_{i,j,k}(\overline{y}_{i..} - \overline{y}_{...})^2$	$I-1$	$\text{MQ}_A = \frac{\text{QS}_A}{\text{FG}_A}$	$F_{\text{Vers},A} = \frac{\text{MQ}_A}{\text{MQ}_E}$
Faktor B	$\text{QS}_B = \sum_{i,j,k}(\overline{y}_{.j.} - \overline{y}_{...})^2$	$J-1$	$\text{MQ}_B = \frac{\text{QS}_B}{\text{FG}_B}$	$F_{\text{Vers},B} = \frac{\text{MQ}_B}{\text{MQ}_E}$
Wechsel-wirkung	$\text{QS}_{AB} = \sum_{i,j,k}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2$	$(I-1)\times(J-1)$	$\text{MQ}_{AB} = \frac{\text{QS}_{AB}}{\text{FG}_{AB}}$	$F_{\text{Vers},AB} = \frac{\text{MQ}_{AB}}{\text{MQ}_E}$
Fehler	$\text{QS}_E = \sum_{i,j,k}(y_{ijk} - \overline{y}_{ij.})^2$	$N-IJ$	$\text{MQ}_E = \frac{\text{QS}_E}{\text{FG}_E}$	
Total	$\text{QS}_T = \sum_{i,j,k}(y_{ijk} - \overline{y}_{...})^2$	$N-1$		

11.4 Beispiele

Beispiel 11.2. ***chemosphere*** **Daten** Octocrylene ist ein organischer UV Filter, welcher in Sonnencremen und als Zusatzstoff in kosmetischen Artikeln enthalten ist. Die Substanz ist eingestuft als reizend und umweltgefährlich (EU-Gefahrstoffkennzeichnung).
Da sich der Stoff schwer abbaut, kann die Octocrylene (Umwelt-)Belastung durch Messungen im Klärschlamm von Kläranlagen geschätzt werden.
Die Studie Plagellat *et al.* (2006) analysierte Octocrylene (`OC`) Konzentrationen aus 24 verschiedenen Kläranlagen (bestehend aus drei verschiedenen Typen `Behandlung`) mit je zwei Proben (`Monat`). Zusätzlich sind Einzugsbereich (`Einwohner`) und Klärschlammmenge (`Produktion`) bekannt. Behandlungstyp `A` sind kleine Anlagen, `B` sind mittelgrosse Anlagen ohne nennenswerte Industrie und `C` mittelgrosse Anlagen mit Industriegewerbe.
R-Code 11.3 bereitet die Daten auf und zeigt eine einfaktorielle ANOVA. R-Code 11.4 zeigt eine zweifaktorielle ANOVA (ohne und mit Interaktion).
Abbildung 11.2 zeigt, warum die Wechselwirkung nicht signifikant ist. Zum ersten scheint der saisonale Effekt der Gruppen A und B sehr ähnlich und zum anderen ist die Variabilität in Gruppe C zu gross. ♣

R-Code 11.3: *UVfilter* Daten: einfaktorielle ANOVA mit `lm`. (See Figure 11.1.)

```
UV <- read.csv('./data/chemosphere.csv')
UV <- UV[,c(1:6,10)]    # reduce to OT
str(UV, strict.width='cut')
## 'data.frame': 24 obs. of  7 variables:
##  $ Behandlung   : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 2 2 ..
##  $ Standort_code: Factor w/ 12 levels "A11","A12","A15",..: 1 2 ..
##  $ Standort     : Factor w/ 12 levels "Chevilly","Cronay",..: 1 ..
##  $ Monat        : Factor w/ 2 levels "jan","jul": 1 1 1 1 1 1 1 ..
##  $ Einwohner    : int  210 284 514 214 674 5700 8460 11300 6500 ..
##  $ Produktion   : num  2.7 3.2 12 3.5 13 80 150 220 80 250 ...
##  $ OT           : int  1853 1274 1342 685 1003 3502 4781 3407 11..
with( UV, table(Behandlung, Monat))
##           Monat
## Behandlung jan jul
##          A   5   5
##          B   3   3
##          C   4   4
options(contrasts=c("contr.sum","contr.sum"))
lmout <- lm( log(OT) ~ Behandlung, data=UV)
summary(lmout)
##
## Call:
## lm(formula = log(OT) ~ Behandlung, data = UV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9524 -0.3468 -0.1359  0.3432  1.2612
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.1216     0.1158  70.154  < 2e-16 ***
## Behandlung1  -0.6398     0.1538  -4.159 0.000445 ***
## Behandlung2   0.4378     0.1747   2.506 0.020490 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.555 on 21 degrees of freedom
## Multiple R-squared:  0.4539,Adjusted R-squared:  0.4019
## F-statistic: 8.728 on 2 and 21 DF,  p-value: 0.001742
boxplot(log(OT)~Behandlung, data=UV)
```

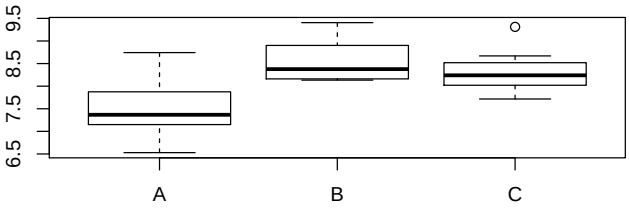


Abbildung 11.1: *UVfilter* Daten: Boxplots nach Behandlung sortiert. (Siehe R-Code 11.3.)

