

ulm university universität
uulm

Angewandte Stochastik

Vorlesungsskript

Prof. Dr. Evgeny Spodarev

Ulm

Sommersemester 2023

Vorwort

Das vorliegende Skript der Vorlesung Angewandte Stochastik gibt eine Einführung in die Problemstellungen der Wahrscheinlichkeitstheorie und der Statistik für Studierende der nicht mathematischen (jedoch mathematisch arbeitenden) Studiengänge wie Elektrotechnik, Informatik, Physik, usw. Es entstand in den Jahren 2005–2013, in denen ich diesen Vorlesungskurs an der Universität Ulm mehrmals gehalten habe.

Ich bedanke mich bei Herrn Tobias Scheinert und Herrn Michael Wiedler für die Umsetzung meiner Vorlesungsnotizen in L^AT_EX.

Ulm, den 5. Juli 2013
Evgeny Spodarev

Inhaltsverzeichnis

Inhaltsverzeichnis	i
1 Einführung	1
1.1 Über den Begriff “Stochastik”	1
1.2 Geschichtliche Entwicklung der Stochastik	2
1.3 Typische Problemstellungen der Stochastik	5
2 Wahrscheinlichkeiten	6
2.1 Ereignisse	7
2.2 Wahrscheinlichkeitsräume	10
2.3 Beispiele	12
2.3.1 Klassische Definition der Wahrscheinlichkeiten	13
2.3.2 Geometrische Wahrscheinlichkeiten	19
2.3.3 Bedingte Wahrscheinlichkeiten	20
3 Zufallsvariablen	26
3.1 Definition und Beispiele	26
3.2 Verteilungsfunktion	27
3.3 Grundlegende Klassen von Verteilungen	30
3.3.1 Diskrete Verteilungen	30
3.3.2 Absolut stetige Verteilungen	34
3.3.3 Mischungen von Verteilungen	40
3.4 Verteilungen von Zufallsvektoren	41
3.5 Stochastische Unabhängigkeit	46
3.5.1 Unabhängige Zufallsvariablen	46
3.6 Funktionen von Zufallsvektoren	48
4 Momente von Zufallsvariablen	53
4.1 Erwartungswert	54
4.2 Varianz	57
4.3 Kovarianz und Korrelationskoeffizient	60
4.4 Höhere und gemischte Momente	61
4.5 Entropie	63
4.6 Ungleichungen	65

5 Grenzwertsätze	67
5.1 Gesetzte der großen Zahlen	67
5.1.1 Schwaches und Starkes Gesetz der großen Zahlen	68
5.1.2 Anwendung der Gesetze der großen Zahlen	69
5.2 Zentraler Grenzwertsatz	70
5.2.1 Klassischer zentraler Grenzwertsatz	71
5.2.2 Konvergenzgeschwindigkeit im zentralen Grenzwertsatz	73
5.2.3 Grenzwertsatz von Lindeberg	73
6 Monte–Carlo–Simulation von Zufallsvariablen	75
6.1 Pseudozufallszahlen	75
6.2 Inversionsmethode	78
6.3 Akzeptanz– und Verwerfungsmethode	79
6.4 Simulation der Normalverteilung	83
6.4.1 Akzeptanz– und Verwerfungsmethode für $N(0,1)$	84
6.4.2 Box–Muller-Transformation	85
6.4.3 Abgeschnittene Normalverteilung	87
6.5 Simulation von diskret verteilten Zufallsvariablen	88
6.6 Monte–Carlo– und Quasi–Monte–Carlo–Integration	91
7 Beschreibende Statistik	92
7.1 Typische Fragestellungen, Aufgaben und Ziele der Statistik .	92
7.2 Statistische Merkmale und ihre Typen	93
7.3 Statistische Daten und Stichproben	94
7.4 Stichprobenfunktionen	95
7.5 Verteilungen und ihre Darstellungen	96
7.5.1 Häufigkeiten und Diagramme	96
7.5.2 Empirische Verteilungsfunktion	98
7.6 Beschreibung von Verteilungen	99
7.6.1 Lagemaße	100
7.6.2 Streuungsmaße	104
7.6.3 Maße für Schiefe und Wölbung	106
7.7 Quantilplots (Quantil-Grafiken)	107
7.8 Dichteschätzung	111
7.9 Beschreibung und Exploration von bivariaten Datensätzen .	113
7.9.1 Zusammenhangsmaße	113
7.9.2 Einfache lineare Regression	117
8 Punktschätzer	125
8.1 Parametrisches Modell	125
8.2 Parametrische Familien von statistischen Prüfverteilungen .	126
8.2.1 Gamma-Verteilung	126
8.2.2 Student-Verteilung (t-Verteilung)	130
8.2.3 Fisher-Snedecor-Verteilung (F-Verteilung)	131

8.3	Punktschätzer und ihre Grundeigenschaften	131
8.3.1	Eigenschaften von Punktschätzern	131
8.3.2	Schätzer des Erwartungswertes und empirische Momente	133
8.3.3	Schätzer der Varianz	135
8.3.4	Eigenschaften der Ordnungsstatistiken	138
8.3.5	Empirische Verteilungsfunktion	140
8.4	Methoden zur Gewinnung von Punktschätzern	143
8.4.1	Momentenschätzer	143
8.4.2	Maximum-Likelihood-Schätzer	145
8.4.3	Bayes-Schätzer	150
8.4.4	Resampling-Methoden zur Gewinnung von Punktschätzern	152
8.5	Weitere Güteeigenschaften von Punktschätzern	157
8.5.1	Ungleichung von Cramér-Rao	157
9	Konfidenzintervalle	163
9.1	Einführung	163
9.2	Ein-Stichproben-Probleme	165
9.2.1	Normalverteilung	165
9.2.2	Konfidenzintervalle aus stochastischen Ungleichungen	167
9.2.3	Asymptotische Konfidenzintervalle	169
9.3	Zwei-Stichproben-Probleme	171
9.3.1	Normalverteilte Stichproben	172
10	Tests statistischer Hypothesen	174
10.1	Allgemeine Philosophie des Testens	174
10.2	Nichtrandomisierte Tests	184
10.2.1	Parametrische Signifikanztests	184
10.3	Randomisierte Tests	186
10.3.1	Grundlagen	187
10.3.2	Neyman-Pearson-Tests bei einfachen Hypothesen . .	188
10.3.3	Einseitige Neyman-Pearson-Tests	194
10.3.4	Unverfälschte zweiseitige Tests	200
10.4	Anpassungstests	206
10.4.1	χ^2 -Anpassungstest	206
11	Lineare Regression	213
11.1	Multivariate Normalverteilung	214
11.1.1	Eigenschaften der multivariaten Normalverteilung .	218
11.1.2	Lineare und quadratische Formen von normalverteilten Zufallsvariablen	219
11.2	Multivariate lineare Regressionsmodelle mit vollem Rang .	227
11.2.1	Methode der kleinsten Quadrate	227

11.2.2 Schätzer der Varianz σ^2	233
11.2.3 Maximum-Likelihood-Schätzer für β und σ^2	234
11.2.4 Tests für Regressionsparameter	237
11.2.5 Konfidenzbereiche	240
11.3 Multivariate lineare Regression mit $\text{Rang}(X) < m$	244
11.3.1 Verallgemeinerte Inverse	244
11.3.2 MKQ-Schätzer für β	246
11.3.3 Erwartungstreue schätzbare Funktionen	248
11.3.4 Normalverteilte Störgrößen	252
11.3.5 Hypothesentests	255
11.3.6 Konfidenzbereiche	257
11.3.7 Einführung in die Varianzanalyse	259
Literatur	262
Index	275

Kapitel 1

Einführung

1.1 Über den Begriff “Stochastik”

Wahrscheinlichkeitsrechnung ist eine Teildisziplin von Stochastik. Dabei kommt das Wort “Stochastik” aus dem Griechischen $\sigma\tau\omega\chi\alpha\sigma\tau\iota\kappa\eta$ - “die Kunst des Vermutens” (von $\sigma\tau\omega\chi\omega\xi$ - “Vermutung, Ahnung, Ziel”). Dieser Begriff wurde von Jacob Bernoulli in seinem Buch “Ars conjectandi” geprägt (1713), in dem das erste Gesetz der großen Zahlen bewiesen wurde.

Stochastik beschäftigt sich mit den Ausprägungen und quantitativen Merkmalen von Zufall. Aber was ist Zufall? Gibt es Zufälle überhaupt? Das ist eine philosophische Frage, auf die jeder seine eigene Antwort suchen muss. Für die moderne Mathematik ist der Zufall eher eine Arbeitshypothese, die viele Vorgänge in der Natur und in der Technik ausreichend gut zu beschreiben scheint. Insbesondere kann der Zufall als eine Zusammenwirkung mehrerer Ursachen aufgefasst werden, die sich dem menschlichen Verstand entziehen (z.B. Brownsche Bewegung). Andererseits gibt es Studienbereiche (wie z.B. in der Quantenmechanik), in denen der Zufall eine essentielle Rolle zu spielen scheint (die Unbestimmtheitsrelation von Heisenberg). Wir werden die Existenz des Zufalls als eine wirkungsvolle Hypothese annehmen, die für viele Bereiche des Lebens zufriedenstellende Antworten liefert.



Abbildung 1.1:
Jacob Bernoulli
(1654-1705)

Stochastik kann man in folgende Gebiete unterteilen:

- Wahrscheinlichkeitsrechnung oder Wahrscheinlichkeitstheorie (Grundlagen)
- Statistik (Umgang mit den Daten)
- Stochastische Prozesse (Theorie zufälliger Zeitreihen und Felder)

- Simulation

Diese Vorlesung ist nur dem ersten Teil gewidmet.

1.2 Geschichtliche Entwicklung der Stochastik

1. Vorgeschichte:

Die Ursprünge der Wahrscheinlichkeitstheorie liegen im Dunklen der alten Zeiten. Ihre Entwicklung ist in der ersten Phase den Glücksspielen zu verdanken. Die ersten Würfelspiele konnte man in Altägypten, I. Dynastie (ca. 3500 v. Chr.) nachweisen. Auch später im klassischen Griechenland und im römischen Reich waren solche Spiele Mode (Kaiser August (63 v. Chr. -14 n. Chr.) und Claudius (10 v.Chr. - 54 n. Chr.).

Gleichzeitig gab es erste wahrscheinlichkeitstheoretische Überlegungen in der Versicherung und im Handel. Die älteste uns bekannte Form der Versicherungsverträge stammt aus dem Babylon (ca. 4-3 T. J. v.Chr., Verträge über die Seetransporte von Gütern). Die ersten Sterbetafeln in der Lebensversicherung stammen von dem römischen Juristen Ulpian (220 v.Chr.). Die erste genau datierte Lebensversicherungspolice stammt aus dem Jahre 1347, Genua.

Der erste Wissenschaftler, der sich mit diesen Aufgabenstellungen aus der Sicht der Mathematik befasst hat, war G. Cardano, der Erfinder der Cardan-Welle. In seinem Buch "Liber de ludo alea" sind zum ersten Mal Kombinationen von Ereignissen beim Würfeln beschrieben, die vorteilhaft für den Spieler sind. Er hat auch als erster die



Abbildung 1.2:
Gerolamo Cardano
(1501-1576)

$$\frac{\text{Anzahl vorteilhafter Ereignisse}}{\text{Anzahl aller Ereignisse}}$$

als Maß für Wahrscheinlichkeit entdeckt.

2. Klassische Wahrscheinlichkeiten (XVII-XVIII Jh.):

Diese Entwicklungsperiode beginnt mit dem Briefwechsel zwischen Blaise Pascal und Pierre de Fermat. Sie diskutierten Probleme, die von Chevalier de Méré (Antoine Gombaud (1607-1684)) gestellt wurden. Anbei ist eines seiner Probleme:

Was ist wahrscheinlicher: mindestens eine 6 bei 4 Würfen eines Würfels oder mindestens ein Paar (6,6) bei 24 Würfen von 2 Würfeln zu bekommen?

$$\begin{aligned} A &= \{\text{mind. eine 6 bei 4 Würfen eines Würfels}\} \\ B &= \{\text{mind. (6,6) bei 24 Würfen von zwei Würfeln}\} \end{aligned}$$

Die Antwort:

$$\begin{aligned} P(\text{mind. eine 6 in 4 Würfen}) \\ = P(A) = 1 - P(\overline{A}) = 1 - P(\text{keine 6}) = 1 - \left(\frac{5}{6}\right)^4 = 0,516 > 0,491 = 1 - \left(\frac{35}{36}\right)^{24} \\ = 1 - P(\overline{B}) = P(B) = P(\text{mind. 1 (6,6) in 24 Würfen von 2 Würfeln}) \end{aligned}$$



Abbildung 1.3: Blaise Pascal (1623-1662), Pierre de Fermat (1601-1665) und Christian Huygens (1629-1695)

Weitere Entwicklung:

1657	Christian Huygens “De Ratiociniis in Ludo Alea” (Operationen mit Wahrscheinlichkeiten)
1713	Jacob Bernoulli “Ars Conjectandi” (Wahrscheinlichkeit eines Ereignisses und Häufigkeit seines Eintretens)

3. *Entwicklung analytischer Methoden (XVIII-XIX Jh.)* von Abraham de Moivre, Thomas Bayes (1702-1761), Pierre Simon de Laplace, Carl Friedrich Gauß, Simeon Denis Poisson (vgl. Abb. 1.4).

Entwicklung der Theorie bezüglich Beobachtungsfehlern und der Theorie des Schießens (Artilleriefeuer). Erste nicht-klassische Verteilungen wie Binomial- und Normalverteilung ($f(x) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x^2}{2}\right)}$, $x \in \mathbb{R}$), Poisson-Verteilung, zentraler Grenzwertsatz von De Moivre.

St.-Petersburger Schule von Wahrscheinlichkeiten:

(P.L. Tschebyschew, A.A. Markow, A.M. Ljapunow)

- Einführung von Zufallsvariablen, Erwartungswerten, Wahrscheinlichkeitsfunktionen, Markow-Ketten, abhängigen Zufallsvariablen.

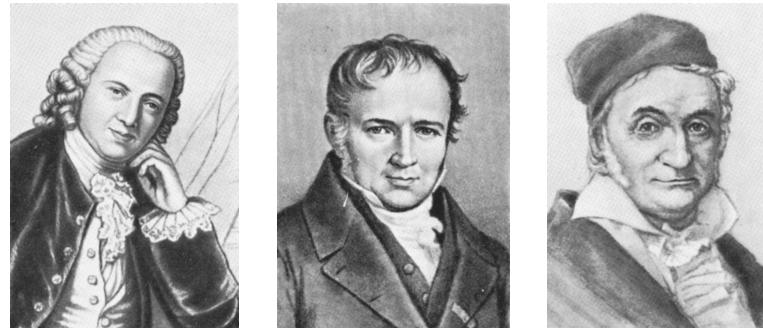


Abbildung 1.4: Abraham de Moivre (1667-1754), Pierre Simon de Laplace (1749-1827) und Karl Friedrich Gauß (1777-1855)



Abbildung 1.5: Simeon Denis Poisson (1781-1840), P. L. Tschebyschew (1821-1894) und A. A. Markow (1856-1922)

4. *Moderne Wahrscheinlichkeitstheorie (XX Jh.)* David Hilbert, 8.8.1900, *II. Mathematischer Kongress in Paris, Problem Nr. 6:*
Axiomatisierung von physikalischen Disziplinen, wie z.B. Wahrscheinlichkeitstheorie.
R. v. Mises: frequentistischer Zugang: $P(A) = \lim_{n \rightarrow \infty} \frac{\#A \text{ in } n \text{ Versionen}}{n}$
Antwort darauf: A.N. Kolmogorow führt Axiome der Wahrscheinlichkeitstheorie basierend auf der Maß- und Integrationstheorie von Borel und Lebesgue (1933) ein.

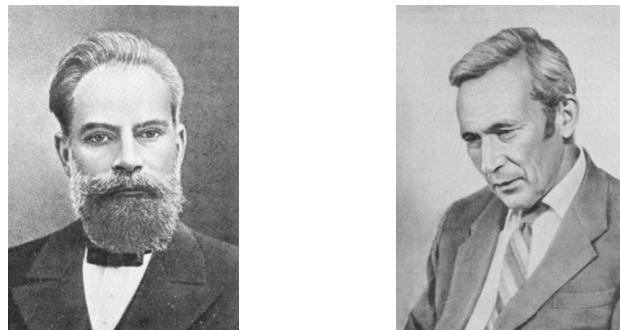
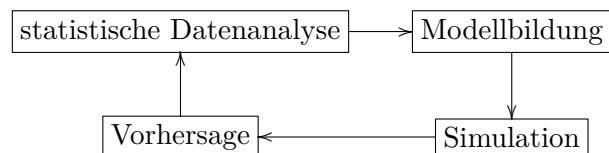


Abbildung 1.6: A. M. Ljapunow (1857-1918) und A. H. Kolmogorow (1903-1987)

1.3 Typische Problemstellungen der Stochastik



1. Modellierung von Zufallsexperimenten, d.h. deren adäquate theoretische Beschreibung.
2. Bestimmung von
 - Wahrscheinlichkeiten von Ereignissen
 - Mittelwerten und Varianzen von Zufallsvariablen
 - Verteilungsgesetzen von Zufallsvariablen
3. Näherungsformel und Lösungen mit Hilfe von Grenzwertsätzen
4. Schätzung von Modellparametern in der Statistik, Prüfung statistischer Hypothesen

Kapitel 2

Wahrscheinlichkeiten

Wahrscheinlichkeitstheorie befasst sich mit (im Prinzip unendlich oft) wiederholbaren Experimenten, in Folge derer ein Ereignis auftreten kann (oder nicht). Solche Ereignisse werden “zufällige Ereignisse” genannt. Sei A ein solches Ereignis. Wenn $n(A)$ die Häufigkeit des Auftretens von A in n Experimenten ist, so hat man bemerkt, dass $\frac{n(A)}{n} \rightarrow c$ für große n ($n \rightarrow \infty$). Diese Konstante c nennt man “Wahrscheinlichkeit von A ” und bezeichnet sie mit $P(A)$.