R-Code 11.4: *UVfilter* Daten: zweifaktorielle ANOVA und zweifaktorielle ANOVA mit Interaktion mit `lm`. (See Figure 11.2.)

```
lmout <- lm( log(OT) ~ Behandlung+Monat, data=UV)
summary(lmout)
##
## Call:
## lm(formula = log(OT) ~ Behandlung + Monat, data = UV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71749 -0.34517 -0.01236  0.16913  1.22361
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.1216     0.1058  76.775  < 2e-16 ***
## Behandlung1  -0.6398     0.1406  -4.551 0.000194 ***
## Behandlung2   0.4378     0.1596   2.743 0.012536 *
## Monat1       -0.2349     0.1035  -2.270 0.034442 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.5071 on 20 degrees of freedom
## Multiple R-squared:  0.5658,Adjusted R-squared:  0.5006
## F-statistic: 8.686 on 3 and 20 DF,  p-value: 0.0006858
summary( aovout <- aov( log(OT) ~ Behandlung*Monat, data=UV))
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## Behandlung        2  5.376  2.6880  11.673 0.000562 ***
## Monat             1  1.325  1.3246   5.752 0.027519 *
## Behandlung:Monat  2  0.998  0.4989   2.166 0.143552
## Residuals        18  4.145  0.2303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
TukeyHSD( aovout, which=c('Behandlung'))
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log(OT) ~ Behandlung * Monat, data = UV)
##
## $Behandlung
##           diff        lwr       upr     p adj
## B-A  1.0776023  0.4451582 1.7100464 0.0010732
## C-A  0.8417391  0.2608021 1.4226761 0.0044606
## C-B -0.2358632 -0.8972891 0.4255627 0.6410487
boxplot(log(OT)~Behandlung, data=UV, col=7, boxwex=.5)
at <- c(0.7, 1.7, 2.7, 1.3, 2.3, 3.3)
boxplot(log(OT)~Behandlung+Monat, data=UV, add=T, at=at, xaxt='n',
        boxwex=.2)
with(UV, interaction.plot(Behandlung, Monat, log(OT),col=2:3))
```

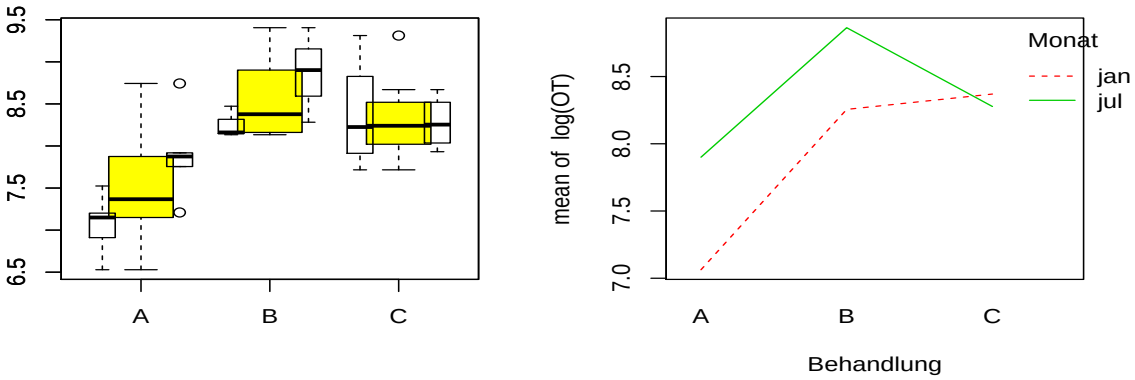


Abbildung 11.2: *UVfilter* Daten: Boxplots und Wechselwirkungsplot (Siehe R-Code 11.4.)

Kapitel 12

Bayes'sche Methoden

Es existieren in der Statistik zwei philosophisch verschiedene Ansätze zur Inferenz: frequentistische und Bayes'sche Inferenz. Die letzten Kapitel haben sich mit dem frequentistischen Ansatz befasst. Im Bayes'schen Ansatz betrachten wir die Parameter als Zufallsvariablen mit einer angebrachten Verteilung, die man a priori, d.h. vor der Datenerhebung festlegt. Das Ziel ist diese Verteilung nach Beobachtung der Daten zu aktualisieren, um dann mit Hilfe dieser a posteriori Verteilung Schlüsse ziehen zu können.

12.1 Terminologie

Die Dichte des Parameters wird Priori-Dichte genannt und ensprechend Priori-Verteilung (*prior probability*, *prior density*, *prior distribution*). Die Dichte der Daten wird Likelihood (*likelihood*) oder Beobachtungsmodell genannt.
Die aktualisierte Dichte wird Posteriori-Dichte genannt (*posterior probability*, *posterior density*, *posterior distribution*).
Grundsätzlich gilt: Die Posteriori-Dichte ist proportional zum Likelihood multipliziert mit der Priori-Dichte, dass heisst

$$\text{Posteriori-Dichte} \propto \text{Likelihood} \times \text{Priori-Dichte} \tag{12.1}$$

Dies kann auf verschiedene Weise ausgedrückt werden:

$$\begin{aligned} f_{\Theta|Y}(\theta \mid y) &\propto f_Y(y) f_\Theta(\theta), \\ f(\theta \mid y) &\propto f(y \mid \theta) f(\theta). \end{aligned}$$

Das Symbol "$\propto$" heisst "proportional zu" und wird verwendet, weil die Normalisierungskonstante nicht berücksichtigt wird:

$$f(\theta \mid y) = \frac{f(y, \theta)}{f(y)} = \frac{f(y \mid \theta) f(\theta)}{f(y)} \propto f(y \mid \theta) f(\theta).$$

Wichtig ist, dass die Normalisierungskonstante frei vom Parameter θ ist.

12.2 Beispiele

Beispiel 12.1. (Beta-Binomial) Sei $Y \sim \mathcal{B}in(n,p)$. Wir beobachten y Erfolge (aus n). Wie im Abschnitt 5.1 gezeigt wurde, $\widehat{p} = y/n$. Wir haben aber oft zusätzliches Wissen über den Parameter p. Z.B. seien die Herbstlämmer einer grossen Schafherde, $n = 150$. Wir zählen die Anzahl der männlichen Lämmer. Es ist somit höchst unwahrscheinlich, dass $p \leq 0.1$. Wir nehmen daher an, dass p betaverteilt ist, d.h.

$$f(p) = c \cdot p^{\alpha-1}(1-p)^{\beta-1}, \qquad p \in [0,1], \alpha > 0, \beta > 0 \tag{12.2}$$

mit c der Normalisierungskonstante. Wir schreiben $p \sim \mathcal{B}eta(\alpha,\beta)$. Abbildung 2.8 zeigt Dichten für verschiedene Paare (α,β).
Die Posterori-Dichte ist somit

$$\propto \binom{n}{y} p^y (1-p)^{n-y} \cdot c\, p^{\alpha-1}(1-p)^{\beta-1} \tag{12.3}$$

$$\propto p^y p^{\alpha-1}(1-p)^{n-y}(1-p)^{\beta-1} \tag{12.4}$$

$$\propto p^{y+\alpha-1}(1-p)^{n-y+\beta-1}, \tag{12.5}$$

welche als eine Betadichte $\mathcal{B}eta(y+\alpha, n-y+\beta)$ identifiziert wird.
Der Erwartungswert einer betaverteilten Zufallsvariablen $\mathcal{B}eta(\alpha,\beta)$ ist $\alpha/(\alpha+\beta)$ (hier die Priori-Dichte). Der Posteriori-Erwartungswert ist somit

$$\mathrm{E}(p \mid Y = y) = \frac{y+\alpha}{n+\alpha+\beta} \tag{12.6}$$

Der Fall $\mathcal{B}eta(1,1)$ entspricht einer Gleichverteilung $\mathcal{U}(0,1)$. Diese Gleichverteilung entspricht aber nicht "informationsfrei". Aus der Gleichung (12.6) geht hervor, dass eine Gleichverteilung als Priori-Verteilung äquivalent ist, zu zwei zusätzlichen Experimenten, von denen eines ein Erfolg ist, d.h., eins von zwei Lämmer ist männlich. □

Beispiel 12.2. (Normal-Normal) Seien $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Nehmen wir an, dass $\mu \sim \mathcal{N}(\eta, \tau^2)$ und σ bekannt. Somit haben wir das Bayes'sche Modell:

$$Y_i \mid \mu \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \qquad i = 1, \ldots, n, \tag{12.7}$$

$$\mu \mid \eta \sim \mathcal{N}(\eta, \tau^2). \tag{12.8}$$

Die Posteriori-Dichte ist somit

$$f(\mu \mid y_1, \ldots, y_n) \propto f(y_1, \ldots, y_n \mid \mu) f(\mu) \tag{12.9}$$

$$\propto \prod_{i=1}^{n} f(y_i \mid \mu) f(\mu) \tag{12.10}$$

$$\propto \prod_{i=1}^{n} \exp\Big(-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}\Big) \exp\Big(-\frac{1}{2}\frac{(\mu-\eta)^2}{\tau^2}\Big) \tag{12.11}$$

$$\propto \exp\Big(-\frac{1}{2}\sum_i \frac{(y_i-\mu)^2}{\sigma^2} - \frac{1}{2}\frac{(\mu-\eta)^2}{\tau^2}\Big), \tag{12.12}$$

wobei die Konstanten $(2\pi\sigma^2)^{-1/2}$ und $(2\pi\tau^2)^{-1/2}$ nicht berücksichtigt werden müssen. Durch Komplettieren des Quadrates in μ erhält man

$$\propto \exp\left(-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\mu-\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{n\overline{y}}{\sigma^2}+\frac{\eta}{\tau^2}\right)\right]^2\right) \tag{12.13}$$

und somit ist die Posteriori-Verteilung

$$\mathcal{N}\left(\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{n\overline{y}}{\sigma^2}+\frac{\eta}{\tau^2}\right),\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\right). \tag{12.14}$$

In anderen Worten, der Posteriori-Erwartungswert

$$\mathrm{E}(\mu \mid y_1,\ldots,y_n) = \eta\frac{\sigma^2}{n\tau^2+\sigma^2}+\overline{y}\frac{n\tau^2}{n\tau^2+\sigma^2} \tag{12.15}$$

ist ein gewichtetes Mittel zwischen dem Priori-Mittel η und dem "Likelihood"-Mittel $\overline{y}$. Je grösser n, je weniger fällt das Priori-Mittel ins Gewicht, da $\sigma^2/(n\tau^2+\sigma^2) \to 0$ für $n \to \infty$. □

Als Kennzahl der Posteriori-Verteilung wird oft der Posteriori-Modus verwendet. Natürlich sind der Posteriori-Median und Posteriori-Erwartungswert intuitive Alternativen. Im frequentistischen Ansatz haben wir Konfidenzintervalle konstruiert, die aber nicht anhand von Wahrscheinlichkeiten interpretiert werden. D.h. das $(1-\alpha)$%-Konfidenzintervall $[b_u, b_o]$ beinhaltet mit einer Häufigkeit von $(1-\alpha)$% den wahren Parameter bei einer unendlichen Wiederholung des Zufallsexperiments. Mit einem Bayes'schen Ansatz können wir jetzt Aussagen über die Parameter mit Wahrscheinlichkeiten machen. Im Beispiel 12.2, basierend auf (12.14) ist

$$\mathrm{P}\left(v^{-1}\left(\frac{n\overline{y}}{\sigma^2}+\frac{\eta}{\tau^2}\right)-z_{1-\alpha/2}v^{-1/2} \le \mu \le v^{-1}\left(\frac{n\overline{y}}{\sigma^2}+\frac{\eta}{\tau^2}\right)+z_{1-\alpha/2}v^{-1/2}\right)=1-\alpha, \tag{12.16}$$

mit $v = \frac{n}{\sigma^2}+\frac{1}{\tau^2}$.