Beispiel: *n*-maliger Münzwurf: (siehe Abbildung 2.1) faire Münze, d.h. $n(A) \approx n(\bar{A})$, $A = \{\text{Kopf}\}$, $\bar{A} = \{\text{Zahl}\}$

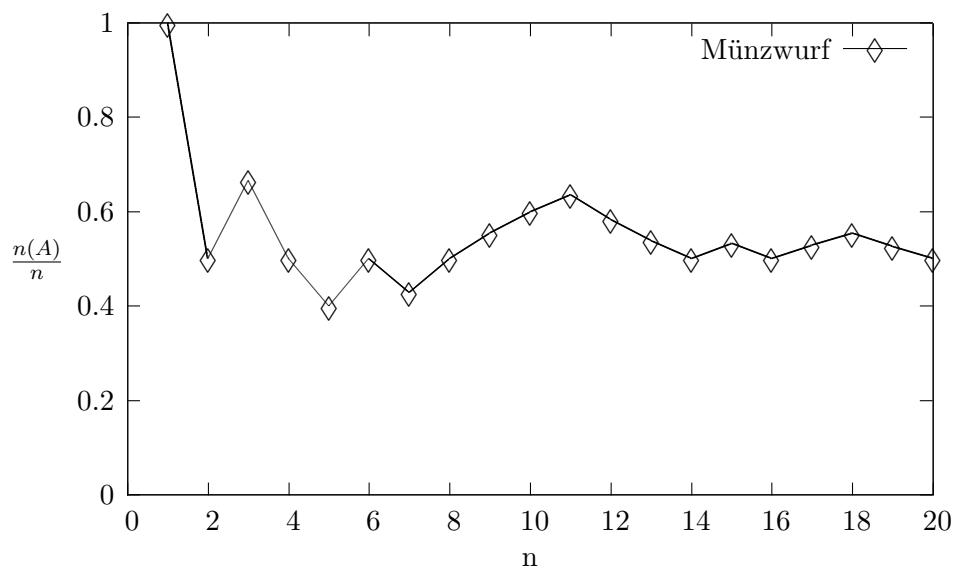


Abbildung 2.1: Relative Häufigkeit $\frac{n(A)}{n}$ des Ereignisses “Kopf” beim n -maligen Münzwurf

Man kann leicht feststellen, dass $\frac{n(A)}{n} \approx \frac{1}{2}$ für große n . $\Rightarrow P(A) = \frac{1}{2}$. Um dies zu verifizieren, hat Buffon in XVIII Jh. 4040 mal eine faire Münze geworfen, davon war 2048 mal Kopf, so dass $\frac{n(A)}{n} = 0,508$. Pierson hat es 24000 mal gemacht: es ergab $n(A) = 12012$ und somit $\frac{n(A)}{n} \approx 0.5005$.

In den Definitionen, die wir bald geben werden, soll diese empirische Begriffsbildung $P(A) = \lim_{n \rightarrow \infty} \frac{n(A)}{n}$ ihren Ausdruck finden. Zunächst definieren wir, für welche Ereignisse A die Wahrscheinlichkeit $P(A)$ überhaupt eingeführt werden kann.

offene Fragen:

1. Was ist $P(A)$?
2. Für welche A ist $P(A)$ definiert?

2.1 Ereignisse

Sei E ein Grundraum und $\Omega \subseteq E$ sei die Menge von Elementarereignissen ω (Grundmenge).

Ω kann als Menge der möglichen Versuchsergebnisse interpretiert werden. Man nennt Ω manchmal auch *Grundgesamtheit* oder *Stichprobenraum*.

Definition 2.1 Eine Teilmenge A von Ω ($A \subset \Omega$) wird *Ereignis* genannt. Dabei ist $\{\omega\} \subset \Omega$ ein *Elementarereignis*, das das Versuchsergebnis ω darstellt. Falls bei einem Versuch das Ergebnis $\omega \in A$ erzielt wurde, so sagen wir, dass A eintritt.

Beispiel 2.2

1. Einmaliges Würfeln: $\Omega = \{1, 2, 3, 4, 5, 6\}$, $E = \mathbb{N}$
2. n -maliger Münzwurf: $\Omega = \{(\omega_1, \dots, \omega_n) : \omega_i \in \{0, 1\}\}$
 $E = \mathbb{N}^n$, $\omega_i = \begin{cases} 1, & \text{falls ein "Kopf" im } i\text{-ten Wurf} \\ 0, & \text{sonst} \end{cases}$

Weiter werden wir E nicht mehr spezifizieren.

Tabelle 2.1: Wahrscheinlichkeitstheoretische Bedeutung von Mengenoperationen

$A = \emptyset$	unmögliches Ereignis
$A = \Omega$	wahres Ereignis
$A \subset B$	aus dem Eintreten von A folgt auch, dass B eintritt.
$A \cap B = \emptyset$	(disjunkte, <i>unvereinbare</i> Ereignisse): A und B können nicht gleichzeitig eintreten.
$A = B \cup C$	Mindestens eines der Ereignisse B und C tritt ein.
$A = \bigcup_{i=1}^n A_i$	Ereignis A = “Es tritt mindestens ein Ereignis A_i ein”
$A = B \cap C$	Ereignis A = “Es treten B und C gleichzeitig ein.”
$A = \bigcap_{i=1}^n A_i$	Ereignis A = “Es treten alle Ereignisse A_1, \dots, A_n ein”
$\bar{A} = A^c$	Das Ereignis A tritt nicht ein.
$A = B \setminus C$	Ereignis A tritt genau dann ein, wenn B eintritt, aber nicht C
$A = B \Delta C$	Ereignis A tritt genau dann ein, wenn B oder C eintreten (<i>nicht gleichzeitig!</i>)

Anmerkung: $A = B \Delta C = (B \setminus C) \cup (C \setminus B)$ (symmetrische Differenz)

Definition 2.3 Ereignisse A_1, A_2, A_3, \dots heißen *paarweise disjunkt* oder *unvereinbar*, wenn $A_i \cap A_j = \emptyset \quad \forall i \neq j$

Beispiel 2.4 *Zweimaliges Würfeln:* $\Omega = \{(\omega_1, \omega_2) : \omega_i \in \{1, \dots, 6\}, i = 1, 2\}$, $A = \text{“die Summe der Augenzahlen ist 6”} = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$. Die Ereignisse $B = \text{“Die Summe der Augenzahlen ist ungerade”}$ und $C = \{(3, 5)\}$ sind unvereinbar.

Oft sind nicht alle Teilmengen von Ω als Ereignisse sinnvoll. Deswegen beschränkt man sich auf ein Teilsystem von Ereignissen mit bestimmten Eigenschaften; und zwar soll dieses Teilsystem abgeschlossen bezüglich Mengenoperationen sein.

Definition 2.5 Eine nicht leere Familie \mathcal{F} von Ereignissen aus Ω heißt *Algebra*, falls

1. $A \in \mathcal{F} \implies \bar{A} \in \mathcal{F}$
2. $A, B \in \mathcal{F} \implies A \cup B \in \mathcal{F}$

Beispiel 2.6 1. Die *Potenzmenge* $\mathcal{P}(\Omega) =$ (die Menge aller Teilmengen

von Ω) ist eine Algebra.

2. Im Beispiel 2.4 ist $\mathcal{F} = \{\emptyset, A, \bar{A}, \Omega\}$ eine Algebra. Dagegen ist $\mathcal{F} = \{\emptyset, A, B, C, \Omega\}$ keine Algebra: z.B. $A \cup C \notin \mathcal{F}$.

Lemma 2.7 (*Eigenschaften einer Algebra:*) Sei \mathcal{F} eine Algebra von Ereignissen aus Ω . Es gelten folgende Eigenschaften:

1. $\emptyset, \Omega \in \mathcal{F}$
2. $A, B \in \mathcal{F} \implies A \setminus B \in \mathcal{F}$
3. $A_1, \dots, A_n \in \mathcal{F} \implies \bigcup_{i=1}^n A_i \in \mathcal{F}$ und $\bigcap_{i=1}^n A_i \in \mathcal{F}$

Beweis

1. $\mathcal{F} \neq \emptyset \implies \exists A \in \mathcal{F} \implies \bar{A} \in \mathcal{F}$ nach Definition $\implies A \cup \bar{A} = \Omega \in \mathcal{F}$;
 $\emptyset = \bar{\Omega} \in \mathcal{F}$.
2. $A, B \in \mathcal{F}, \quad A \setminus B = A \cap \bar{B} = \overline{(A \cup B)} \in \mathcal{F}$.
3. *Induktiver Beweis:*
 $n = 2: A, B \in \mathcal{F} \implies A \cap B = \overline{(A \cup B)} \in \mathcal{F}$
 $n = k \mapsto n = k + 1: \quad \bigcap_{i=1}^{k+1} A_i = (\bigcap_{i=1}^k A_i) \cap A_{k+1} \in \mathcal{F}$.

□

Für die Entwicklung einer gehaltvollen Theorie sind aber Algebren noch zu allgemein. Manchmal ist es auch notwendig, unendliche Vereinigungen $\bigcup_{i=1}^{\infty} A_i$ oder unendliche Schnitte $\bigcap_{i=1}^{\infty} A_i$ zu betrachten, um z.B. Grenzwerte von Folgen von Ereignissen definieren zu können. Dazu führt man Ereignissysteme ein, die σ -Algebren genannt werden:

Definition 2.8

1. Eine Algebra \mathcal{F} heißt σ -Algebra, falls aus $A_1, A_2, \dots \in \mathcal{F}$ folgt, dass $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.
2. Das Paar (Ω, \mathcal{F}) heißt *Messraum*, falls \mathcal{F} eine σ -Algebra der Teilmengen von Ω ist.

Beispiel 2.9 1. $\mathcal{F} = \mathcal{P}(\Omega)$ ist eine σ -Algebra.

2. *Beispiel einer Algebra \mathcal{F} , die keine σ -Algebra ist:*

Sei \mathcal{F} die Klasse von Teilmengen aus $\Omega = \mathbb{R}$, die aus endlichen Vereinigungen von disjunkten Intervallen der Form $(-\infty, a]$, $(b, c]$ und (d, ∞) $a, b, c, d \in \mathbb{R}$, besteht. Offensichtlich ist \mathcal{F} eine Algebra. Dennoch ist \mathcal{F} keine σ -Algebra, denn $[b, c] = \underbrace{\bigcap_{n=1}^{\infty} (b - \frac{1}{n}, c]}_{\in \mathcal{F}} \notin \mathcal{F}$.

2.2 Wahrscheinlichkeitsräume

Auf einem Messraum (Ω, \mathcal{F}) wird ein *Wahrscheinlichkeitsmaß* durch folgende *Axiome von Kolmogorow* eingeführt:

Definition 2.10

1. Die Mengenfunktion $P : \mathcal{F} \rightarrow [0, 1]$ heißt *Wahrscheinlichkeitsmaß* auf \mathcal{F} , falls
 - (a) $P(\Omega) = 1$ (*Normiertheit*)
 - (b) $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$, A_n paarweise disjunkt $\implies P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ (σ -*Additivität*)
2. Das Tripel (Ω, \mathcal{F}, P) heißt *Wahrscheinlichkeitsraum*.
3. $\forall A \in \mathcal{F}$ heißt $P(A)$ *Wahrscheinlichkeit* des Ereignisses A .

Bemerkung 2.11 1. Nachfolgend werden nur solche $A \subset \Omega$ *Ereignisse* genannt, die zu der ausgewählten σ -Algebra \mathcal{F} von Ω gehören. Alle anderen Teilmengen $A \subset \Omega$ sind demnach *keine* Ereignisse.

2. \mathcal{F} kann nicht immer als $\mathcal{P}(\Omega)$ gewählt werden. Falls Ω endlich oder abzählbar ist, ist dies jedoch möglich. Dann kann $P(A)$ auf $(\Omega, \mathcal{P}(\Omega))$ als $P(A) = \sum_{\omega \in A} P(\{\omega\})$ definiert werden (klassische Definition der Wahrscheinlichkeiten).

Falls z.B. $\Omega = \mathbb{R}$ ist, dann kann \mathcal{F} meistens nicht mehr als $\mathcal{P}(\mathbb{R})$ gewählt werden, weil ansonsten \mathcal{F} eine große Anzahl von *pathologischen* Ereignissen enthalten würde, für die z.B. der Begriff der Länge nicht definiert ist.

Definition 2.12 Sei \mathcal{U} eine beliebige Klasse von Teilmengen aus Ω . Dann ist durch

$$\sigma(\mathcal{U}) = \bigcap_{\mathcal{U} \subset \mathcal{F}, \mathcal{F} - \sigma\text{-Alg. von } \Omega} \mathcal{F}$$

eine σ -Algebra gegeben, die *minimale σ -Algebra, die \mathcal{U} enthält*, genannt wird.

Übungsaufgabe 2.13 Zeigen Sie bitte, dass die in Definition 2.12 definierte Klasse $\sigma(\mathcal{U})$ tatsächlich eine σ -Algebra darstellt.

Definition 2.14 Sei $\Omega = \mathbb{R}^d$. Sei \mathcal{U} = Klasse aller offenen Teilmengen von \mathbb{R}^d . Dann heißt $\sigma(\mathcal{U})$ die *Borel σ -Algebra* auf \mathbb{R}^d und wird mit $\mathfrak{B}_{\mathbb{R}^d}$ bezeichnet. Elemente von $\mathfrak{B}_{\mathbb{R}^d}$ heißen *Borel-Mengen*. Diese Definition kann auch für einen beliebigen topologischen Raum Ω (nicht unbedingt \mathbb{R}^d) gegeben werden.

Übungsaufgabe 2.15 Zeigen Sie, dass \mathcal{B}_{R^d} alle $\{x\}, x \in \mathbb{R}^d$, alle geschlossenen und insbesondere kompakten Teilmengen von \mathbb{R}^d enthält, z.B. $[a, b]^d \in \mathcal{B}_{R^d}$, $a \leq b$.

Satz 2.16 Sei (Ω, \mathcal{F}, P) ein Wahrscheinlichkeitsraum und $A_1, \dots, A_n, A, B \subset \mathcal{F}$. Dann gilt:

1. $P(\emptyset) = 0$.
2. $P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$, falls A_i paarweise disjunkt sind.
3. Falls $A \subset B$, dann ist $P(B \setminus A) = P(B) - P(A)$.

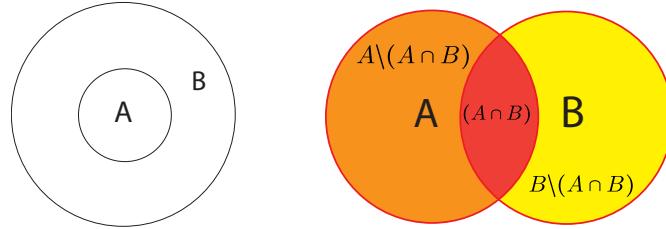


Abbildung 2.2: Illustration zu Satz 2.16, 3) und 4)

4. $P(A \cup B) = P(A) + P(B) - P(A \cap B), \quad \forall A, B \in \mathcal{F}$.
5. $P(\overline{B}) = 1 - P(B)$

Beweis

1. $\emptyset = \bigcup_{i=1}^{\infty} \emptyset \implies P(\emptyset) = \sum_{i=1}^{\infty} P(\emptyset) \leq 1 \implies P(\emptyset) = 0$
2. $P(\bigcup_{i=1}^n A_i) = P(\bigcup_{i=1}^n A_i \cup \bigcup_{k=n+1}^{\infty} \emptyset) = \sum_{i=1}^n P(A_i) + \sum_{i=n+1}^{\infty} 0 = \sum_{i=1}^n P(A_i)$
3. $P(B) = P(A \cup (B \setminus A)) = P(A) + P(B \setminus A) \implies$ geht.
4. Benutze 2), 3) und $A \cup B = [A \setminus (A \cap B)] \cup [A \cap B] \cup [B \setminus (A \cap B)]$

$$\begin{aligned} P(A \cup B) &= P(A) - P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

(vgl. Abb. 2.2).

5. $\overline{B} = \Omega \setminus B \Rightarrow P(\overline{B}) = \underbrace{P(\Omega)}_{=1} - P(B) = 1 - P(B)$.

□

Folgerung 2.17

Es gelten folgende Eigenschaften von P für $A_1, \dots, A_n, A, B \in \mathcal{F}$:

1. $A \subseteq B \implies P(A) \leq P(B)$ (Monotonie)
2. $P(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i), \quad \forall A_1, \dots, A_n \in \mathcal{F}$ (Subadditivität)
3. *Siebformel:*

$$P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n (-1)^{i-1} \sum_{1 \leq k_1 < \dots < k_i \leq n} P(A_{k_1} \cap \dots \cap A_{k_i})$$

Beweis

1. (Folgt aus Satz 2.16) $P(B) = P(A) + \underbrace{P(B \setminus A)}_{\geq 0} \geq P(A)$
2. Induktion nach n : $n=2$: $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2)$
3. Induktion nach n : Der Fall $n = 2$ folgt aus Satz 2.16, 4). Der Rest ist klar.

□

Übungsaufgabe 2.18 Führen Sie die Induktion bis zum Ende durch.

Die Eigenschaft $P(\bigcup_{n=1}^k A_n) \leq \sum_{n=1}^k P(A_n)$ heißt *Subadditivität des Wahrscheinlichkeitsmaßes P* . Diese Eigenschaft gilt jedoch auch für unendlich viele A_n , wie folgendes Korollar zeigt:

Folgerung 2.19 σ -Subadditivität: Sei (Ω, \mathcal{F}, P) ein Wahrscheinlichkeitsmaß und $\{A_n\}_{n \in \mathbb{N}}$ eine Folge von Ereignissen. Dann gilt $P(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} P(A_n)$, wobei die rechte Seite nicht unbedingt endlich sei soll (dann ist die Aussage trivial).

Definition 2.20 1. Ereignisse A und B heißen (*stochastisch*) *unabhängig*, falls $P(A \cap B) = P(A) \cdot P(B)$.

2. Eine Folge von Ereignissen $\{A_n\}_{n \in \mathbb{N}}$ (diese Folge kann auch endlich viele Ereignisse enthalten!) heißt (*stochastisch*) *unabhängig in ihrer Gesamtheit*, falls $\forall n \forall i_1 < i_2 < \dots < i_n$

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_n}) = \prod_{k=1}^n P(A_{i_k}).$$

2.3 Beispiele

In diesem Abschnitt betrachten wir die wichtigsten Beispiele für Wahrscheinlichkeitsräume (Ω, \mathcal{F}, P) . Wir beginnen mit:

2.3.1 Klassische Definition der Wahrscheinlichkeiten

Hier wird ein Grundraum Ω mit $|\Omega| < \infty$ ($|A| = \#A$ – die Anzahl von Elementen in A) betrachtet. Dann kann \mathcal{F} als $\mathcal{P}(\Omega)$ gewählt werden. Die klassische Definition von Wahrscheinlichkeiten geht von der Annahme aus, dass alle Elementarereignisse ω gleich wahrscheinlich sind:

Definition 2.21

1. Ein Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) mit $|\Omega| < \infty$ ist *endlicher Wahrscheinlichkeitsraum*.
2. Ein endlicher Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) mit $\mathcal{F} = \mathcal{P}(\Omega)$ und

$$\forall \omega \in \Omega \quad P(\{\omega\}) = \frac{1}{|\Omega|}$$

heißt *Laplacescher Wahrscheinlichkeitsraum*. Das eingeführte Maß heißt *klassisches* oder *laplacesches Wahrscheinlichkeitsmaß*.