Auf diesen Überlegungen wird das Intervall R mit

$$\int_R f(\theta \mid y_1,\ldots,y_n)d\theta = 1-\alpha \tag{12.17}$$

$(1-\alpha)$-Kredibilitätsintervall für θ bezüglich der Posteriori-Dichte $f(\theta \mid y_1,\ldots,y_n)$ genannt. $1-\alpha$ ist das Kredibilitätsniveau oder der Glaubwürdigkeitsgrad des Intervalls.
Da das Kredibilitätsintervall für ein festes α nicht eindeutig bestimmt ist, wird oft das "kürzeste" benutzt. Dieses ist das sogenannte HPD-Intervall (*highest posterior density interval*). Eine ausführliche Diskussion ist in Held (2008) zu finden. Die Kredibilitätsintervalle werden oft numerisch bestimmt.

12.3 Wahl der Priori-Verteilung

Die Wahl der Priori-Verteilung gehört zur Modellwahl wie die Wahl der Likelihood-Verteilung.
Die Beispiele im letzten Abschnitt waren so, dass die Posteriori- und die Priori-Verteilung zur selben Klasse gehörten. Das ist natürlich kein Zufall, da wir die für den Likelihood sogenannte konjugierte Priori-Verteilung gewählt haben.

Mit anderen Priori-Verteilung haben wir unter Umständen Verteilungen, die wir nicht mehr "erkennen" und gegebenenfalls eine Normalisierungskonstante explizit berechnen müssten.

Die Wahl der Priori-Verteilung führt immer wieder zu Diskussionen, und wir verweisen hier auf Held (2008) für mehr Details.

Kapitel 13

Vertiefung: Monte-Carlo Methoden

Mehrere Beispiele im letzten Kapitel waren so, dass die Posteriori- und die Priori-Verteilung dieselbe Verteilung (mit unterschiedlichen Parametern) hatten. Das ist natürlich kein Zufall, da wir die für die Likelihood sogenannte konjugierte Priori-Verteilung gewählt haben.
Mit anderen Priori-Verteilung haben wir unter Umständen unbekannte, "komplizierte" Posteriori-Verteilungen, von denen wir beispielsweise den Erwartungswert nicht mehr kennen. Oft ist dessen Berechnung (Integral) zu komplex, und so betrachten wir hier klassische (Monte-Carlo) Simulationsansätze, um dieses Problem zu lösen.
Im allgemeinen wird bei einer Monte-Carlo-Simulation (*Monte Carlo simulation*) durch eine grosse Anzahl von Realisationen gleicher Zufallszahlen ein komplexes Problem numerisch gelöst. Dabei wird vor allem das Gesetz der grossen Zahlen ausgenutzt.

13.1 Monte Carlo Integration

Wir betrachten zuerst, wie wir auf einfache Art Integrale annähern können. Sei $f_X(x)$ eine Dichte und $g(x)$ eine "beliebige" Funktion. Gesucht ist der Erwartungswert von $g(X)$, d.h.,

$$\mathrm{E}\left(g(X)\right) = \int_{\mathbb{R}} g(x) f_X(x)\,\mathsf{d}x. \tag{13.1}$$

Eine Annäherung dieses Integrals ist (Momentenmethode)

$$\mathrm{E}\left(g(X)\right) \int_{\mathbb{R}} g(x) f_X(x)\,\mathsf{d}x \approx \widehat{\mathrm{E}\left(g(X)\right)} = \frac{1}{n}\sum_{i=1}^{n} g(x_i), \tag{13.2}$$

wobei $x_1, \ldots, x_n$ eine Stichprobe aus $f_X(x)$ ist.

Beispiel 13.1. Sei $g(x_1, x_2)$ die Indikatorfunktion mit Menge $x_1^2 + x_2^2 \leq 1$. Wir haben folgende Annäherung der Zahl π (Fläche eines Kreises mit Radius eins)

$$\pi = \int_{x_1^2+x_2^2\leq 1} g(x_1, x_2)\,\mathsf{d}x_1\,\mathsf{d}x_2 \approx 4 \times \frac{1}{n}\sum_{i=1}^{n} g(x_{1,i}, x_{2,i}), \tag{13.3}$$

wobei $x_{1,i}$ und $x_{2,i}$, $i = 1 \ldots, n$ zwei Stichproben aus $U(0, 1)$ sind.
Es ist zu beachten, dass die Konvergenz recht langsam ist, siehe Abbildung 13.1. ♣

R-Code 13.1 Annäherung von π mit Hilfe von Monte Carlo Integration (Siehe Abbildung 13.1.)

```
set.seed(14)
m <- 49
n <- round( 10+1.4^(1:m))
piapprox <- numeric(m)
for (i in 1:m) {
   st <- matrix( runif( 2*n[i]), ncol=2)
   piapprox[i] <- 4*mean( rowSums( st^2)<= 1)
}
plot( n, abs( piapprox-pi)/pi, log='xy', type='l')
lines( n, 1/sqrt(n), col=2, lty=2)
cbind( n=n[1:7*7], pi.approx=piapprox[1:7*7], rel.error=
      abs( piapprox[1:7*7]-pi)/pi, abs.error=abs( piapprox[1:7*7]-pi))
##             n pi.approx    rel.error    abs.error
## [1,]       21  2.476190 0.2118040914 0.6654021774
## [2,]      121  3.008264 0.0424396812 0.1333281908
## [3,]     1181  3.163421 0.0069481243 0.0218281762
## [4,]    12358  3.166208 0.0078353477 0.0246154707
## [5,]   130171  3.140331 0.0004016619 0.0012618579
## [6,]  1372084  3.142408 0.0002595935 0.0008155372
## [7,] 14463522  3.140567 0.0003265614 0.0010259230
```

13.2 Verwerfungsmethode

Wenn keine Methode existiert, Realisationen direkt aus einer Verteilung mit gegebener Dichtefunktion $f_Y(y)$ zu ziehen, kann die Verwerfungsmethode (*rejection sampling*) angewendet werden. Dabei werden Werte von einer bekannten Dichte $f_Z(z)$ gezogen und durch Verwerfen von unpassenden Werten Realisationen von $f_Y(y)$ generiert. Die Verwerfungsmethode kann auch angewandt werden, wenn die Normalisierungskonstante von $f_Y(y)$ nicht bekannt ist. Wir schreiben $f_Y(y) = c \cdot f^*(y)$.
Das Vorgehen ist wie folgt: Finde ein $m < \infty$, so dass $f^*(y) \leq m \cdot f_Z(y)$. Dann ziehe $\tilde{y}$ von $f_Z(y)$ und u von einer standard Uniform-Verteilung. Falls $u \leq f^*(\tilde{y})/\big(m \cdot f_Z(\tilde{y})\big)$ wird $\tilde{y}$ als simulierter Wert von $f_Y(y)$ akzeptiert, sonst wird $\tilde{y}$ verworfen und nicht weiter beachtet. Aus Effizienzgründen sollte die Konstante m möglichst klein gewählt werden.

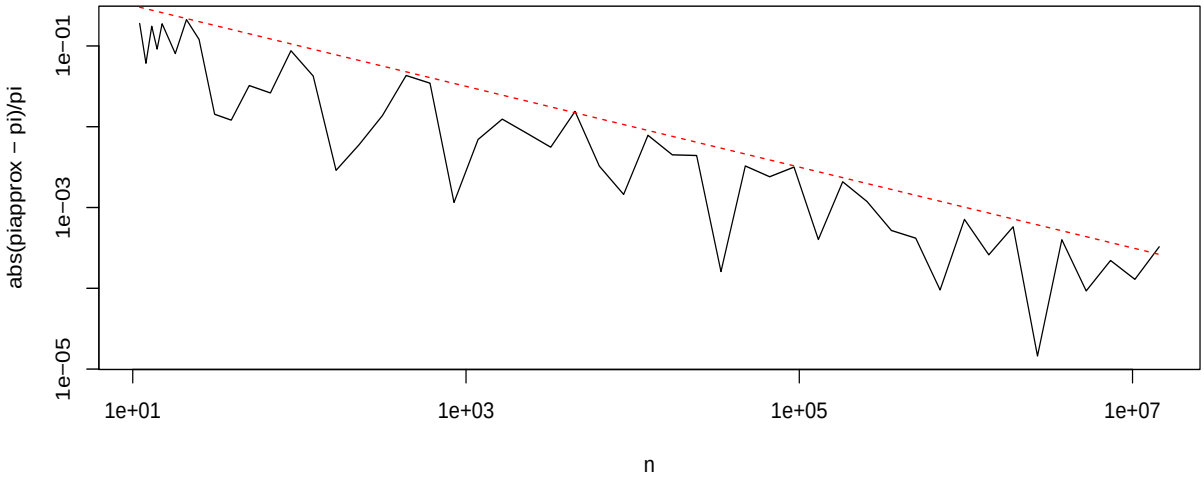


Abbildung 13.1: Konvergenz der Annäherung von π: der relative Fehler als Funktion von n. (Siehe R-Code 13.1.)

Beispiel 13.2. Wir ziehen Werte aus einer $\mathcal{B}eta(6,3)$ Verteilung mit der Verwerfungsmethode und brauchen dazu eine Uniformverteilung. Das heisst, $f_Y(y) = c \cdot y^{6-1}(1-y)^{3-1}$ (die Normalisierungskonstante c ist $1/\beta(6,3) = 1/168$) und $f_Z(y) = 1_{0\leq y\leq 1}(y)$. Wir wählen $m = 0.02$, was die Bedingung $f^*(y) \leq m \cdot f_Z(y)$ erfüllt.
Eine Implementation des Beispieles ist in R-Code 13.2 gegeben. Der R-Code kann bezüglich Geschwindigkeit optimiert werden. Er ist dann allerdings schwieriger zu lesen. Die Abbildung 13.2 zeigt ein Histogramm und die Dichte der simulierten Werte. ♣