Bemerkung 2.22 Für die klassische Definition der Wahrscheinlichkeit sind alle Elementarereignisse $\{\omega\}$ gleich wahrscheinlich:

$\forall \omega \in \Omega \quad P(\{\omega\}) = \frac{1}{|\Omega|}$. Nach der Additivität von Wahrscheinlichkeitsmaßen gilt:

$$P(A) = \frac{|A|}{|\Omega|} \quad \forall A \subset \Omega.$$

(Beweis: $P(A) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{\omega \in A} \frac{1}{|\Omega|} = \frac{|A|}{|\Omega|}$). Dabei heißt

$$P(A) = \frac{\text{Anzahl günstiger Fälle}}{\text{Anzahl aller Fälle}}.$$

Beispiel 2.23

1. *Problem von Galilei:*

Ein Landsknecht hat Galilei (manche sagen, es sei Huygens passiert) folgende Frage gestellt: Es werden 3 Würfel gleichzeitig geworfen. Was ist wahrscheinlicher: Die Summe der Augenzahlen ist 11 oder 12? Nach Beobachtung sollte 11 öfter vorkommen als 12. Doch ist es tatsächlich so?

- Definieren wir den Wahrscheinlichkeitsraum $\Omega = \{\omega = (\omega_1, \omega_2, \omega_3) : \omega_i \in \{1, \dots, 6\}\}, |\Omega| = 6^3 = 216 < \infty, \mathcal{F} = \mathcal{P}(\Omega)$; sei

$$\begin{aligned} B &:= \{\text{Summe der Augenzahlen } 11\} \\ &= \{\omega \in \Omega : \omega_1 + \omega_2 + \omega_3 = 11\} \\ C &:= \{\text{Summe der Augenzahlen } 12\} \\ &= \{\omega \in \Omega : \omega_1 + \omega_2 + \omega_3 = 12\}. \end{aligned}$$

- *Lösung des Landknechtes:* 11 und 12 können folgendermaßen in die Summe von 3 Summanden zerlegt werden:

$$\begin{aligned} 11 &= 1 + 5 + 5 = 1 + 4 + 6 = 2 + 3 + 6 = 2 + 4 + 5 = 3 + 3 + 5 = \\ 3 + 4 + 4 &\implies |B| = 6 \implies P(B) = \frac{6}{6^3} = \frac{1}{36} \\ 12 &= 1 + 5 + 6 = 2 + 4 + 6 = 2 + 5 + 5 = 3 + 4 + 5 = 3 + 3 + 6 = \\ 4 + 4 + 4 &\implies |C| = 6 \implies P(C) = \frac{6}{6^3} = \frac{1}{36}. \end{aligned}$$

Dies entspricht jedoch nicht der Erfahrung.

Die Antwort von Galilei war, dass der Landsknecht mit nicht unterscheidbaren Würfeln gearbeitet hat, somit waren Kombinationen wie (1,5,5), (5,1,5) und (5,5,1) identisch und wurden nur einmal gezählt. In der Tat ist es anders: Jeder Würfel hat eine Nummer, ist also von den anderen Zwei zu unterscheiden. Daher gilt $|B| = 27$ und $|C| = 25$, was daran liegt, das (4,4,4) nur einmal gezählt wird. Also

$$P(B) = \frac{|B|}{|\Omega|} = \frac{27}{216} > P(C) = \frac{|C|}{|\Omega|} = \frac{25}{216}$$

- 2. *Geburtstagsproblem:* Es gibt n Studenten in einem Jahrgang an der Uni, die dieselbe Vorlesung Angewandte Stochastik besuchen. Wie groß ist die Wahrscheinlichkeit, dass mindestens 2 Studenten den Geburtstag am selben Tag feiern? Sei $M = 365$ = Die Anzahl der Tage im Jahr. Dann gilt

$$\Omega = \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in \{1, \dots, M\}\}, |\Omega| = M^n < \infty.$$

n	4	16	22	23	40	64
$P(A_n)$	0,016	0,284	0,476	0,507	0,891	0,997

Tabelle 2.2: Geburtstagsproblem

Sei

$$\begin{aligned} A_n &= \{\text{min. 2 Studenten haben am gleichen Tag Geb.}\} \subset \Omega, \\ A_n &= \{\omega \in \Omega : \exists i, j \in \{1, \dots, n\}, i \neq j : \omega_i = \omega_j\}, \\ P(A_n) &= ? \end{aligned}$$

Ansatz: $P(A_n) = 1 - P(\bar{A}_n)$, wobei

$$\bar{A}_n = \{\omega \in \Omega : \omega_i \neq \omega_j \quad \forall i \neq j \text{ in } \omega = (\omega_1, \dots, \omega_n)\}$$

$|\bar{A}_n| = M(M-1)(M-2) \dots (M-n+1)$. Somit gilt

$$\begin{aligned} P(\bar{A}_n) &= \frac{M(M-1) \dots (M-n+1)}{M^n} \\ &= \left(1 - \frac{1}{M}\right) \left(1 - \frac{2}{M}\right) \dots \left(1 - \frac{n-1}{M}\right) \end{aligned}$$

und

$$P(A_n) = 1 - \left(1 - \frac{1}{M}\right) \dots \left(1 - \frac{n-1}{M}\right)$$

Für manche n gibt Tabelle 2.2 die numerischen Wahrscheinlichkeiten von $P(A_n)$ an.

Es gilt offensichtlich $P(A_n) \approx 1$ für $n \rightarrow M$. Interessanterweise ist $P(A_n) \approx 0,5$ für $n = 23$. Dieses Beispiel ist ein Spezialfall eines so genannten *Urnenmodells*: In einer Urne liegen M durchnummerierte Bälle. Aus dieser Urne werden n Stück auf gut Glück mit Zurücklegen entnommen. Wie groß ist die Wahrscheinlichkeit, dass in der Stichprobe mindestens 2 Gleiche vorkommen?

3. *Urnenmodelle*: In einer Urne gibt es M durchnummerierte Bälle. Es werden n Stück „zufällig“ entnommen. Das Ergebnis dieses Experiments ist eine Stichprobe (j_1, \dots, j_n) , wobei j_m die Nummer des Balls in der m -ten Ziehung ist. Es werden folgende Arten der Ziehung betrachtet:

- mit Zurücklegen
- ohne Zurücklegen

- mit Reihenfolge
- ohne Reihenfolge

Das Geburtstagsproblem ist somit ein Urnenproblem mit Zurücklegen und mit Reihenfolge.

Demnach werden auch folgende Grundräume betrachtet:

- (a) Ziehen mit Reihenfolge und mit Zurücklegen:

$$\Omega = \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in \underbrace{\{1, \dots, M\}}_{=K}\} = K^n, \quad |\Omega| = M^n$$

- (b) Ziehen mit Reihenfolge und ohne Zurücklegen:

$$\Omega = \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in K, \omega_i \neq \omega_j, i \neq j\},$$

$$|\Omega| = M(M-1)\dots(M-n+1) = \frac{M!}{(M-n)!}$$

Spezialfall: $M=n$ (Permutationen): $\Rightarrow |\Omega| = M!$

- (c) Ziehen ohne Reihenfolge und mit Zurücklegen:

$$\Omega = \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in K, \omega_1 \leq \omega_2 \leq \dots \leq \omega_n\}$$

Dies ist äquivalent zu der Verteilung von n Teilchen auf M Zellen ohne Reihenfolge \iff das Verteilen von $M-1$ Trennwänden der Zellen unter n Teilchen (vgl. Abb. 2.3). Daher ist

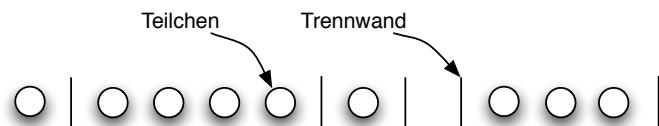


Abbildung 2.3: Ziehen ohne Reihenfolge und mit Zurücklegen

$$|\Omega| = \frac{(M+n-1)!}{n!(M-1)!} = \binom{M+n-1}{n}$$

- (d) Ziehen ohne Reihenfolge und ohne Zurücklegen:

$$\Omega = \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in K, \omega_1 < \omega_2 < \dots < \omega_n\}$$

$$|\Omega| = \frac{M!}{(M-n)!n!} = \binom{M}{n}$$

Auswahl von n aus M Kugeln in einer Urne	mit Zurücklegen	ohne Zurücklegen	
mit Reihenfolge	M^n (<i>Maxwell–Boltzmann–Statistik</i>)	$\frac{M!}{(M-n)!}$	unterscheidbare Teilchen
ohne Reihenfolge	$\binom{M+n-1}{n}$ (<i>Bose–Einstein–Statistik</i>)	$\binom{M}{n}$ (<i>Fermi–Dirac–Statistik</i>)	nicht unterscheidbare Teilchen
	mit Mehrfachbelegung	ohne Mehrfachbelegung	Verteilung von n Teilchen auf M Zellen.

Tabelle 2.3: Die Potenz $|\Omega|$ der Grundgesamtheit Ω in Urnenmodellen.

Ein Experiment der Mehrfachziehung aus einer Urne entspricht der Verteilung von n (unterschiedlichen oder nicht unterscheidbaren) Teilchen (wie z.B. Elektronen, Protonen, usw.) auf M Energieebenen oder Zellen (mit oder ohne Mehrfachbelegung dieser Ebenen) in der statistischen Physik. Die entsprechenden Namen der Modelle sind in Tabelle 2.3 zusammengeführt. So folgen z.B. Elektronen, Protonen und Neutronen der so genannten *Fermi–Dirac–Statistik* (nicht unterscheidbare Teilchen ohne Mehrfachbelegung). Photonen und Prionen folgen der *Bose–Einstein–Statistik* (nicht unterscheidbare Teilchen mit Mehrfachbelegung). Unterscheidbare Teilchen, die dem *Exklusionsprinzip von Pauli* folgen (d.h. ohne Mehrfachbelegung), kommen in der Physik nicht vor.

4. *Lotterie–Beispiele:* ein Urnenmodell ohne Reihenfolge und ohne Zurücklegen; In einer Lotterie gibt es M Lose (durchnummiert von 1 bis M), davon n Gewinne ($M \geq 2n$). Man kauft n Lose. Mit welcher Wahrscheinlichkeit gewinnt man mindestens einen Preis?
Laut Tabelle 2.3 ist $|\Omega| = \binom{M}{n}$,

$$\Omega = \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \neq \omega_j, i \neq j, \omega_i \in \{1 \dots M\}\}.$$

Sei $A = \{\text{es gibt mind. 1 Preis}\}$.

$$\begin{aligned}
P(A) &= 1 - P(\bar{A}) = 1 - P(\text{es werden keine Preise gewonnen}) \\
n &= 1 - \frac{\binom{M-n}{n}}{|\Omega|} = 1 - \frac{\frac{(M-n)!}{n!(M-2n)!}}{\frac{M!}{n!(M-n)!}} \\
&= 1 - \frac{(M-n)(M-n-1)\dots(M-2n+1)}{M(M-1)\dots(M-n+1)} \\
&= 1 - \left(1 - \frac{n}{M}\right) \left(1 - \frac{n}{M-1}\right) \dots \left(1 - \frac{n}{M-n+1}\right).
\end{aligned}$$

Um ein Beispiel zu geben, sei $M = n^2$. Dann gilt:

$P(A) \xrightarrow{n \rightarrow \infty} 1 - e^{-1} \approx 0,632$, denn $e^x = \lim_{n \rightarrow \infty} (1 + \frac{x}{n})^n$. Die Konvergenz ist schnell, $P(A) = 0,670$ schon für $n = 10$.

5. *Hypergeometrische Verteilung:* Nehmen wir jetzt an, dass M Kugeln in der Urne zwei Farben tragen können: schwarz und weiß. Seien S schwarze und W weiße Kugeln gegeben ($M = S + W$). Wie groß ist die Wahrscheinlichkeit, dass aus n zufällig entnommenen Kugeln (ohne Reihenfolge und ohne Zurücklegen) s schwarz sind?

Sei $A = \{\text{unter } n \text{ entnommenen Kugeln } s \text{ schwarze}\}$. Dann ist

$$P(A) = \frac{\binom{S}{s} \binom{W}{n-s}}{\binom{M}{n}}.$$

Diese Wahrscheinlichkeiten bilden die so genannte *hypergeometrische Verteilung*.

Um ein numerisches Beispiel zu geben, seien 36 Spielkarten gegeben. Sie werden zufällig in zwei gleiche Teile aufgeteilt. Wie groß ist die Wahrscheinlichkeit, dass die Anzahl von roten und schwarzen Karten in diesen beiden Teilen gleich ist?

Lösung: hypergeometrische Wahrscheinlichkeiten mit $M = 36$, $S = W = n = 18$, $s = \frac{18}{2} = 9$, $w = s = 9$. Dann ist

$$P(A) = \frac{\binom{18}{9} \binom{18}{9}}{\binom{36}{18}} = \frac{(18!)^4}{36!(9!)^4}.$$

Wenn man die Formel von Stirling

$$n! \approx \sqrt{2\pi n} n^n e^{-n}$$

benutzt, so kommt man auf

$$P(A) \approx \frac{(\sqrt{2\pi}18 \cdot 18^{18}e^{-18})^4}{\sqrt{2\pi}36 \cdot 36^{36}e^{-36}(\sqrt{2\pi}9 \cdot 9^9e^{-9})^4} \approx \frac{2}{\sqrt{18\pi}} \approx \frac{4}{15} \approx 0.26$$

2.3.2 Geometrische Wahrscheinlichkeiten

Hier sei ein Punkt π zufällig auf eine beschränkte Teilmenge Ω von \mathbb{R}^d geworfen. Wie groß ist die Wahrscheinlichkeit, dass π die Teilmenge $A \subset \Omega$ trifft? Um dieses Experiment formalisieren zu können, dürfen wir nur solche Ω und A zulassen, für die der Begriff des d -dimensionalen Volumens (Lebesgue-Maß) wohl definiert ist. Daher werden wir nur Borelsche Teilmengen von \mathbb{R}^d betrachten. Also sei $\Omega \in \mathcal{B}_{\mathbb{R}^d}$ und $|\cdot|$ das Lebesgue-Maß auf \mathbb{R}^d , $|\Omega| < \infty$. Sei $\mathcal{F} = \mathcal{B}_{\mathbb{R}^d} \cap \Omega$ (vgl. Abb. 2.4).

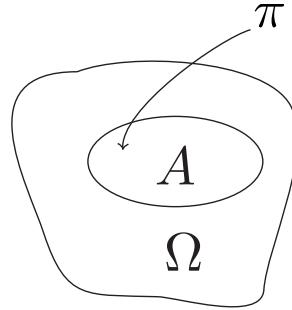


Abbildung 2.4: Zufälliger Punkt π auf Ω .

Definition 2.24

- Das Wahrscheinlichkeitsmaß auf (Ω, \mathcal{F}) gegeben durch

$$P(A) = \frac{|A|}{|\Omega|}, \quad A \in \mathcal{F}$$

heißt *geometrische Wahrscheinlichkeit* auf Ω .

- Das Tripel (Ω, \mathcal{F}, P) heißt *geometrischer Wahrscheinlichkeitsraum*.

Beispiel 2.25

Die Koeffizienten p und q einer quadratischen Gleichung $x^2 + px + q = 0$ werden zufällig im Intervall $(0, 1)$ gewählt. Wie groß ist die Wahrscheinlichkeit, dass die Lösungen x_1, x_2 dieser Gleichung reelle Zahlen sind?

Hier ist $\Omega = \{(p, q) : p, q \in (0, 1)\} = (0, 1)^2$, $\mathcal{F} = \mathcal{B}_{\mathbb{R}^2} \cap \Omega$.

$$A = \{x_1, x_2 \in \Omega\} = \{(p, q) \in \Omega : p^2 \geq 4q\},$$

denn $x_1, x_2 \in \mathbb{R}$ genau dann, wenn die Diskriminante $D = p^2 - 4q \geq 0$. Also gilt $A = \{(p, q) \in [0, 1]^2 : q \leq \frac{1}{4}p^2\}$ und

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\int_0^1 \frac{1}{4}p^2 dp}{1} = \frac{1}{12},$$

vgl. Abb. 2.5.

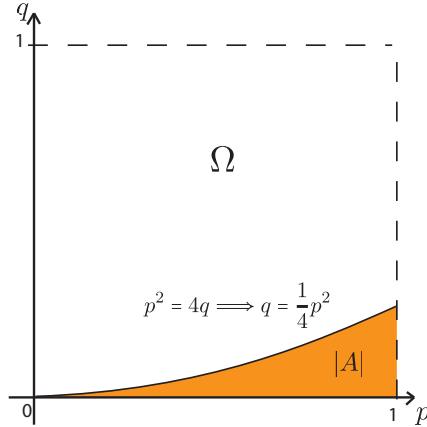


Abbildung 2.5: Wahrscheinlichkeit für reelle Lösungen einer quadratischen Gleichung

2.3.3 Bedingte Wahrscheinlichkeiten

Um den Begriff der bedingten Wahrscheinlichkeit intuitiv einführen zu können, betrachten wir zunächst das Beispiel der klassischen Wahrscheinlichkeiten: Sei (Ω, \mathcal{F}, P) ein Laplacescher Wahrscheinlichkeitsraum mit $|\Omega| = N$. Seien A und B Ereignisse aus \mathcal{F} . Dann gilt

$$P(A) = \frac{|A|}{N}, \quad P(A \cap B) = \frac{|A \cap B|}{N}.$$

Wie groß ist die Wahrscheinlichkeit $P(A|B)$ von A unter der Bedingung, dass B eintritt?