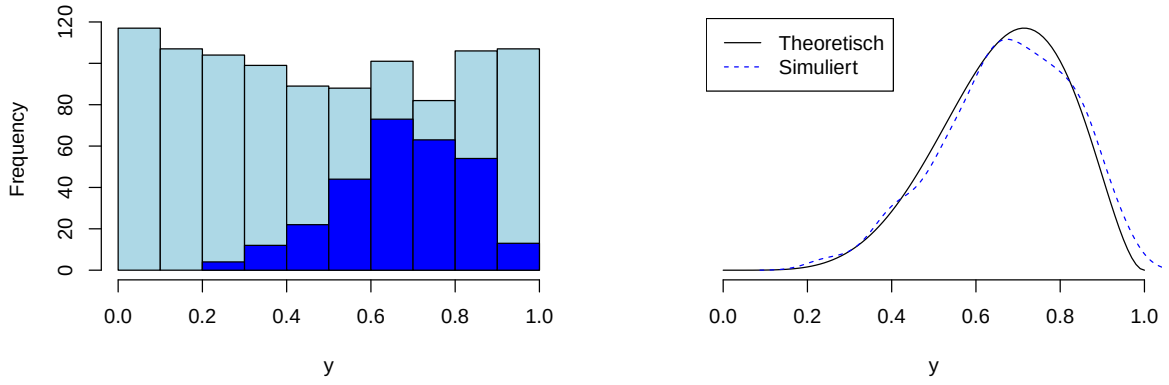


Abbildung 13.2: Links sind die Histogramme von den simulierten Werten von $f_Z(y)$ (hellblau) und $f_Y(y)$ (dunkelblau) dargestellt. Rechts sind die theoretische und die simulierte Dichten abgebildet. (Siehe R-code 13.2 und Beispiel 13.2.)

R-Code 13.2 Verwerfungsmethode (Siehe Abbildung 13.2)

```
set.seed( 14)
n.sim <- 1000
m <- 0.02
fst <- function(y) y^( 6-1) * (1-y)^(3-1)
f_Z <- function(y) ifelse( y >= 0 & y <= 1, 1, 0)
result <- sample <- rep( NA, n.sim)
for (i in 1:n.sim){
    sample[i] <- runif(1)
    u <- runif(1)
    if( u < fst( sample[i]) /( m * f_Z( sample[i])) ) # if accepted ...
        result[i] <- sample[i]
}
mean( !is.na(result))  # proportion of accepted samples
## [1] 0.285
result <- result[ !is.na(result)]
## Figures
hist( sample, xlab="y", main="", col="lightblue")
hist( result, add=TRUE, col=4)
curve( dbeta(x, 6, 3), frame =FALSE, ylab="", xlab='y', yaxt="n")
lines( density( result), lty=2, col=4)
legend( "topleft", legend=c("Theoretisch", "Simuliert"),
       lty=1:2, col=c(1,4))
```

13.3 Gibbs Sampling

Die Idee von Gibbs Sampling ist es, die Posteri-Verteilung mittels einer Markov Kette zu simulieren. Die gehört deshalb zu der Familie der Markov Ketten Monte Carlo Methoden (*Markov chain Monte Carlo, MCMC*). Wir illustrieren das Prinzip anhand einer gemeinsamen Verteilung $f(\theta_1, \theta_2 \mid y)$ mit den zwei Parameter θ_1 und θ_2. Der Gibbs Sampler reduziert das Problem auf zwei eindimensionale Simulationen. Anschliessend wird abwechslungsweise von $f(\theta_1 \mid \theta_2, y)$ und $f(\theta_2 \mid \theta_1, y)$ gezogen, wobei jeweils auf die vorgängig gezogenen Werte von θ_1 und θ_2 bedingt wird.

In vielen Fällen muss man die Gibbs sampler nicht selber programmieren und kann auf einen vorprogrammierten sampler zurückgreiffen. Wir benützen den Sampler JAGS (*Just Another Gibbs sampler*). Eine Alternative zu JAGS ist BUGS (*Bayesian inference Using Gibbs Sampling*) die in zwei Hauptversionen WinBUGS and OpenBUGS vertrieben wird. Zu allen gibt es R-Interface Pakete (`rjags`, `R2OpenBUGS`).

Bei den MCMC Methoden kann es vorkommen, dass der Sampler nicht oder noch nicht

konvergiert. In so einem Fall kann die Posteriori-Verteilung *nicht* mit den simulierten Werten approximiert werden. Es ist deshalb wichtig, die simulierten Werte auf auffällige Muster zu untersuchen. Dazu verwendet man häufig diagnostische Figuren, zum Beispiel den "Trace-plot" wie in Abbildungen 13.3 (links) dargestellt. Weitere Information zu diesen diagnostischen Figuren gibt es in Lunn *et al.* (2012). Hier findet sich zudem eine kurze Einführung in Bayesianische Statistik und zahlreiche Beispiele von (Open)BUGS Modellen.

Beispiel 13.3. Die folgenden drei R-Codes geben einen kleinen Einblick in die Markov-Ketten-Monte-Carlo Methode mit JAGS. R-Code 13.3 implementiert das Normal-normal Model mit $y = 1$, $n = 1$, mit bekannter Varianz, R-Code 13.4 das Normal-normal Model mit $n = 10$ mit bekannter Varianz und R-Code 13.5 das Norma-normal-gamma Model mit $n = 10$ und unbekannter Varianz. Die entsprechenden Abbildungen 13.3, 13.4 und 13.5 geben die empirischen und exakten Dichten der Posteriori, Priori und der Likelihood wieder. ♣

R-Code 13.3: JAGS sampler für Normal-normal Model, mit $n = 1$. (See Figure 13.3.)

```
require( rjags)
writeLines("model {
              y ~ dnorm( mu, 1)
              mu ~ dnorm( 0, 1/1.2)     # here Precision = 1/Variance
           }",   con="jags01.txt")

jagsModel <- jags.model( 'jags01.txt', data=list( 'y'=1))
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1
##    Unobserved stochastic nodes: 1
##    Total graph size: 7
##
## Initializing model
postSamples <- coda.samples( jagsModel, 'mu', n.iter=2000)

plot( postSamples, density=FALSE, main="", auto.layout=FALSE)
plot( postSamples, trace=FALSE, main="", auto.layout=FALSE,
      xlim=c(-2,3), ylim=c(0,.56))
```

```
rug( summary( postSamples )$quantiles, lwd=2, tick=.2)

y <- seq(-3,to=4, length=100)
lines( y, dnorm(y, 1, 1), col=3)
lines( y, dnorm(y, 0, sqrt(1.2)   ), col=4)
lines( y, dnorm(y, 1/(1+1/1.2), sqrt(1/(1+1/1.2))), col=2)
```

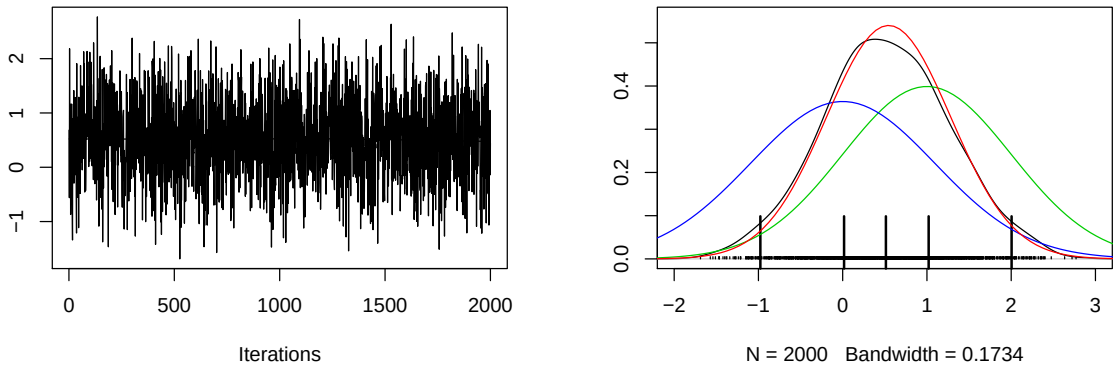


Abbildung 13.3: Links: trace-plot von der Posteriori $\mu \mid y = 1$. Rechts: Empirische Dichten des MCMC basierten Posteriors (schwarz), exakt (rot), Prori (blau), Likelihood (grün). (Siehe R-Code 13.3.)

R-Code 13.4: JAGS sampler für Normal-normal Model, mit $n = 10$. (See Figure 13.4.)

```
set.seed( 4)
n <- 10
obs <- rnorm( n, 1, 1)

writeLines("model {
              for (i in 1:n) {
                 y[i] ~ dnorm( mu, 1)
              }
              mu ~ dnorm( 0, 1/2)
           }",   con="jags02.txt")

jagsModel <- jags.model( 'jags02.txt', data=list('y'=obs, 'n'=n),
                         quiet=T)
postSamples <- coda.samples(jagsModel, 'mu', n.iter=2000)
```

```
plot( postSamples, density=FALSE, main="", auto.layout=FALSE)
plot( postSamples, trace=FALSE, main="", auto.layout=FALSE,
      xlim=c(-.5, 3), ylim=c(0, 1.3))
rug( obs, col=3)
rug( summary( postSamples )$quantiles, lwd=2, tick=.2)

y <- seq(-.7, to=3.5, length=100)
lines( y, dnorm( y, mean(obs), sqrt(1/n)), col=3)
lines( y, dnorm( y, 0, sqrt(2)   ), col=4)
lines( y, dnorm( y, n*mean(obs)/(n+1/2), sqrt(1/(n+1/2))), col=2)
```

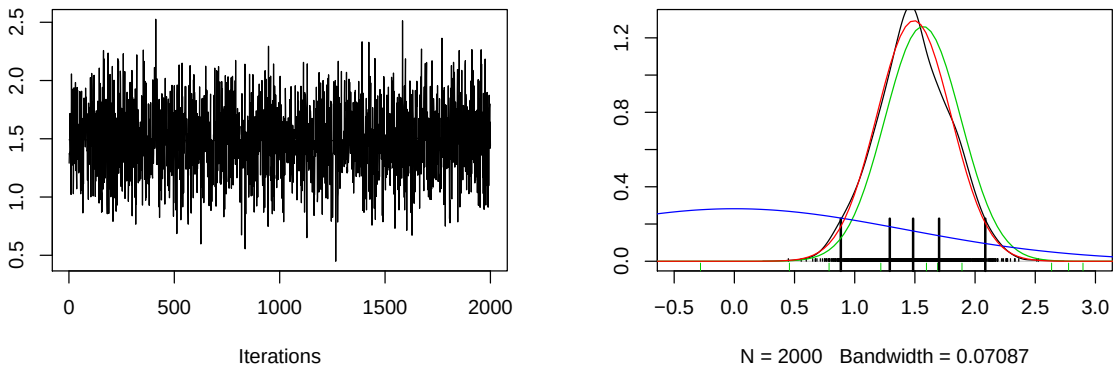


Abbildung 13.4: Links: trace-plot von der Posteriori $\mu \mid y = y_1, \ldots, y_n$. Rechts: Empirische Dichten des MCMC basierten Posteriors (schwarz), exakt (rot), Prori (blau), Likelihood (grün). (Siehe R-Code 13.4.)