Da B eingetreten ist, ist die Gesamtanzahl aller Elementarereignisse hier gleich $|B|$. Die Elementarereignisse, die zu A beim Eintreten von B führen, liegen alle in $A \cap B$. Somit ist die Anzahl der “günstigen” Fälle hier $|A \cap B|$ und wir bekommen

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{|A \cap B|/N}{|B|/N} = \frac{P(A \cap B)}{P(B)}.$$

Dies führt zu folgender Definition:

Definition 2.26

Sei (Ω, \mathcal{F}, P) ein beliebiger Wahrscheinlichkeitsraum, $A, B \in \mathcal{F}$, $P(B) > 0$. Dann ist die *bedingte Wahrscheinlichkeit* von A unter der Bedingung B gegeben durch

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Diese Definition kann in Form des sogenannten Multiplikationssatzes gegeben werden:

$$P(A \cap B) = P(A|B) \cdot P(B).$$

Übungsaufgabe 2.27 Zeigen Sie, dass $P(\cdot|B)$ für $B \in \mathcal{F}$, $P(B) > 0$ ein Wahrscheinlichkeitsmaß auf (Ω, \mathcal{F}) ist.

Satz 2.28 Seien $A_1, \dots, A_n \in \mathcal{F}$ Ereignisse mit $P(A_1 \cap \dots \cap A_{n-1}) > 0$, dann gilt $P(A_1 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \dots \dots P(A_n|A_1 \cap \dots \cap A_{n-1})$.

Übungsaufgabe 2.29 Beweisen Sie den Satz 2.28.

Beweisidee: Induktion bezüglich n.

An dieser Stelle sollte man zu den stochastisch unabhängigen Ereignissen zurückkehren. A und B sind nach Definition 2.20 unabhängig, falls $P(A \cap B) = P(A) \cdot P(B)$. Dies ist äquivalent zu $P(A|B) = P(A)$, falls $P(B) > 0$. Es sei allerdings an dieser Stelle angemerkt, dass die Definition 2.20 allgemeiner ist, weil sie auch den Fall $P(B) = 0$ zulässt.

Übungsaufgabe 2.30 Zeigen Sie folgendes:

1. Seien $A, B \in \mathcal{F}$. A und B sind (stochastisch) unabhängig genau dann, wenn A und \bar{B} oder $(\bar{A}$ und $\bar{B})$ unabhängig sind.
2. Seien $A_1, \dots, A_n \in \mathcal{F}$. Ereignisse A_1, \dots, A_n sind stochastisch unabhängig in ihrer Gesamtheit genau dann, wenn B_1, \dots, B_n unabhängig in ihrer Gesamtheit sind, wobei $B_i = A_i$ oder $B_i = \bar{A}_i$ für $i = 1, \dots, n$.
3. Seien $A, B_1, B_2 \in \mathcal{F}$ mit $B_1 \cap B_2 = \emptyset$. Sei A und B_1 , A und B_2 unabhängig. Zeigen Sie, dass A und $B_1 \cup B_2$ ebenfalls unabhängig sind.

Bemerkung 2.31 Der in Definition 2.20 gegebene Begriff der stochastischen Unabhängigkeit ist viel allgemeiner als die sogenannte Unabhängigkeit im Sinne des Gesetzes von Ursache und Wirkung. In den folgenden Beispielen wird man sehen, dass zwei Ereignisse stochastisch unabhängig sein können, obwohl ein kausaler Zusammenhang zwischen ihnen besteht. Somit ist die stochastische Unabhängigkeit allgemeiner, und nicht an das Gesetz von Ursache und Wirkung gebunden. In der Praxis allerdings ist man gut beraten, Ereignisse, die keinen kausalen Zusammenhang haben als stochastisch unabhängig zu deklarieren.

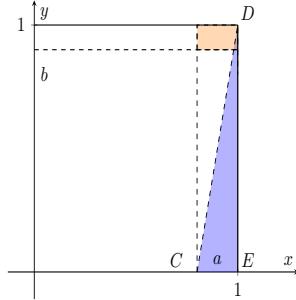
Beispiel 2.32 1. *Abhängige und unabhängige Ereignisse:*

Es werde ein Punkt $\pi = (X, Y)$ zufällig auf $[0, 1]^2$ geworfen. $\Omega = [0, 1]^2$, $\mathcal{F} = \mathcal{B}_{\mathbb{R}^2} \cap [0, 1]$. Betrachten wir $A = \{X \geq a\}$ und $B = \{Y \geq b\}$.

Dann gilt

$$\begin{aligned} P(A \cap B) &= P(X \geq a, Y \geq b) \\ &= \frac{(1-a)(1-b)}{1} \\ &= P(A) \cdot P(B), \end{aligned}$$

insofern sind A und B stochastisch unabhängig. Allerdings kann für $B = \{\pi \in \Delta CDE\}$ leicht gezeigt werden, dass A und B voneinander abhängig sind.



2. Es können $n + 1$ Ereignisse konstruiert werden, die abhängig sind, wobei beliebige n von ihnen unabhängig sind, $\forall n \in \mathbb{N}$.
3. *Kausale und stochastische Unabhängigkeit:*

Auf das Intervall $[0, 1]$ wird auf gut Glück ein Punkt Π geworfen. Sei x die Koordinate von Π in $[0, 1]$. Betrachten wir die binäre Zerlegung der Zahl x :

$$x = \sum_{k=1}^{\infty} \frac{a_k}{2^k}, \quad a_n \in \{0, 1\}.$$

Dann ist klar, dass es einen starken kausalen Zusammenhang zwischen $\{a_n\}_{n=1}^{\infty}$ gibt, weil sie alle durch x verbunden sind. Man kann jedoch zeigen, dass die Ereignisse $B_k = \{a_k = j\}, k \in \mathbb{N}$ für alle $j = 0, 1$ unabhängig in ihrer Gesamtheit sind, und dass $P(a_k = j) = 1/2 \quad \forall k \in \mathbb{N}, j = 0, 1$.

Definition 2.33 Sei $\{B_n\}$ eine endliche oder abzählbare Folge von Ereignissen aus \mathcal{F} . Sie heißt eine *messbare Zerlegung* von Ω , falls

1. B_n paarweise disjunkt sind: $B_i \cap B_j = \emptyset, \quad i \neq j$
2. $\bigcup_n B_n = \Omega$
3. $P(B_n) > 0 \quad \forall n$.

Satz 2.34 (*Formel der totalen Wahrscheinlichkeit, Bayes'sche Formel*):

Sei $\{B_n\} \subset \mathcal{F}$ eine messbare Zerlegung von Ω und $A \in \mathcal{F}$ ein beliebiges Ereignis, dann gilt

1. Die Formel der totalen Wahrscheinlichkeit:

$$P(A) = \sum_n P(A|B_n) \cdot P(B_n)$$

2. Bayes'sche Formel:

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{\sum_n P(A|B_n) \cdot P(B_n)} \quad \forall i$$

falls $P(A) > 0$. Die Summen in 1) und 2) können endlich oder unendlich sein, je nach Anzahl der B_n .

Beweis 1. Da $\Omega = \bigcup_n B_n$, ist $A = A \cap \Omega = A \cap (\bigcup_n B_n) = \bigcup_n (A \cap B_n)$ eine disjunkte Vereinigung von Ereignissen $A \cap B_n$, und es gilt

$$P(A) = P\left(\bigcup_n (A \cap B_n)\right) \underset{\sigma\text{-Add. v. } P}{=} \sum_n P(A \cap B_n) \underset{\text{S. 2.28}}{=} \sum_n P(A|B_n)P(B_n)$$

2.

$$P(B_i|A) \underset{\text{Def. 2.26}}{=} \frac{P(B_i \cap A)}{P(A)} \underset{\text{S. 2.28 u. 2.34 1}}{=} \frac{P(A|B_i) \cdot P(B_i)}{\sum_n P(A|B_n)P(B_n)}$$

□

Bemerkung 2.35 Die Ereignisse B_n heißen oft ‘‘Hypothesen’’. Dann ist $P(B_n)$ die so genannte *a-priori-Wahrscheinlichkeit von B_n* , also vor dem ‘‘Experiment’’ A . Die Wahrscheinlichkeiten $P(B_n|A)$ werden als Wahrscheinlichkeiten des Auftretens von B_n ‘‘nach dem Experiment A ’’ interpretiert. Daher heißen sie auch oft ‘‘*a-posteriori-Wahrscheinlichkeiten von B_n* ’’. Die Formel von Bayes verbindet also die a-posteriori-Wahrscheinlichkeiten mit den a-priori-Wahrscheinlichkeiten.

Beispiel 2.36

1. *Routing-Problem:*

Im Internet muss ein Paket von Rechner S (Sender) auf den Rechner E (Empfänger) übertragen werden. In Abb. 2.6 ist die Geometrie des Computernetzes zwischen S und E schematisch dargestellt, wobei R_1, R_2, R_3 und R_4 (und andere Knoten des Graphen) jeweils andere Rechner sind, die sich an der Übertragung beteiligen können. Wir gehen davon aus, dass die Richtung der weiteren Übertragung des Paketes in den Knoten zufällig aus allen möglichen Knoten gewählt wird (mit gleicher Wahrscheinlichkeit). So ist z.B.

$$P\left(\underbrace{\text{von } S \text{ wird Router } R_i \text{ gewählt}}_{=A_i}\right) = \frac{1}{4} \quad i = 1, \dots, n.$$

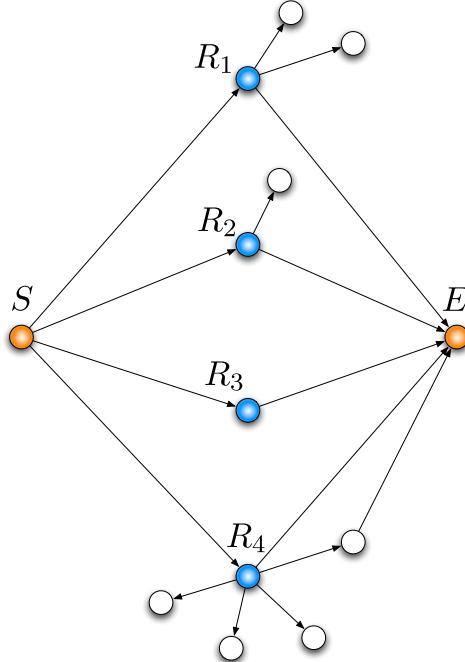


Abbildung 2.6: Routing–Problem: Computernetzwerk

Offensichtlich stellen die Ereignisse A_1, A_2, A_3, A_4 eine messbare Zerlegung von Ω dar. Nach Satz 2.34, 1) gilt also für $A = \{\text{das Paket erreicht } E \text{ aus } S\}$

$$P(A) = \sum_{i=1}^4 P(A|A_i) \cdot P(A_i) = \frac{1}{4} \sum_{i=1}^4 P(A|A_i).$$

Dabei können $P(A|A_i)$ aus dem Graphen eindeutig bestimmt werden:

$$\begin{aligned} P(A|A_1) &= \frac{1}{3}, & P(A|A_2) &= \frac{1}{2}, \\ P(A|A_3) &= 1, & P(A|A_4) &= \frac{2}{5}. \end{aligned}$$

Es gilt also

$$P(A) = \frac{1}{4} \left(\frac{1}{3} + \frac{1}{2} + 1 + \frac{2}{5} \right) = \frac{67}{120} \approx 0,5.$$

2. In einer Urne gibt es zwei Münzen. Die erste ist fair (Wahrscheinlichkeit des Kopfes und der Zahl = $\frac{1}{2}$), die zweite ist allerdings nicht fair mit $P(\text{Kopf}) = \frac{1}{3}$. Aus der Urne wird eine Münze zufällig genommen

und geworfen. In diesem Wurf ist das Ereignis Kopf. Wie groß ist die Wahrscheinlichkeit, dass die Münze fair war?

Sei

$$A_1 = \{\text{Faire Münze ausgewählt}\}$$

$$A_2 = \{\text{Nicht faire Münze ausgewählt}\}$$

$$A = \{\text{Es kommt Kopf im Münzwurf}\}$$

$$P(A_1|A) = ?$$

Dann gilt $P(A_1) = P(A_2) = \frac{1}{2}$, $P(A|A_1) = \frac{1}{2}$, $P(A|A_2) = \frac{1}{3}$, daher gilt nach der Bayesschen Formel

$$P(A_1|A) = \frac{P(A_1) \cdot P(A|A_1)}{P(A_1) \cdot P(A|A_1) + P(A_2) \cdot P(A|A_2)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2} \cdot (\frac{1}{2} + \frac{1}{3})} = \frac{3}{5}.$$

Kapitel 3

Zufallsvariablen

3.1 Definition und Beispiele

Definition 3.1 1. Eine Abbildung $X : \Omega \rightarrow \mathbb{R}$ heißt *Zufallsvariable*, falls sie $\mathcal{B}_{\mathbb{R}}$ -messbar ist, mit anderen Worten,

$$\forall B \in \mathcal{B}_{\mathbb{R}} \quad X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}.$$

2. Eine Abbildung $X : \Omega \rightarrow \mathbb{R}^n$, $n \geq 1$ heißt *Zufallsvektor*, falls sie $\mathcal{B}_{\mathbb{R}^n}$ -messbar ist, mit anderen Worten,

$$\forall B \in \mathcal{B}_{\mathbb{R}^n} \quad X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}.$$

Offensichtlich bekommt man aus Definition 3.1, 2) auch 3.1, 1) für $n = 1$.

Beispiel 3.2 1. *Indikator-Funktion eines Ereignisses:*

Sei (Ω, \mathcal{F}, P) ein Wahrscheinlichkeitsraum und A ein Ereignis aus \mathcal{F} .

Betrachten wir

$$X(\omega) = I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}.$$

Diese Funktion von ω nennt man *Indikator-Funktion des Ereignisses*

A. Sie ist offensichtlich messbar und somit eine Zufallsvariable:

$$X^{-1}(B) = \begin{cases} A & \text{falls } 1 \in B, 0 \notin B \\ \bar{A} & \text{falls } 1 \notin B, 0 \in B \\ \Omega & \text{falls } 0, 1 \in B \\ \emptyset & \text{falls } 0, 1 \notin B \end{cases} \in \mathcal{F} \quad \forall B \in \mathcal{B}_{\mathbb{R}}$$

2. *n-maliger Münzwurf:*

Sei $\Omega = \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in \{0, 1\}\} = \{0, 1\}^n$ mit

$$\omega_i = \begin{cases} 1, & \text{falls Kopf im i-ten Münzwurf} \\ 0, & \text{sonst} \end{cases}$$

für $i = 1, \dots, n$. Sei $\mathcal{F} = \mathcal{P}(\Omega)$. Definieren wir

$$X(\omega) = X((\omega_1, \dots, \omega_n)) = \sum_{i=1}^n \omega_i$$

als die Anzahl der Köpfe im n -maligen Münzwurf, so kann man zeigen, dass X \mathcal{F} -messbar ist und somit eine Zufallsvariable ist.

Satz 3.3 Eine Abbildung $X : \Omega \rightarrow \mathbb{R}$ ist genau dann eine Zufallsvariable, wenn $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F} \quad \forall x \in \mathbb{R}$.

3.2 Verteilungsfunktion

Definition 3.4 Sei (Ω, \mathcal{F}, P) ein beliebiger Wahrscheinlichkeitsraum und $X : \Omega \rightarrow \mathbb{R}$ eine Zufallsgröße.

1. Die Funktion $F_X(x) = P(\{\omega \in \Omega : X(\omega) \leq x\})$, $x \in \mathbb{R}$ heißt *Verteilungsfunktion* von X . Offensichtlich ist $F_X : \mathbb{R} \rightarrow [0, 1]$.
2. Die Mengenfunktion $P_X : \mathcal{B}_{\mathbb{R}} \rightarrow [0, 1]$ gegeben durch

$$P_X(B) = P(\{\omega \in \Omega : X(\omega) \in B\}), B \in \mathcal{B}_{\mathbb{R}}$$

heißt *Verteilung* von X .

Bemerkung 3.5 Folgende gekürzte Schreibweise wird benutzt:

$$F_X(x) = P(X \leq x), \quad P_X(B) = P(X \in B).$$

Beispiel 3.6

Hier geben wir Verteilungsfunktionen für Zufallsvariablen aus dem Beispiel 3.2 an.

1. *Indikator-Funktion:*

Sei $X(\omega) = I_A(\omega)$. Dann ist

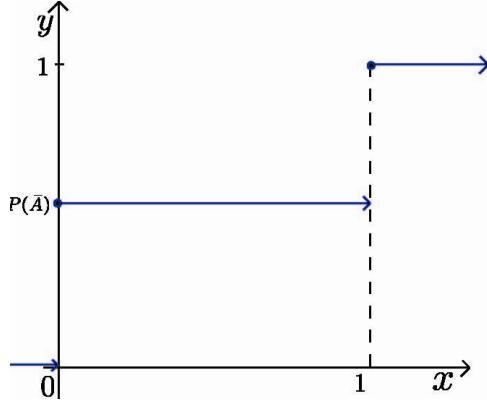
$$F_X(x) = P(I_A \leq x) = \begin{cases} 1, & x \geq 1, \\ P(\overline{A}), & x \in [0, 1], \\ 0, & x < 0 \end{cases}$$

vgl. Abb. 3.1.

2. *n -maliger Münzwurf:*

Sei X = Anzahl Kopf in n Münzwürfen. $P(\text{Kopf in einem Wurf}) = p$, $p \in (0, 1)$. Dann gilt

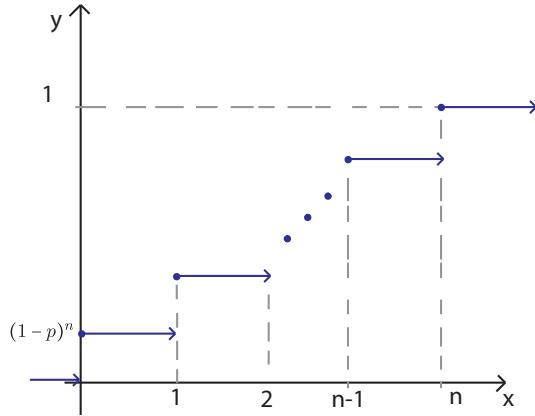
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \dots, n,$$

Abbildung 3.1: Verteilungsfunktion von I_A

und somit

$$F_X(x) = P(X \leq x) = \sum_{0 \leq k \leq [x]} P(x = k) = \sum_{k=0}^{[x]} \binom{n}{k} p^k (1-p)^{n-k},$$

$\forall x \in [0, n]$, vgl. Abb. 3.2. Es gilt

Abbildung 3.2: Verteilungsfunktion einer $Bin(n, p)$ -Verteilung

$$\begin{aligned} F_X(0) &= P(X \leq 0) = P(X = 0) = (1-p)^n, \\ F_X(x) &= P(X \leq x) = 0 \quad \text{für } x < 0, \\ F_X(n) &= P(X \leq n) = 1. \end{aligned}$$

Diese Verteilung wird später *Binomial-Verteilung* mit Parametern n, p genannt: $Bin(n, p)$

Satz 3.7 Sei X eine beliebige Zufallsvariable und $F_X : \mathbb{R} \rightarrow [0, 1]$ ihre Verteilungsfunktion. F_X besitzt folgende Eigenschaften:

1. *Asymptotik*: $\lim_{x \rightarrow -\infty} F_X(x) = 0$, $\lim_{x \rightarrow +\infty} F_X(x) = 1$.
2. *Monotonie*: $F_X(x) \leq F_X(x+h)$, $\forall x \in \mathbb{R}, h \geq 0$.
3. *Rechtsseitige Stetigkeit*: $\lim_{x \rightarrow x_0+0} F_X(x) = F_X(x_0)$ $\forall x_0 \in \mathbb{R}$.