R-Code 13.5: JAGS sampler für Normal-normal-gamma Model, mit $n = 10$. (See Figure 13.5.)

```
nu <- 0      # hyperparameters
tau2 <- 1.2

writeLines("model {
              for (i in 1:n) {
                 y[i] ~ dnorm( mu, kappa)
              }
              mu ~ dnorm( nu, 1/tau2)
              kappa ~ dgamma(1, .2)
```

```
            }",    con="jags03.txt")

jagsModel <- jags.model('jags03.txt', quiet=T,
                data=list('y'=obs, 'n'=n, 'nu'=nu, 'tau2'=tau2))

postSamples <- coda.samples(jagsModel, c('mu','kappa'), n.iter=2000)

plot( postSamples[,"mu"], trace=FALSE, auto.layout=FALSE,
      xlim=c(-1,3), ylim=c(0, 1.2))

y <- seq( -2, to=5, length=100)
lines( y, dnorm(y, mean(obs), sd(obs)/sqrt(n)), col=3)
lines( y, dnorm(y, 0, sqrt(1.2)   ), col=4)
plot( postSamples[,"kappa"], trace=FALSE, auto.layout=FALSE)
y <- seq( 0, to=5, length=100)
lines( y, dgamma( y, 1, .2), type='l', col=4)
lines( y, dgamma( y, n/2+1, n*var(obs)/2 ), col=3)
```

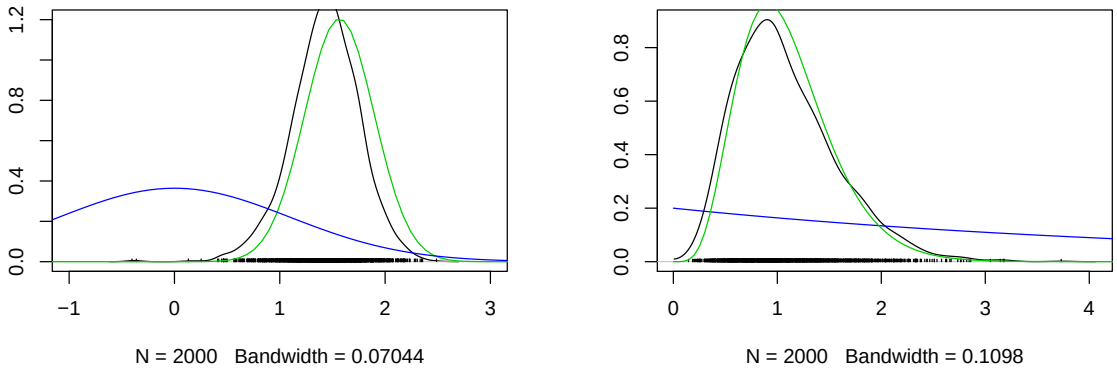


Abbildung 13.5: Empirische Posteriori-Dichten von $\mu \mid y_1, \ldots, y_n$ (links) und von $\kappa = 1/\sigma^2 \mid y_1, \ldots, y_n$ (rechts), MCMC basiert (schwarz), prori (blau), likelihood (grün). (Siehe R-Code 13.5.)

Literaturverzeichnis

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-building and Response Surfaces*. Wiley. 103

Cleveland, W. S. (1993). *Visualizing Data*. Hobart Press, Summit, New Jersey, U.S.A. 108

Dalal, S. R., Fowlkes, E. B., and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-challenger prediction of failure. *Journal of the American Statistical Association*, **84**, 945–957. 98

Fahrmeir, L., Kneib, T., and Lang, S. (2009). *Regression: Modelle, Methoden und Anwendungen*. Springer, 2 edition. 89

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, **17**, 347–388. 50

Furrer, R. and Genton, M. G. (1999). Robust spatial data analysis of lake Geneva sediments with S+SpatialStats. *Systems Research and Information Science*, **8**, 257–272. 118

Gemeinde Staldenried (1994). Verwaltungsrechnung 1993 Voranschlag 1994. 5

Held, L. (2008). *Methoden der statistischen Inferenz: Likelihood und Bayes*. Springer, Heidelberg. 139, 140

Hüsler, J. and Zimmermann, H. (2010). *Statistische Prinzipien für medizinische Projekte*. Huber, 5 edition. 29

Kupper, T., De Alencastro, L., Gatsigazi, R., Furrer, R., Grandjean, D., and J., T. (2008). Concentrations and specific loads of brominated flame retardants in sewage sludge. *Chemosphere*, **71**, 1173–1180. 127

Landesman, R., Aguero, O., Wilson, K., LaRussa, R., Campbell, W., and Penaloza, O. (1965). The prophylactic use of chlorthalidone, a sulfonamide diuretic, in pregnancy. *J. Obstet. Gynaecol.*, **72**, 1004–1010. 55

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Texts in Statistical Science. Chapman & Hall/CRC. 145

Petersen, K. B. and Pedersen, M. S. (2008). The Matrix Cookbook. Version 2008-11-14, http://matrixcookbook.com. 85

Plagellat, C., Kupper, T., Furrer, R., de Alencastro, L. F., Grandjean, D., and Tarradellas, J. (2006). Concentrations and specific loads of UV filters in sewage sludge originating from a monitoring network in Switzerland. *Chemosphere*, **62**, 915–925. 132

Rudolf, M. and Kuhlisch, W. (2008). *Biostatistik: Eine Einführung für Biowissenschaftler*. Pearson Studium. 40

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, **103**, 677–680. 1, 2

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, **46**, 234–240. 116

Index von statistischen Tests

Index von Datensätzen

Index englischer Begriffe