Bemerkung 3.8

1. Im Satz 3.7 wurde gezeigt, dass eine Verteilungsfunktion F_X monoton nicht-fallend, rechtsseitig stetig und beschränkt auf $[0, 1]$ ist. Diese Eigenschaften garantieren, dass F_X höchstens abzählbar viele Sprungstellen haben kann. In der Tat kann F_X wegen $F_X \uparrow$ und $0 \leq F_X \leq 1$ nur eine endliche Anzahl von Sprungstellen mit Sprunghöhe $> \varepsilon$ besitzen, $\forall \varepsilon > 0$. Falls ε_n die Menge \mathbb{Q} aller rationaler Zahlen durchläuft, wird somit gezeigt, dass die Anzahl aller möglichen Sprungstellen höchstens abzählbar sein kann. Die Grafik einer typischen Verteilungsfunktion ist in Abb. 3.3 dargestellt.

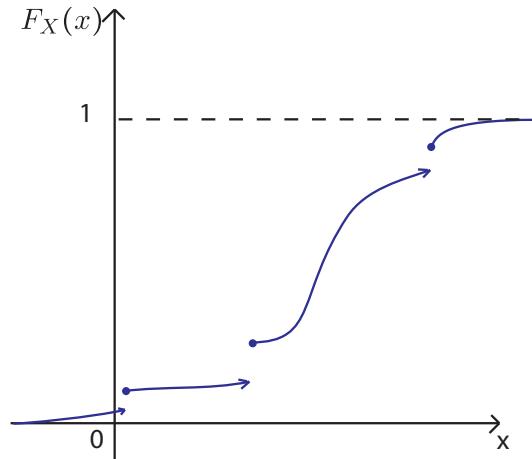


Abbildung 3.3: Typische Verteilungsfunktion

2. Mit Hilfe von F_X können folgende Wahrscheinlichkeiten leicht berechnet werden: $\forall a, b$ gilt: $-\infty \leq a < b \leq +\infty$

$$\begin{aligned} P(a < X \leq b) &= F_X(b) - F_X(a), \\ P(a \leq X \leq b) &= F_X(b) - \lim_{x \rightarrow a-0} F_X(x), \end{aligned}$$

denn

$$\begin{aligned} P(a < X \leq b) &= P(\{X \leq b\} \setminus \{X \leq a\}) = P(X \leq b) - P(X \leq a) \\ &= F_X(b) - F_X(a), \\ P(a \leq X \leq b) &= P(X \leq b) - P(X < a) = F_X(b) - \lim_{x \rightarrow a-0} F_X(x) \end{aligned}$$

mit $P(X < a) = P(X \leq a) - P(X = a) = \lim_{x \rightarrow a-0} F_X(x)$ nach Stetigkeit von P_X .

Da $P(X < a) = F(X \leq a) - P(X = a)$ gilt, ist somit

$$\lim_{x \rightarrow a-0} F_X(x) \neq F_X(a)$$

und F_X im Allgemeinen nicht linksseitig stetig.

Übungsaufgabe 3.9 Drücken Sie die Wahrscheinlichkeiten $P(a < X < b)$ und $P(a \leq X < b)$ mit Hilfe von F_X aus.

Satz 3.10 Falls eine Funktion $F(x)$ die Eigenschaften 1) bis 3) des Satzes 3.7 erfüllt, dann existiert ein Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) und eine Zufallsvariable X , definiert auf diesem Wahrscheinlichkeitsraum, derart, dass $F_X(x) = F(x)$, $\forall x \in \mathbb{R}$.

Satz 3.11 Die Verteilung P_X einer Zufallsvariable X wird eindeutig durch die Verteilungsfunktion F_X von X bestimmt.

3.3 Grundlegende Klassen von Verteilungen

In diesem Abschnitt werden wir Grundtypen von Verteilungen betrachten, die dem Schema aus Abbildung 3.4 zu entnehmen sind.

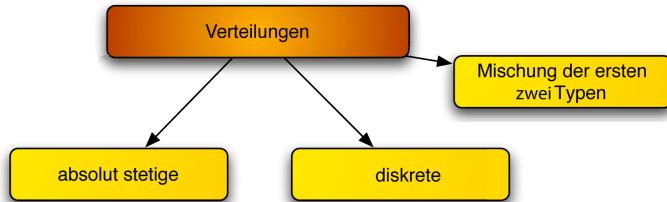


Abbildung 3.4: Verteilungstypen

3.3.1 Diskrete Verteilungen

Definition 3.12 1. Die Verteilung einer Zufallsvariablen X heißt *diskret*, falls eine höchstens abzählbare Teilmenge $C \subset \mathbb{R}$ (Wertebereich

von X) mit $P(X \in C) = 1$ existiert. Manchmal wird auch die Zufallsvariable X selbst als diskret bezeichnet.

2. Falls X eine diskrete Zufallsvariable mit Wertebereich $C = \{x_1, x_2, x_3, \dots\}$ ist, dann heißt $\{p_k\}$ mit $p_k = P(X = x_k)$, $k = 1, 2, \dots$ Wahrscheinlichkeitsfunktion oder Zähldichte von X .

Bemerkung 3.13

1. Beispiele für diskrete Wertebereiche C sind $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \{0, 1, \dots, n\}, n \in \mathbb{N}$.
2. Für die Zähldichte $\{p_k\}$ einer diskreten Zufallsvariable X gilt offenbar $0 \leq p_k \leq 1 \quad \forall n$ und $\sum_k p_k = 1$. Diese Eigenschaften sind für eine Zähldichte charakteristisch.
3. Die Verteilung P_X einer diskreten Zufallsvariable X wird eindeutig durch ihre Zähldichte $\{p_k\}$ festgelegt:

$$\begin{aligned} P_X(B) &= P_X(B \cap C) = P_X(\bigcup_{x_i \in B} \{x_i\}) = \sum_{x_i \in B} P(X = x_i) \\ &= \sum_{x_i \in B} p_i, \quad B \in \mathcal{B}_{\mathbb{R}}. \end{aligned}$$

Insbesondere gilt $F_X(x) = \sum_{x_k \leq x} p_k \implies P_X$ festgelegt nach Satz 3.11.

Wichtige diskrete Verteilungen:

Die Beispiele 3.2 und 3.6 liefern uns zwei wichtige diskrete Verteilungen mit Wertebereichen $\{0, 1\}$ und $\{0, 1, \dots, n\}$. Das sind

1. *Bernoulli-Verteilung:*

$X \sim Ber(p)$, $p \in [0, 1]$ (abkürzende Schreibweise für “Zufallsvariable X ist Bernoulli-verteilt mit Parameter p ”), falls

$$X = \begin{cases} 1, & \text{mit Wahrscheinlichkeit } p, \\ 0, & \text{mit Wahrscheinlichkeit } 1-p. \end{cases}$$

Dann gilt $C = \{0, 1\}$ und $p_0 = 1 - p$, $p_1 = p$ (vgl. Beispiel 3.2, 1) mit $X = I_A$.

2. *Binomialverteilung:*

$X \sim Bin(n, p)$, $p \in [0, 1], n \in \mathbb{N}$, falls $C = \{0, \dots, n\}$ und

$$P(X = k) = p_k = \binom{n}{k} \cdot p^k (1 - p)^{n-k}, \quad k = 0, \dots, n.$$

Interpretation:

$X = \#\{\text{Erfolge in einem } n \text{ mal unabhängig wiederholten Versuch}\}$, wobei $p = \text{Erfolgswahrscheinlichkeit in einem Versuch}$ (vgl. Beispiel 3.6, 2) mit $X = \#\{\text{Kopf}\}$.

3. *Geometrische Verteilung:*

$X \sim Geo(p)$, $p \in [0, 1]$, falls $C = \mathbb{N}$, und

$$P(X = k) = p_k = (1 - p)p^{k-1}, \quad k \in \mathbb{N}.$$

Interpretation: $X = \#\{\text{unabhängige Versuche bis zum ersten Erfolg}\}$, wobei $1 - p = \text{Erfolgswahrscheinlichkeit in einem Versuch}$.

4. *Hypergeometrische Verteilung:*

$X \sim HG(M, S, n)$, $M, S, n \in \mathbb{N}$, $S, n \leq M$, falls

$$X : \Omega \rightarrow \{0, 1, 2, \dots, \min\{n, S\}\}$$

und

$$P(X = k) = p_k = \frac{\binom{S}{k} \binom{M-S}{n-k}}{\binom{M}{n}}, \quad k = 0, 1, \dots, \min\{n, S\}.$$

Interpretation: Urnenmodell aus Beispiel 2.23, 5) mit

$X = \#\{\text{schwarze Kugeln bei } n \text{ Entnahmen aus einer Urne}\}$

mit insgesamt S schwarzen und $M - S$ weißen Kugeln.

5. *Gleichverteilung:*

$X \sim U\{x_1, \dots, x_n\}$, $n \in \mathbb{N}$, falls $X : \Omega \rightarrow \{x_1, \dots, x_n\}$ mit

$$p_k = P(X = x_k) = \frac{1}{n}, \quad k = 1, \dots, n$$

(wobei U in der Bezeichnung von Englischen ‘‘uniform’’ kommt).

Interpretation: $\{p_k\}$ ist eine Laplacesche Verteilung (klassische Definition von Wahrscheinlichkeiten, vgl. Abschnitt 2.3.1).

6. *Poisson-Verteilung:*

$X \sim Poisson(\lambda)$, $\lambda > 0$, falls $X : \Omega \rightarrow \{0, 1, 2, \dots\} = \mathbb{N} \cup \{0\}$ mit

$$p_k = P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \in \mathbb{N} \cup \{0\}.$$

Interpretation: $X = \#\{\text{Ereignisse im Zeitraum } [0, 1]\}$, λ ist die Rate (Häufigkeit), mit der Ereignisse passieren können, wobei

$$P(1 \text{ Ereignis tritt während } \Delta t \text{ ein}) = \lambda |\Delta t| + o(|\Delta t|),$$

$$P(> 1 \text{ Ereignis tritt während } \Delta t \text{ ein}) = o(|\Delta t|), \quad |\Delta t| \rightarrow 0$$

und $\#\{\text{Ereignisse in Zeitintervall } \Delta t_i\}$, $i = 1, \dots, n$ sind unabhängig, falls Δt_i , $i = 1, \dots, n$ disjunkte Intervalle aus \mathbb{R} sind. Hier $|\Delta t|$ ist die

Länge des Intervalls Δt .

z. B.

$$X = \#\{ \text{Schäden eines Versicherers in einem Geschäftsjahr} \}$$

$$X = \#\{ \text{Kundenanrufe eines Festnetzanbieters an einem Tag} \}$$

$$X = \#\{ \text{Elementarteilchen in einem Geiger-Zähler in einer Sekunde} \}.$$

Satz 3.14 (Approximationssatz)

1. *Binomiale Approximation:* Die hypergeometrische Verteilung $HG(M, S, n)$ kann für $M, S \rightarrow \infty, \frac{S}{M} \rightarrow p$ durch eine $Bin(n, p)$ -Verteilung approximiert werden: Für $X \sim HG(M, S, n)$ gilt

$$p_k = P(X = k) = \frac{\binom{S}{k} \binom{M-S}{n-k}}{\binom{M}{n}} \xrightarrow[M, S \rightarrow \infty, \frac{S}{M} \rightarrow p]{} \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \dots, n$$

2. *Poissonsche Approximation oder Gesetz der seltenen Ereignisse:* Die Binomialverteilung $Bin(n, p)$ kann für $n \rightarrow \infty, p \rightarrow 0, np \rightarrow \lambda$ durch eine Poisson-Verteilung $Poisson(\lambda)$ approximiert werden:

$$X \sim Bin(n, p), \quad p_k = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \xrightarrow[n \rightarrow \infty, p \rightarrow 0, np \rightarrow \lambda]{} e^{-\lambda} \frac{\lambda^k}{k!}$$

mit $k = 0, 1, 2, \dots$

Beweis 1. Falls $M, S \rightarrow \infty, \frac{S}{M} \rightarrow p \in (0, 1)$, dann gilt

$$\begin{aligned} \frac{\binom{S}{k} \binom{M-S}{n-k}}{\binom{M}{n}} &= \frac{\frac{S!}{k!(S-k)!} \cdot \frac{(M-S)!}{(M-S-n+k)!(n-k)!}}{\frac{M!}{n!(M-n)!}} \\ &= \frac{n!}{(n-k)!k!} \cdot \underbrace{\frac{S}{M}}_{\rightarrow p} \underbrace{\frac{(S-1)}{(M-1)}}_{\rightarrow p} \dots \underbrace{\frac{(S-k+1)}{(M-k+1)}}_{\rightarrow p} \cdot \\ &\quad \cdot \underbrace{\frac{(M-S)}{(M-k)}}_{\rightarrow 1-p} \underbrace{\frac{(M-S-1)}{(M-1)}}_{\rightarrow 1-p} \dots \underbrace{\frac{(M-S-n+k+1)}{(M-n+1)}}_{\rightarrow 1-p} \\ &\xrightarrow[M, S \rightarrow \infty, \frac{S}{M} \rightarrow p]{} \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

2. Falls $n \rightarrow \infty, p \rightarrow 0, np \rightarrow \lambda > 0$, dann gilt

$$\begin{aligned} \binom{n}{k} p^k (1-p)^{n-k} &= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= \frac{1}{k!} \underbrace{\frac{n(n-1)\dots(n-k+1)}{n^k}}_{\rightarrow 1} \cdot \underbrace{(np)^k}_{\rightarrow \lambda^k} \underbrace{\frac{(1-p)^n}{(1-p)^k}}_{\rightarrow e^{-\lambda}} \\ &\rightarrow e^{-\lambda} \frac{\lambda^k}{k!} \text{ für } n \rightarrow \infty, p \rightarrow 0, np \rightarrow \lambda, \end{aligned}$$

weil

$$\frac{(1-p)^n}{(1-p)^k} \underset{n \rightarrow \infty}{\sim} \frac{(1 - \frac{\lambda}{n})^n}{1} \underset{n \rightarrow \infty}{\longrightarrow} e^{-\lambda}, \text{ da } p \sim \frac{\lambda}{n} (n \rightarrow \infty).$$

□

Bemerkung 3.15 1. Die Aussage 1) aus Satz 3.14 wird dann verwendet, wenn M und S in $HG(M, S, n)$ -Verteilung groß werden ($n < 0, 1 \cdot M$). Dabei wird die direkte Berechnung von hypergeometrischen Wahrscheinlichkeiten umständlich.

2. Genauso wird die Poisson-Approximation verwendet, falls n groß und p entweder bei 0 oder bei 1 liegt. Dann können binomiale Wahrscheinlichkeiten nur schwer berechnet werden.
3. Bei allen diskreten Verteilungen ist die zugehörige Verteilungsfunktion eine stückweise konstante Treppenfunktion (vgl. Bsp. 1, 2 im Abschnitt 3.4).

3.3.2 Absolut stetige Verteilungen

Im Gegensatz zu diskreten Zufallsvariablen ist der Wertebereich einer absolut stetigen Zufallsvariablen überabzählbar.

Definition 3.16 Die Verteilung einer Zufallsvariablen X heißt *absolut stetig*, falls die Verteilungsfunktion von F_X folgende Darstellung besitzt:

$$F_X(x) = \int_{-\infty}^x f_X(y) dy, \quad x \in \mathbb{R}, \tag{3.1}$$

wobei $f_X : \mathbb{R} \rightarrow \mathbb{R}_+ = [0, \infty)$ eine Lebesgue-integrierbare Funktion auf \mathbb{R} ist, die *Dichte* der Verteilung von X heißt und das Integral in (3.1) als Lebesgue-Integral zu verstehen ist.

Daher wird oft abkürzend gesagt, dass die Zufallsvariable X absolut stetig (verteilt) mit Dichte f_X ist.

Im folgenden Satz zeigen wir, dass die Verteilung P_X einer absolut stetigen Zufallsvariablen eindeutig durch ihre Dichte f_X bestimmt wird:

Satz 3.17 Sei X eine Zufallsvariable mit Verteilung P_X .

1. X ist absolut stetig verteilt genau dann, wenn

$$P_X(B) = \int_B f_X(y) dy, \quad B \in \mathcal{B}_{\mathbb{R}}. \quad (3.2)$$

2. Seien X und Y absolut stetige Zufallsvariablen mit Dichten f_X, f_Y und Verteilungen P_X und P_Y . Es gilt $P_X = P_Y$ genau dann, wenn $f_X(x) = f_Y(x)$ für fast alle $x \in \mathbb{R}$, d.h. für alle $x \in \mathbb{R} \setminus A$, wobei $A \in \mathcal{B}_{\mathbb{R}}$ und $\int_A dy = 0$ (das Lebesgue-Maß von A ist Null).

Bemerkung 3.18 (*Eigenschaften der absolut stetigen Verteilungen*): Sei X absolut stetig verteilt mit Verteilungsfunktion F_X und Dichte f_X .

1. Für die Dichte f_X gilt: $f_X(x) \geq 0 \quad \forall x$ und $\int_{-\infty}^{\infty} f_X(x) dx = 1$ (vgl. Abb. 3.5).

Diese Eigenschaften sind charakteristisch für eine Dichte, d.h. eine

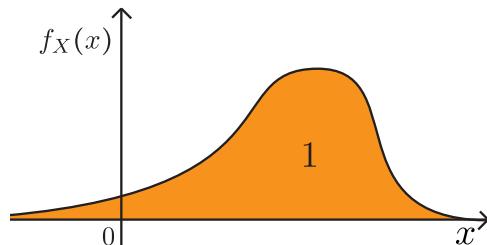


Abbildung 3.5: Die Fläche unter dem Graphen einer Dichtenfunktion ist gleich eins.

beliebige Funktion f , die diese Eigenschaften erfüllt, ist die Dichte einer absolut stetigen Verteilung.

2. Es folgt aus (3.2), dass

- (a) $P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(y) dy, \quad \forall a < b, a, b \in \mathbb{R}$
- (b) $P(X = x) = \int_{\{x\}} f_X(y) dy = 0, \quad \forall x \in \mathbb{R},$
- (c) $f_X(x)\Delta x$ als Wahrscheinlichkeit $P(X \in [x, x + \Delta x])$ interpretiert werden kann, falls f_X stetig in der Umgebung von x und Δx klein ist.

In der Tat, mit Hilfe des Mittelwertsatzes bekommt man

$$\begin{aligned}
 P(X \in [x, x + \Delta x]) &= \int_x^{x+\Delta x} f_X(y) dy \\
 &= f_X(\xi) \cdot \Delta x, \quad \xi \in (x, x + \Delta x) \\
 &\stackrel{\Delta x \rightarrow 0}{=} (f_X(x) + o(1))\Delta x \\
 &= f_X(x) \cdot \Delta x + o(\Delta x),
 \end{aligned}$$

weil $\xi \rightarrow x$ für $\Delta x \rightarrow 0$ und f_X stetig in der Umgebung von x ist.

3. Es folgt aus 2b, dass die Verteilungsfunktion F_X von X eine stetige Funktion ist. F_X kann keine Sprünge haben, weil die Höhe eines Sprunges von F_X in x genau $P(X = x) = 0$ darstellt (vgl. Abb. 3.6).

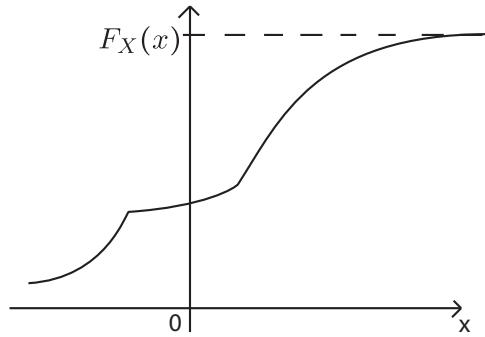


Abbildung 3.6: Eine absolut stetige Verteilungsfunktion

4. Sehr oft wird f_X als (stückweise) stetig angenommen. Dann ist das Integral in Definition 3.16 das (uneigentliche) Riemann–Integral. F_X ist im Allgemeinen nur an jeder Stetigkeitsstelle von ihrer Dichte f_X differenzierbar.
5. In den Anwendungen sind Wertebereiche aller Zufallsvariablen endlich. Somit könnte man meinen, dass für Modellierungszwecke nur diskrete Zufallsvariablen genügen. Falls der Wertebereich einer Zufallsvariable X jedoch sehr viele Elemente x enthält, ist die Beschreibung dieser Zufallsvariable mit einer absolut stetigen Verteilung günstiger, denn man braucht nur eine Funktion f_X (Dichte) anzugeben, statt sehr viele Einzelwahrscheinlichkeiten $p_k = P(X = x_k)$ aus den Daten zu schätzen.

Wichtige absolut stetige Verteilungen

1. *Normalverteilung (Gauß-Verteilung):*

$X \sim N(\mu, \sigma^2)$ für $\mu \in \mathbb{R}$ und $\sigma^2 > 0$, falls

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

(vgl. Abb. 3.7).

μ heißt der *Mittelwert* von X und σ die *Standardabweichung* bzw.

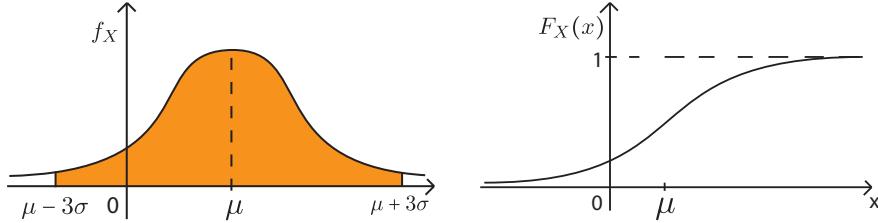


Abbildung 3.7: Dichte und Verteilungsfunktion der $N(\mu, \sigma^2)$ -Verteilung

Streuung, denn es gilt die sogenannte “3 σ -Regel” (Gauß, 1821):

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \geq 0,9973$$

Spezialfall $N(0, 1)$: In diesem Fall sieht die Dichte f_X folgendermaßen aus:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}.$$

Interpretation:

X = Messfehler einer physikalischen Größe μ , σ = Streuung des Messfehlers. Die Verteilungsfunktion $F_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$ kann nicht analytisch berechnet werden (vgl. Abb. 3.7).

2. *Gleichverteilung auf $[a, b]$:*

$X \sim U[a, b]$, $a < b$, $a, b \in \mathbb{R}$, falls

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0, & \text{sonst} \end{cases} \quad (\text{vgl. Abb. 3.8}).$$

Interpretation:

X = Koordinate eines zufällig auf $[a, b]$ geworfenen Punktes (geometrische Wahrscheinlichkeit). Für $F_X(x)$ gilt:

$$F_X(x) = \begin{cases} 1, & x \geq b, \\ \frac{x-a}{b-a}, & x \in [a, b], \\ 0, & x < a \end{cases} \quad (\text{vgl. Abb. 3.8}).$$

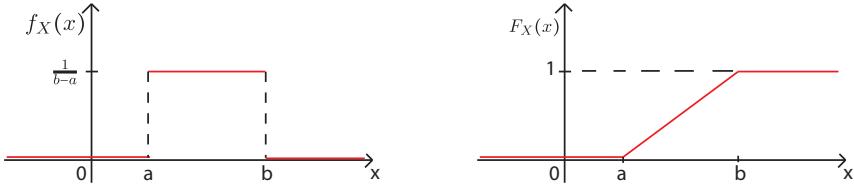


Abbildung 3.8: Dichte und Verteilungsfunktion der Gleichverteilung $U[a, b]$.

3. Exponentialverteilung:

$X \sim Exp(\lambda)$ für $\lambda > 0$, falls

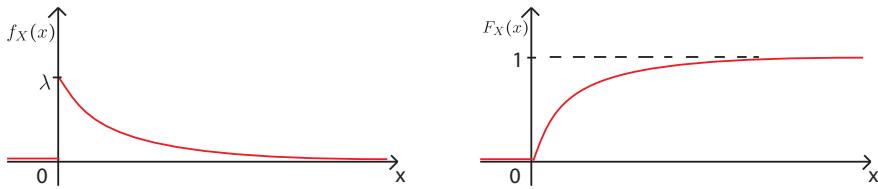


Abbildung 3.9: Dichte und Verteilungsfunktion der Exponentialverteilung $Exp(\lambda)$.

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{sonst} \end{cases} \quad (\text{vgl. Abb. 3.9}).$$

Interpretation:

X = Zeitspanne der fehlerfreien Arbeit eines Geräts, z.B. eines Netz-servers oder einer Glühbirne, λ = Alterungsrate des Geräts. $F_X(x)$ hat folgende Gestalt:

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0 \end{cases} \quad (\text{vgl. Abb. 3.9}).$$

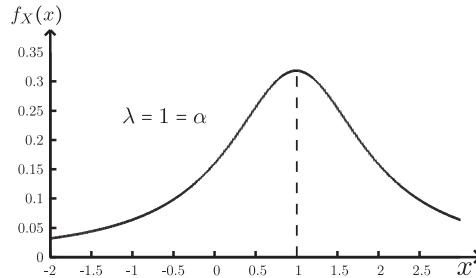
4. Cauchy-Verteilung:

$X \sim Cauchy(\alpha, \lambda)$, falls für $\lambda > 0, \alpha \in \mathbb{R}$

$$f_X(x) = \frac{\lambda}{\pi(\lambda^2 + (x - \alpha)^2)}, \quad x \in \mathbb{R}, \quad \text{vgl. Abb. 3.10}$$

Die Verteilungsfunktion der Cauchy-Verteilung ist dabei

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x - \alpha}{\lambda}\right), \quad x \in \mathbb{R}.$$

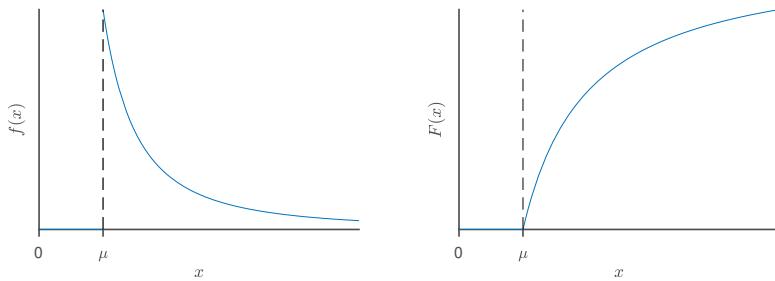
Abbildung 3.10: Dichte der $\text{Cauchy}(\alpha, \lambda)$ -Verteilung*Interpretation:*

Diese Verteilung beschreibt z.B. die Positionen der radioaktiven Teilchen in einem Detektor, sowie die Energie der instabilen Zustände in Kernspaltungsreaktionen (Gesetz von Lorenz).

5. Pareto-Verteilung:

$X \sim \text{Pareto}(\alpha, \mu)$, $\alpha, \mu > 0$, falls

$$f_X(x) = \frac{\alpha\mu^\alpha}{x^{\alpha+1}} I_{[\mu, \infty)}(x), \quad F_X(x) = \left(1 - \frac{\mu^\alpha}{x^\alpha}\right) I_{[\mu, \infty)}(x).$$

Abbildung 3.11: Dichte und Verteilungsfunktion der Pareto(α, μ)-Verteilung.*Interpretation:*

$X = \text{Schadenshöhe einer Police eines Feuerversicherers}$. Da $P(X > x) = (\mu/x)^\alpha$, $x \rightarrow \infty$, nur langsam gegen Null geht (verglichen mit der $N(0, 1)$ -Verteilung), spricht man hier von einer Verteilung mit *schwerem Tailverhalten* oder von einem *gefährlichen Risiko X*.

6. Fréchet-Verteilung:

$X \sim \text{Fréchet}(\mu, \sigma, \alpha)$, $\alpha, \sigma > 0$, $\mu \in \mathbb{R}$, falls

$$f_X(x) = \alpha\sigma^\alpha(x - \mu)^{-\alpha} e^{-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}} I_{[\mu, \infty)}(x).$$

Damit ist die Verteilungsfunktion

$$F_X(x) = e^{-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}} I_{[\mu, \infty)}(x).$$

Die Standard-Fréchet-Verteilung hat Parameterwerte $\sigma = 1$, $\mu = 0$:

$$F_X(x) = e^{-x^{-\alpha}} I_{[0, \infty)}(x).$$

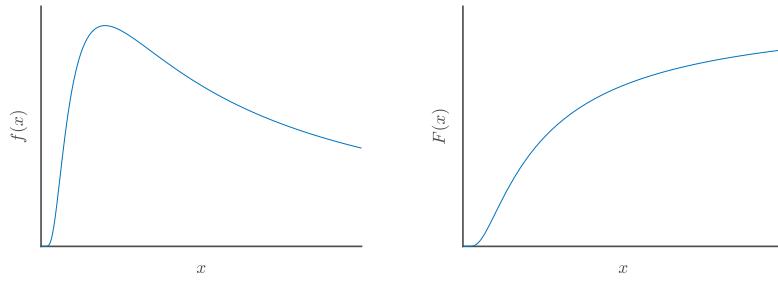


Abbildung 3.12: Dichte- und Verteilungsfunktion der Standard-Fréchet-Verteilung.

Interpretation:

X approximiert normiertes Maximum von n unabhängigen Beobachtungen, die Cauchy- oder Pareto-verteilt sind. Es ist eine der drei möglichen Extremwertverteilungen, zusammen mit der Gumbel- und Weibull-Verteilung.

3.3.3 Mischungen von Verteilungen

Definition 3.19 Sei $\{F_n\}_{n=1}^{\infty}$ eine Folge von Verteilungsfunktionen und sei $\{p_n\}_{n=1}^{\infty}$ eine Zähldichte einer diskreten Zufallsvariablen M . Die Verteilungsfunktion

$$F(x) = \sum_{n=1}^{\infty} p_n F_n(x) \tag{3.3}$$

heißt **Mischung von Verteilungsfunktionen** F_n mit Gewichten p_n .

Übungsaufgabe 3.20 Zeigen Sie, dass F aus (3.3) eine gültige Verteilungsfunktion ist.

Bemerkung 3.21 a) Falls $p_n = 0$ für $n > N$, $N \in \mathbb{N}$, spricht man von einer **endlichen Mischung**

$$F(x) = \sum_{n=1}^N p_n F_n(x)$$

- b) Die Zufallsvariable X mit Verteilungsfunktion F aus (3.3) kann wie folgt simuliert werden:
1. Simuliere die Zufallsvariable M . Sei $M = n$ ihre Realisierung.
 2. Simuliere die Zufallsvariable X mit Verteilungsfunktion F_n .
- c) Generell können Mischungen einer parametrischen Familie $\{F_\mu\}$ von Verteilungen bzgl. des Parameters $\mu \in \mathbb{R}$ definiert werden, wenn μ als Realisierung einer Zufallsvariablen M mit der Verteilungsfunktion Φ_M aufgefasst wird:

$$F(x) = \int_{\mathbb{R}} F_\mu(x) d\Phi_M(\mu), \quad x \in \mathbb{R}.$$

Dabei wird die Borel-Messbarkeit von $F_\mu(x)$ bzgl. μ für alle $x \in \mathbb{R}$ vorausgesetzt.

3.4 Verteilungen von Zufallsvektoren

In der Definition 3.1, 2) wurden Zufallsvektoren bereits eingeführt. Sei (Ω, \mathcal{F}, P) ein beliebiger Wahrscheinlichkeitsraum und $X : \Omega \rightarrow \mathbb{R}^n$ ein n -dimensionaler Zufallsvektor. Bezeichnen wir seine Koordinaten als (X_1, \dots, X_n) . Dann folgt aus Definition 3.1, 2), dass X_i , $i = 1, \dots, n$ Zufallsvariablen sind. Umgekehrt kann man einen beliebigen Zufallsvektor X definieren, indem man seine Koordinaten $X_1 \dots X_n$ als Zufallsvariablen auf demselben Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) einführt (Übungsaufgabe).

Definition 3.22 Sei $X = (X_1, \dots, X_n)$ ein Zufallsvektor auf (Ω, \mathcal{F}, P) .

1. Die *Verteilung* von X ist die Mengenfunktion $P_X : \mathcal{B}_{\mathbb{R}^n} \rightarrow [0, 1]$ mit $P_X(B) = P(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\})$, $B \in \mathcal{B}_{\mathbb{R}^n}$.
2. Die *Verteilungsfunktion* $F_X : \mathbb{R}^n \rightarrow [0, 1]$ von X ist gegeben durch $F_X(x_1, \dots, x_n) = P(X_1 \leq x_1, \dots, X_N \leq x_n)$ $x_1, \dots, x_n \in \mathbb{R}$. Sie heißt manchmal auch die *gemeinsame* oder die *multivariate Verteilungsfunktion* von X , um sie von folgenden *marginalen Verteilungsfunktionen* zu unterscheiden.

3. Sei $\{i_1, \dots, i_k\}$ ein Teilvektor von $\{1, \dots, n\}$. Die multivariate Verteilungsfunktion F_{i_1, \dots, i_k} des Zufallsvektors $(X_{i_1}, \dots, X_{i_k})$ heißt *marginalen Verteilungsfunktion* von X . Insbesondere für $k = 1$ und $i_1 = i$ spricht man von den so genannten *Randverteilungen*:

$$F_{X_i}(x) = P(X_i \leq x), \quad i = 1, \dots, n.$$

Satz 3.23 (*Eigenschaften multivariater Verteilungsfunktionen*):

Sei $F_X : \mathbb{R}^n \rightarrow [0, 1]$ die Verteilungsfunktion eines Zufallsvektors $X = (X_1, \dots, X_n)$. Dann gelten folgende Eigenschaften:

1. *Asymptotik*:

$$\lim_{x_i \rightarrow -\infty} F_X(x_1, \dots, x_n) = 0, \quad \forall i = 1, \dots, n \quad \forall x_1, \dots, x_n \in \mathbb{R},$$

$$\lim_{x_1, \dots, x_n \rightarrow +\infty} F_X(x_1, \dots, x_n) = 1,$$

$$\lim_{x_j \rightarrow +\infty \forall j \notin \{i_1, \dots, i_k\}} F_X(x_1, \dots, x_n) = F_{(X_{i_1}, \dots, X_{i_k})}(x_{i_1}, \dots, x_{i_k}),$$

wobei $F_{(X_{i_1}, \dots, X_{i_k})}(x_{i_1}, \dots, x_{i_k})$ die Verteilungsfunktion der marginalen Verteilung von

$$(X_{i_1} \dots X_{i_k})\{i_1, \dots, i_k\} \subset \{1, \dots, n\} \text{ ist.}$$

Insbesondere gilt

$$\lim_{x_j \rightarrow +\infty, j \neq i} F_X(x_1, \dots, x_n) = F_{X_i}(x_i), \quad \forall i = 1, \dots, n$$

(Randverteilungsfunktion).

2. *Monotonie*: $\forall (x_1, \dots, x_n) \in \mathbb{R}^n \quad \forall h_1, \dots, h_n \geq 0$

$$F_X(x_1 + h_1, \dots, x_n + h_n) \geq F_X(x_1, \dots, x_n).$$

3. *Rechtsseitige Stetigkeit*:

$$F_X(x_1, \dots, x_n) = \lim_{y_i \rightarrow x_i + 0, i=1, \dots, n} F_X(y_1, \dots, y_n) \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Definition 3.24 Die Verteilung eines Zufallsvektors $X = (X_1, \dots, X_n)$ heißt

1. *diskret*, falls eine höchstens abzählbare Menge $C \subseteq \mathbb{R}^n$ existiert, für die $P(X \in C) = 1$ gilt. Die Familie von Wahrscheinlichkeiten

$$\{P(X = x), x \in C\}$$

heißt dann *Wahrscheinlichkeitsfunktion* oder *Zähldichte* von X .

2. *absolut stetig*, falls eine Funktion $f_X : \mathbb{R}^n \rightarrow [0, 1]$ existiert, die Lebesgue-integrierbar auf \mathbb{R}^n ist und für die gilt

$$F_X(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f_X(y_1, \dots, y_n) dy_n \dots dy_1,$$

$\forall (x_1, \dots, x_n) \in \mathbb{R}^n$. f_X heißt *Dichte* der gemeinsamen Verteilung von X .

Lemma 3.25 Sei $X = (X_1, \dots, X_n)$ ein diskreter (bzw. absolut stetiger) Zufallsvektor mit Zähldichte $P(X = x)$ (bzw. Dichte $f_X(x)$). Dann gilt:

1. Die Verteilung P_X von X ist gegeben durch

$$P_X(B) = \sum_{x \in B} P(X = x) \text{ bzw. } P_X(B) = \int_B f_X(x) dx, \quad B \in \mathcal{B}_{\mathbb{R}^n}.$$

2. Die Koordinaten X_i , $i = 1, \dots, n$ sind ebenfalls diskrete bzw. absolut stetige Zufallsvariablen mit der Randzähldichte

$$P(X_i = x) =$$

$$= \sum_{(y_1, \dots, y_{i-1}, x, y_{i+1}, \dots, y_n) \in C} P(X_1 = y_1, \dots, X_{i-1} = y_{i-1}, X_i = x, X_{i+1} = y_{i+1}, \dots, X_n = y_n)$$

bzw. Randdichte

$$f_{X_i}(x) = \int_{\mathbb{R}^{n-1}} f_X(y_1, \dots, y_{i-1}, x, y_{i+1}, \dots, y_n) dy_1 \dots dy_{i-1} dy_{i+1} \dots dy_n$$

$$\forall x \in \mathbb{R}.$$

Beweis

- Folgt aus dem eindeutigen Zusammenhang zwischen einer Verteilung und ihrer Verteilungsfunktion.
- Die Aussage für diskrete Zufallsvektoren ist trivial. Sei nun $X = (X_1, \dots, X_n)$ absolut stetig. Dann folgt aus Satz 3.23

$$\begin{aligned} F_{X_i}(x) &= \lim_{y_j \rightarrow +\infty, j \neq i} F_X(x_1 \dots x_{i-1}, x, x_{i+1}, \dots, x_n) \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \int_{-\infty}^x \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_X(y_1, \dots, y_n) dy_n \dots dy_1 \\ &\stackrel{\text{S. v. Fubini}}{=} \int_{-\infty}^x \underbrace{\left(\int_{\mathbb{R}^{n-1}} f_X(y_1, \dots, y_n) dy_1 \dots dy_{i-1} dy_{i+1} \dots dy_n \right)}_{f_{X_i}(y_i)} dy_i \end{aligned}$$

Somit ist X_i absolut stetig verteilt mit Dichte f_{X_i} .

□

Beispiel 3.26 Verschiedene Zufallsvektoren:

1. *Polynomiale Verteilung:*

$X = (X_1, \dots, X_k) \sim \text{Polynom}(n, p_1, \dots, p_k)$, $n \in \mathbb{N}$, $p_i \in [0, 1]$,
 $i = 1, \dots, k$, $\sum_{i=1}^n p_i = 1$, falls X diskret verteilt ist mit Zähldichte

$$P(X = x) = P(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$$

$\forall x = (x_1, \dots, x_k)$ mit $x_i \in \mathbb{N} \cup \{0\}$ und $\sum_{i=1}^k x_i = n$. Die polynomiale Verteilung ist das k -dimensionale Analogon der Binomialverteilung. So sind die Randverteilungen von $X_i \sim \text{Bin}(n, p_i)$, $i = 1, \dots, k$. (Bitte prüfen Sie dies als Übungsaufgabe!). Es gilt $P(\sum_{i=1}^k X_i = n) = 1$.

Interpretation:

Es werden n Versuche durchgeführt. In jedem Versuch kann eines aus insgesamt k Merkmalen auftreten. Sei p_i die Wahrscheinlichkeit des Auftretens von Merkmal i in einem Versuch. Sei

$$X_i = \#\{\text{Auftretens von Merkmal } i \text{ in } n \text{ Versuchen}\}, \quad i = 1, \dots, k.$$

Dann ist $X = (X_1, \dots, X_k) \sim \text{Polynom}(n, p_1, \dots, p_k)$.

2. *Gleichverteilung:*

$X \sim \mathcal{U}(A)$, wobei $A \subset \mathbb{R}^n$ eine beschränkte Borel–Teilmenge von \mathbb{R}^n ist, falls $X = (X_1, \dots, X_n)$ absolut stetig verteilt mit der Dichte

$$f_X(x_1, \dots, x_n) = \begin{cases} \frac{1}{|A|}, & (x_1, \dots, x_n) \in A, \\ 0, & \text{sonst,} \end{cases}$$

ist (vgl. Abb. 3.13). Im Spezialfall $A = \prod_{i=1}^n [a_i, b_i]$ (Parallelepiped) sind alle Randverteilungen von X_i ebenso Gleichverteilungen:

$$X_i \sim U[a_i, b_i], \quad i = 1, \dots, n.$$

Interpretation:

$X = (X_1, \dots, X_n)$ sind Koordinaten eines zufälligen Punktes, der gleichwahrscheinlich auf A geworfen wird. Dies ist die geometrische Wahrscheinlichkeit, denn $P(X \in B) = \int_B f_X(y) dy = \frac{|B|}{|A|}$ für $B \in \mathcal{B}_{\mathbb{R}^n} \cap A$.

3. *Multivariate Normalverteilung:*

$X = (X_1, \dots, X_n) \sim N(\mu, K)$, $\mu \in \mathbb{R}^n$, K eine positiv definite $(n \times n)$ –Matrix, falls X absolut stetig verteilt mit Dichte

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \cdot \det K}} \exp\left\{-\frac{1}{2}(x - \mu)^T K^{-1}(x - \mu)\right\}, \quad x \in \mathbb{R}^n$$

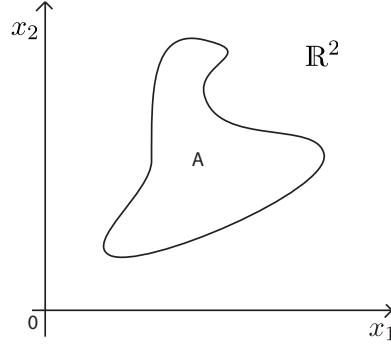


Abbildung 3.13: Wertebereich A einer zweidimensionalen Gleichverteilung

ist.

Spezialfall zweidimensionale Normalverteilung:

Falls $n = 2$ und

$$\mu = (\mu_1, \mu_2), K = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

dann gilt $\det K = \sigma_1^2\sigma_2^2(1 - \rho^2)$ und

$$f_X(x_1, x_2) = \frac{1}{\sqrt{1 - \rho^2} \cdot 2\pi\sigma_1\sigma_2} \times \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left(\frac{(x_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right) \right\},$$

$(x_1, x_2) \in \mathbb{R}^2$, weil

$$K^{-1} = \frac{1}{1 - \rho^2} \begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}, \quad (\text{vgl. Abb. 3.14}).$$

Übungsaufgabe 3.27 Zeigen Sie, dass $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2$. Diese Eigenschaft der Randverteilungen gilt für alle $n \geq 2$. Somit ist die multivariate Normalverteilung ein mehrdimensionales Analogon der eindimensionalen $N(\mu, \sigma^2)$ -Verteilung.

Interpretation:

Man feuert eine Kanone auf das Ziel mit Koordinaten (μ_1, μ_2) . Dann sind $X = (X_1, X_2)$ die Koordinaten des Treffers. Durch die Streuung wird $(X_1, X_2) = (\mu_1, \mu_2)$ nur im Mittel. σ_1^2 und σ_2^2 sind Maße für die Genauigkeit des Feuers.

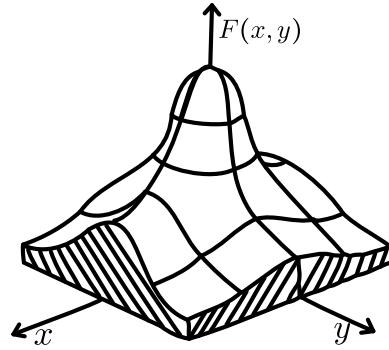


Abbildung 3.14: Grafik der Dichte einer zweidimensionalen Normalverteilung

3.5 Stochastische Unabhängigkeit

3.5.1 Unabhängige Zufallsvariablen

Definition 3.28

- Seien X_1, \dots, X_n Zufallsvariablen definiert auf dem Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) . Sie heißen *unabhängig*, falls

$$F_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdot \dots \cdot F_{X_n}(x_n), \quad x_1, \dots, x_n \in \mathbb{R}$$

oder äquivalent dazu,

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = P(X_1 \leq x_1) \cdot \dots \cdot P(X_n \leq x_n).$$

- Sei $\{X_n\}_{n \in \mathbb{N}}$ eine Folge von Zufallsvariablen definiert auf dem Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) . Diese Folge besteht aus *unabhängigen Zufallsvariablen*, falls $\forall k \in \mathbb{N} \ \forall i_1 < i_2 < \dots < i_k \ X_{i_1}, X_{i_2}, \dots, X_{i_k}$ unabhängige Zufallsvariablen (im Sinne der Definition 1) sind.

Lemma 3.29 Die Zufallsvariablen X_1, \dots, X_n sind genau dann unabhängig, wenn für alle $B_1, \dots, B_n \in \mathcal{B}_{\mathbb{R}}$ gilt

$$P(X_1 \in B_1, \dots, X_n \in B_n) = P(X_1 \in B_1) \cdot \dots \cdot P(X_n \in B_n) \quad (3.4)$$

Satz 3.30 (Charakterisierung der Unabhängigkeit von Zufallsvariablen)

- Sei (X_1, \dots, X_n) ein diskret verteilter Zufallsvektor mit dem Wertebereich C . Seine Koordinaten X_1, \dots, X_n sind genau dann unabhängig, wenn

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i)$$

$$\forall (x_1, \dots, x_n) \in C.$$

2. Sei (X_1, \dots, X_n) ein absolut stetiger Zufallsvektor mit der Dichte $f_{(X_1 \dots X_n)}$ und Randdichten f_{X_i} . Seine Koordinaten X_1, \dots, X_n sind genau dann unabhängig, wenn

$$f_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$$

für fast alle $(x_1, \dots, x_n) \in \mathbb{R}^n$.

Beispiel 3.31

1. *Multivariate Normalverteilung:*

Mit Hilfe des Satzes 3.30 kann gezeigt werden, dass die Komponenten X_1, \dots, X_n von

$$X = (X_1, \dots, X_n) \sim N(\mu, K)$$

genau dann unabhängig sind, wenn $k_{ij} = 0$, $i \neq j$, wobei $K = (k_{ij})_{i,j=1}^n$. Insbesondere gilt im zweidimensionalen Fall (vgl. Bsp. 3 Seite 44), dass X_1 und X_2 unabhängig sind, falls $\rho = 0$.

Übungsaufgabe 3.32 Zeigen Sie es!

2. *Multivariate Gleichverteilung:*

Die Komponenten des Vektors $X = (X_1, \dots, X_n) \sim \mathcal{U}(A)$ sind genau dann unabhängig, falls $A = \prod_{i=1}^n [a_i, b_i]$ ist. In der Tat gilt dann

$$f_X(x_1, \dots, x_n) = \begin{cases} \frac{1}{|A|} = \frac{1}{\prod_{i=1}^n (b_i - a_i)} = \prod_{i=1}^n \frac{1}{b_i - a_i} = \prod_{i=1}^n f_{X_i}(x_i), & x \in A \\ 0, & \text{sonst} \end{cases}$$

mit $x = (x_1, \dots, x_n)$, wobei

$$f_X(x_i) = \begin{cases} 0, & \text{falls } x_i \notin [a_i, b_i] \\ \frac{1}{b_i - a_i}, & \text{sonst.} \end{cases}$$

Implizit haben wir an dieser Stelle benutzt, dass

$$X_i \sim \mathcal{U}[a_i, b_i], i = 1, \dots, n$$

(Herleitung: $\int_{\mathbb{R}^{n-1}} f_X(x) dx_n \dots dx_{i+1} dx_{i-1} \dots dx_1 = \frac{1}{b_i - a_i}, x_i \in [a_i, b_i]$).

Übungsaufgabe 3.33

Zeigen Sie die Notwendigkeit der Bedingung $A = \prod_{i=1}^n [a_i, b_i]!$

3. Es gibt Beispiele von Zufallsvariablen X_1 und X_2 , die stochastisch abhängig von einander sind, so dass X_1^2 stochastisch unabhängig von X_2^2 ist.

Unterschied: Kausale bzw. stochastische Abhängigkeit!

Definition 3.34 Eine Funktion $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $n, m \in \mathbb{N}$ heißt *Borelsche Funktion*, falls sie $\mathcal{B}_{\mathbb{R}^n}$ -messbar ist, d.h. $\forall B \in \mathcal{B}_{\mathbb{R}^m}$ ist $\varphi^{-1}(B) \in \mathcal{B}_{\mathbb{R}^n}$.

Bemerkung 3.35

1. Sei $X = (X_1, \dots, X_n)^T$ ein Zufallsvektor, und $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ sei Borelsch. Dann $Y = (Y_1, \dots, Y_m)^T = \phi(X)$ ebenfalls ein Zufallsvektor.
2. Seien X_1, X_2 Zufallsvariablen auf (Ω, \mathcal{F}, P) und $\varphi_1, \varphi_2 : \mathbb{R} \rightarrow \mathbb{R}$ Borelsche Funktionen. Falls X_1 und X_2 stochastisch unabhängig sind, dann sind auch $\varphi_1(X_1)$ und $\varphi_2(X_2)$ stochastisch unabhängig.

3.6 Funktionen von Zufallsvektoren

Lemma 3.36 Jede stetige Funktion $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ ist Borel-messbar. Insbesondere sind Polynome $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ der Form

$$\varphi(x_1, \dots, x_n) = \sum_{i=0}^k a_i x_1^{m_{1i}} \dots x_n^{m_{ni}},$$

$k \in \mathbb{N}$, $a_0, a_1, \dots, a_n \in \mathbb{R}$, $m_{1i}, \dots, m_{ni} \in \mathbb{N} \cup \{0\}$, $i = 1, \dots, k$ Borel-messbar.

Satz 3.37 (Transformationssatz)

Sei X eine Zufallsvariable auf dem Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) .

1. Falls die Abbildung $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ stetig und streng monoton wachsend ist, dann gilt $F_{\varphi(X)}(x) = F_X(\varphi^{-1}(x)) \quad \forall x \in \mathbb{R}$, wobei φ^{-1} die Umkehrfunktion von φ ist. Falls $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ stetig und streng monoton fallend ist, dann gilt $F_{\varphi(X)}(x) = 1 - F_X(\varphi^{-1}(x)) + P(X = \varphi^{-1}(x))$, $x \in \mathbb{R}$.
2. Falls X absolut stetig verteilt mit Dichte f_X ist und $C \subset \mathbb{R}$ eine offene Menge mit $P(X \in C) = 1$ ist, dann ist $\varphi(X)$ absolut stetig verteilt mit Dichte $f_{\varphi(X)}(y) = f_X(\varphi^{-1}(y)) \cdot |(\varphi^{-1})'(y)|$, $y \in \varphi(C)$, falls φ eine auf C stetig differenzierbare Funktion mit $\varphi'(x) \neq 0$, $x \in C$ ist.

Beweis

1. Falls φ monoton wachsend ist, gilt für $x \in \mathbb{R}$, dass $F_{\varphi(X)}(x) = P(\varphi(X) \leq x) = P(X \leq \varphi^{-1}(x)) = F_X(\varphi^{-1}(x))$. Für φ monoton fallend gilt für $x \in \mathbb{R}$

$$\begin{aligned} F_{\varphi(X)}(x) &= P(\varphi(X) \leq x) = P(X \geq \varphi^{-1}(x)) \\ &= 1 - P(X < \varphi^{-1}(x)) = 1 - F_X(\varphi^{-1}(x)) + P(X = \varphi^{-1}(x)). \end{aligned}$$

2. Nehmen wir o.B.d.A. an, dass $C = \mathbb{R}$ und $\varphi'(x) > 0 \forall x \in \mathbb{R}$.
Für $\varphi'(x) < 0$ verläuft der Beweis analog. Aus Punkt 1) folgt

$$\begin{aligned} F_{\varphi(X)}(x) &= F_X(\varphi^{-1}(x)) \\ &= \int_{-\infty}^{\varphi^{-1}(x)} f_X(y) dy \\ &\underset{\varphi^{-1}(t)=y}{=} \int_{-\infty}^x f_X(\varphi^{-1}(t)) \cdot |(\varphi^{-1})'(t)| dt, \quad x \in \mathbb{R}. \end{aligned}$$

Hieraus folgt $f_{\varphi(X)}(t) = f_X(\varphi^{-1}(t)) \cdot |(\varphi^{-1})'(t)|$, $t \in \mathbb{R}$.

□

Satz 3.38 (lineare Transformation)

Sei $X : \Omega \rightarrow \mathbb{R}$ eine Zufallsvariable und $a, b \in \mathbb{R}$, $a \neq 0$. Dann gilt Folgendes:

1. Die Verteilungsfunktion der Zufallsvariable $aX + b$ ist gegeben durch

$$F_{aX+b}(x) = \begin{cases} F_X\left(\frac{x-b}{a}\right), & a > 0 \\ 1 - F_X\left(\frac{x-b}{a}\right) + P\left(X = \frac{x-b}{a}\right), & a < 0. \end{cases}$$

2. Falls X absolut stetig verteilt mit Dichte f_X ist, dann ist $aX + b$ ebenfalls absolut stetig verteilt mit Dichte

$$f_{aX+b}(x) = \frac{1}{|a|} f_X\left(\frac{x-b}{a}\right).$$

Beweis 1. Der Fall $a > 0$ ($a < 0$) folgt aus dem Satz 3.37, 1), weil $\varphi(x) = aX + b$ stetig und monoton wachsend bzw. fallend ist.

2. Folgt aus dem Satz 3.37, 2), weil $\varphi(x) = aX + b$ stetig differenzierbar auf $C = \mathbb{R}$ (offen) mit $\varphi'(x) = a \neq 0$ ist.

□

Satz 3.39 (Quadrierung)

Sei X eine Zufallsvariable auf (Ω, \mathcal{F}, P) .

1. Die Verteilungsfunktion von X^2 ist gegeben durch

$$F_{X^2}(x) = \begin{cases} F_X(\sqrt{x}) - F_X(-\sqrt{x}) + P(X = -\sqrt{x}), & \text{falls } x \geq 0 \\ 0, & \text{sonst}. \end{cases}$$

2. Falls X absolut stetig verteilt mit Dichte f_X ist, dann ist auch X^2 absolut stetig verteilt mit Dichte

$$f_{X^2}(x) = \begin{cases} \frac{1}{2\sqrt{x}} (f_X(\sqrt{x}) + f_X(-\sqrt{x})), & x > 0 \\ 0, & \text{sonst}. \end{cases}$$

Beweis 1. Für $x < 0$ gilt $F_{X^2}(x) = P(X^2 \leq x) = 0$.

Für $x \geq 0$ gilt

$$\begin{aligned} F_{X^2}(x) &= P(X^2 \leq x) = P(|X| \leq \sqrt{x}) \\ &= P(-\sqrt{x} \leq X \leq \sqrt{x}) = P(X \leq \sqrt{x}) - P(X < -\sqrt{x}) \\ &= F_X(\sqrt{x}) - F_X(-\sqrt{x}) + P(X = -\sqrt{x}). \end{aligned}$$

2. Wegen 1) gilt $F_{X^2}(x) = F_X(\sqrt{x}) - F_X(-\sqrt{x})$, da im absolut stetigen Fall $P(X = -\sqrt{x}) = 0 \quad \forall x \geq 0$. Daher gilt

$$\begin{aligned} F_{X^2}(x) &= \int_{-\sqrt{x}}^{\sqrt{x}} f_X(y) dy = \\ &= \int_0^{\sqrt{x}} f_X(y) dy + \int_{-\sqrt{x}}^0 f_X(y) dy \\ &\stackrel{y=\sqrt{t} \text{ oder } y=-\sqrt{t}}{=} \int_0^x \frac{1}{2\sqrt{t}} f_X(\sqrt{t}) dt + \int_0^x \frac{1}{2\sqrt{t}} f_X(-\sqrt{t}) dt \\ &= \int_0^x \frac{1}{2\sqrt{t}} (f_X(\sqrt{t}) + f_X(-\sqrt{t})) dt, \quad \forall x \geq 0. \end{aligned}$$

Daher gilt die Aussage 2) des Satzes. \square

Beispiel 3.40

1. Falls $X \sim N(0, 1)$, dann ist $Y = \mu + \sigma X \sim N(\mu, \sigma^2)$.
2. Falls $X \sim N(\mu, \sigma^2)$, dann heißt die Zufallsvariable $Y = e^X$ lognormalverteilt mit Parametern μ und σ^2 . Diese Verteilung wird sehr oft in ökonometrischen Anwendungen benutzt.

Zeigen Sie, dass die Dichte von Y durch

$$f_Y(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

gegeben ist.

3. Falls $X \sim N(0, 1)$, dann heißt $Y = X^2 \sim \chi_1^2$ -verteilt (Chi-Quadrat-Verteilung) mit einem Freiheitsgrad.

Zeigen Sie, dass die Dichte von Y durch

$$f_Y(x) = \begin{cases} \frac{1}{\sqrt{2\pi x}} e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

gegeben ist.

Die χ^2 -Verteilung wird in der Statistik sehr oft als die sogenannte *Prüfverteilung* bei der Konstruktion von statistischen Tests und Konfidenzintervallen verwendet.

Satz 3.41 (*Summe von unabhängigen Zufallsvariablen*)

Sei $X = (X_1, X_2)$ ein absolut stetiger Zufallsvektor mit Dichte f_X . Dann gilt Folgendes:

1. Die Zufallsvariable $Y = X_1 + X_2$ ist absolut stetig mit Dichte

$$f_Y(x) = \int_{\mathbb{R}} f_X(y, x-y) dy, \quad \forall x \in \mathbb{R}. \quad (3.5)$$

2. Falls X_1 und X_2 stochastisch unabhängig sind, dann heißt der Spezialfall

$$f_Y(x) = \int_{\mathbb{R}} f_{X_1}(y) f_{X_2}(x-y) dy, \quad x \in \mathbb{R}$$

von (3.5) *Faltungsformel*.

Beweis

2) ergibt sich aus 1) für $f_X(x, y) = f_{X_1}(x) \cdot f_{X_2}(y) \quad \forall x, y \in \mathbb{R}$.

Beweisen wir also 1):

$$\begin{aligned} P(Y \leq t) &= P(X_1 + X_2 \leq t) = \int_{(x,y) \in \mathbb{R}^2: x+y \leq t} f_X(x, y) dx dy \\ &= \underbrace{\int_{-\infty}^{\infty} \int_{-\infty}^{t-x} f_X(x, y) dy dx}_{y \mapsto z = x+y} \\ &\stackrel{=}{} \underbrace{\int_{-\infty}^{\infty} \int_{-\infty}^t f_X(x, z-x) dz dx}_{\substack{\text{Fubini} \\ y \mapsto z = x+y}} \\ &\stackrel{=}{} \underbrace{\int_{-\infty}^t \int_{\mathbb{R}} f_X(x, z-x) dx dz}_{t \in \mathbb{R}} = f_Y(z). \end{aligned}$$

Somit ist $f_Y(z) = \int_{\mathbb{R}} f_X(x, z-x) dx$ die Dichte von $Y = X_1 + X_2$. \square

Folgerung 3.42 (*Faltungsstabilität der Normalverteilung*):

Falls die Zufallsvariablen X_1, \dots, X_n stochastisch unabhängig mit

$$X_i \sim N(\mu_i, \sigma_i^2) \quad i = 1, \dots, n$$

sind, dann gilt

$$X_1 + \dots + X_n \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

Satz 3.43 (Produkt und Quotient von Zufallsvariablen):

Sei $X = (X_1, X_2)$ ein absolut stetiger Zufallsvektor mit Dichte f_X . Dann gilt Folgendes:

1. Die Zufallsvariable $Y = X_1 \cdot X_2$ und $Z = \frac{X_1}{X_2}$ sind absolut stetig verteilt mit Dichten

$$f_Y(x) = \int_{\mathbb{R}} \frac{1}{|t|} f_X(x/t, t) dt,$$

bzw.

$$f_Z(x) = \int_{\mathbb{R}} |t| f_X(x \cdot t, t) dt, x \in \mathbb{R}.$$

2. Falls X_1 und X_2 stochastisch unabhängig sind, dann gilt der Spezialfall der obigen Formeln

$$f_Y(x) = \int_{\mathbb{R}} \frac{1}{|t|} f_{X_1}(x/t) f_{X_2}(t) dt,$$

bzw.

$$f_Z(x) = \int_{\mathbb{R}} |t| f_{X_1}(x \cdot t) f_{X_2}(t) dt, x \in \mathbb{R}.$$

Beispiel 3.44 Zeigen Sie, dass $X_1/X_2 \sim \text{Cauchy}(0,1)$, falls $X_1, X_2 \sim N(0, 1)$ und stochastisch unabhängig sind:

$$f_{X_1/X_2}(x) = \frac{1}{\pi(x^2 + 1)}, \quad x \in \mathbb{R}.$$

Bemerkung 3.45 Da X_1 und X_2 absolut stetig verteilt sind, tritt das Ereignis $\{X_2 = 0\}$ mit Wahrscheinlichkeit Null ein. Daher ist $X_1(\omega)/X_2(\omega)$ wohl definiert für fast alle $\omega \in \Omega$. Für $\omega \in \Omega : X_2(\omega) = 0$ kann $X_1(\omega)/X_2(\omega)$ z.B. als 1 definiert werden. Dies ändert den Ausdruck der Dichte von X_1/X_2 nicht.

Kapitel 4

Momente von Zufallsvariablen

Weitere wichtige Charakteristiken von Zufallsvariablen sind ihre so genannten *Momente*, darunter der Erwartungswert und die Varianz. Zusätzlich wird in diesem Kapitel die Kovarianz von zwei Zufallsvariablen als Maß ihrer Abhängigkeit diskutiert. Um diese Charakteristiken einführen zu können, brauchen wir die Definition des Lebesgue–Integrals auf beliebigen messbaren Räumen.

Beispiel 4.1 Sei X eine diskrete Zufallsvariable mit dem endlichen Wertebereich $C = \{x_1, \dots, x_n\}$ und Zähldichte $\{p_i\}_{i=1}^n$, wobei $p_i = P(X = x_i)$, $i = 1, \dots, n$. Wie soll der Mittelwert von X berechnet werden? Aus der Antike sind drei Ansätze zur Berechnung des Mittels von n Zahlen bekannt:

- das arithmetische Mittel: $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$
- das geometrische Mittel: $\bar{g}_n = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$
- das harmonische Mittel: $\bar{h}_n = \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \right)^{-1}$

Um \bar{g}_n und \bar{h}_n berechnen zu können, ist die Bedingung $x_i > 0$ bzw. $x_i \neq 0$ $i = 1, \dots, n$ eine wichtige Voraussetzung. Wir wollen jedoch diese Mittel für beliebige Wertebereiche einführen. Somit fallen diese zwei Möglichkeiten schon weg. Beim arithmetischen Mittel werden beliebige x_i zugelassen, jedoch alle gleich gewichtet, unabhängig davon, ob der Wert x_{i_0} wahrscheinlicher als alle anderen Werte ist und somit häufiger in den Experimenten vorkommt.

Deshalb ist es naheliegend, das gewichtete Mittel $\sum_{i=1}^n x_i \omega_i$ zu betrachten, $\forall i \omega_i \geq 0$, $\sum_{i=1}^n \omega_i = 1$, wobei das Gewicht ω_i die relative Häufigkeit des

Vorkommens des Wertes x_i in den Experimenten ausdrückt. Somit ist es natürlich, $\omega_i = p_i$ zu setzen, $i = 1, \dots, n$, und schreiben $EX = \sum_{i=1}^n x_i p_i$. Dieses Mittel wird “Erwartungswert der Zufallsvariable X” genannt. Der Buchstabe “ \mathbb{E} ” kommt aus dem Englischen: “Expectation”. Für die Gleichverteilung auf $\{x_1, \dots, x_n\}$, d.h. $p_i = \frac{1}{n}$, stimmt $\mathbb{E}X$ mit dem arithmetischen Mittel \bar{x}_n überein. Wie wir bald sehen werden, kann

$$\mathbb{E}X = \sum_{i=1}^n x_i P(X = x_i)$$

geschrieben werden.

4.1 Erwartungswert

Somit können wir folgende Definition angeben:

Definition 4.2

- Sei X eine diskret verteilte Zufallsvariable mit Wertebereich C und Zähldichte $P_X(x)$. Der Erwartungswert von X ist definiert als

$$\mathbb{E}X = \sum_{x \in C} x P_X(x),$$

falls $\sum_{x \in C} |x| P_X(x) < \infty$.

- Sei X absolut stetig verteilt mit Dichte f_X . Der Erwartungswert von X ist definiert als

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

falls $\underbrace{\int_{\mathbb{R}} |x| f_X(x) dx}_{\mathbb{E}|X|} < \infty$

Satz 4.3 (Eigenschaften des Erwartungswertes)

Seien X, Y integrierbare Zufallsvariablen auf dem Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) . Dann gilt Folgendes:

- Falls $X(\omega) = I_A(\omega)$ für ein $A \in \mathcal{F}$, dann gilt $\mathbb{E}X = P(A)$.
- Falls $X(\omega) \geq 0$, für fast alle $\omega \in \Omega$, dann ist $\mathbb{E}X \geq 0$.
- Additivität:* für beliebige $a, b \in \mathbb{R}$ gilt $\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$.

4. *Monotonie:* Falls $X \geq Y$ für fast alle $\omega \in \Omega$ (man sagt dazu “*fast sicher*” und schreibt “*f.s.*”), dann gilt $\mathbb{E}X \geq \mathbb{E}Y$.
Falls $0 \leq X \leq Y$ fast sicher und lediglich vorausgesetzt wird, dass Y integrierbar ist, dann ist auch X integrierbar.
5. $|\mathbb{E}X| \leq \mathbb{E}|X|$.
6. Falls X fast sicher auf Ω beschränkt ist, dann ist X integrierbar.
7. Falls X und Y unabhängig sind, dann gilt $\mathbb{E}(XY) = \mathbb{E}X \cdot \mathbb{E}Y$.
8. Falls $X \geq 0$ fast sicher und $\mathbb{E}X = 0$, dann gilt $X = 0$ fast sicher.

Bemerkung 4.4

1. Aus dem Satz 4.3, 3) und 7) folgt per Induktion, dass
 - (a) Falls X_1, \dots, X_n integrierbare Zufallsvariablen sind und $a_1, \dots, a_n \in \mathbb{R}$, dann ist $\sum_{i=1}^n a_i X_i$ eine integrierbare Zufallsvariable und es gilt

$$\mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mathbb{E}X_i .$$
 - (b) Falls X_1, \dots, X_n zusätzlich unabhängig sind und $\mathbb{E}|X_1 \dots X_n| < \infty$, dann gilt

$$\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathbb{E}X_i .$$
2. Die Aussage 7) des Satzes 4.3 gilt nicht in umgekehrte Richtung:
aus $\mathbb{E}(XY) = \mathbb{E}X \cdot \mathbb{E}Y$ folgt im Allgemeinen nicht die Unabhängigkeit von Zufallsvariablen X und Y . Als Illustration betrachten wir folgendes Beispiel:
 - (a) Seien X_1, X_2 unabhängige Zufallsvariablen mit $\mathbb{E}X_1 = \mathbb{E}X_2 = 0$. Setzen wir $X = X_1$ und $Y = X_1 \cdot X_2$. X und Y sind abhängig und dennoch

$$\mathbb{E}(XY) = \mathbb{E}(X_1^2 X_2) = \mathbb{E}X_1^2 \cdot \mathbb{E}X_2 = 0 = \mathbb{E}X \cdot \mathbb{E}Y = 0 \cdot \mathbb{E}Y = 0 .$$
 - (b) Falls der Zufallsvektor (X, Y) normalverteilt ist, dann sind X und Y unabhängig genau dann, wenn $\mathbb{E}(XY) = \mathbb{E}X \cdot \mathbb{E}Y$.

Folgerung 4.5

1. Falls X absolut stetig verteilt mit Dichte f_X ist, dann gilt

$$\mathbb{E}g(X) = \int_{\mathbb{R}^n} g(x) f_X(x) dx .$$

2. Falls X diskret verteilt mit dem Wertebereich $C = \{x_1, x_2, \dots\} \subset \mathbb{R}^n$ ist, dann gilt

$$\mathbb{E}g(X) = \sum_i g(x_i)P(X = x_i) = \sum_{x \in C} g(x)P(X = x).$$

Beispiele für die Berechnung des Erwartungswertes:

1. Poisson-Verteilung: Sei $X \sim \text{Poisson}(\lambda)$, $\lambda > 0$. Dann gilt

$$\begin{aligned} \mathbb{E}X &= \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} ke^{-\lambda} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1} \cdot \lambda}{(k-1)!} \\ &\stackrel{k-1=n}{=} e^{-\lambda} \cdot \lambda \cdot \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda} \lambda e^{\lambda} = \lambda \implies \mathbb{E}X = \lambda. \end{aligned}$$

2. Normalverteilung: Sei $X \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Zeigen wir, dass $\mathbb{E}X = \mu$.

$$\begin{aligned} \mathbb{E}X &= \int_{\mathbb{R}} xf_X(x) dx = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \int_{-\infty}^{\infty} xe^{-\left(\frac{x-\mu}{\sigma}\right)^2 \cdot \frac{1}{2}} dx \\ &\stackrel{y=\frac{x-\mu}{\sigma}}{=} \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} (\sigma y + \mu) e^{-\frac{y^2}{2}} dy \\ &= \frac{\sigma}{\sqrt{2\pi}} \underbrace{\int_{\mathbb{R}} ye^{-\frac{y^2}{2}} dy}_{=0} + \frac{\mu}{\sqrt{2\pi}} \underbrace{\int_{\mathbb{R}} e^{-\frac{y^2}{2}} dy}_{=\sqrt{2\pi}} = \mu, \end{aligned}$$

weil

$$\int_{\mathbb{R}} ye^{-\frac{y^2}{2}} dy \stackrel{t=\frac{y^2}{2}}{=} \left(\int_{-\infty}^0 + \int_0^{+\infty} \right) e^{-t} dt = - \left(\int_0^{+\infty} - \int_0^{+\infty} e^{-t} \right) = 0;$$

$$\begin{aligned} \left(\int_{\mathbb{R}} e^{-\frac{y^2}{2}} dy \right)^2 &= \int_{\mathbb{R}} e^{-\frac{x^2}{2}} dx \cdot \int_{\mathbb{R}} e^{-\frac{y^2}{2}} dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}} dx dy \\ &\stackrel{(x,y) \mapsto \text{Polarkoordinat. } (r,\varphi)}{=} \int_0^{2\pi} \int_0^{+\infty} e^{-\frac{r^2}{2}} r dr d\varphi \\ &= 2\pi \cdot \int_0^{+\infty} e^{-\frac{r^2}{2}} d\left(\frac{r^2}{2}\right) \\ &\stackrel{\frac{r^2}{2}=t}{=} 2\pi(-1) \int_0^{+\infty} d(e^{-t}) = 2\pi \end{aligned}$$

$$\implies \int_{\mathbb{R}} e^{-\frac{y^2}{2}} dy = \sqrt{2\pi} \implies E(X) = \mu.$$

4.2 Varianz

Neben dem ‘‘Mittelwert’’ einer Zufallsvariablen, den der Erwartungswert repräsentiert, gibt es weitere Charakteristiken, die für die praktische Beschreibung der zufälligen Vorgänge in der Natur und Technik sehr wichtig sind. Die Varianz z.B. beschreibt die Streuung der Zufallsvariablen um ihren Mittelwert. Sie wird als mittlere quadratische Abweichung vom Erwartungswert eingeführt:

Definition 4.6 Sei X eine Zufallsvariable mit $E(X^2) < \infty$.

1. Die *Varianz* der Zufallsvariablen X wird als $\text{Var } X = E(X - EX)^2$ definiert.
2. $\sqrt{\text{Var } X}$ heißt *Standardabweichung* von X .
3. Seien X und Y zwei Zufallsvariablen mit $E|XY| < \infty$. Die Größe $\text{Cov}(X, Y) = E(X - EX)(Y - EY)$ heißt *Kovarianz* der Zufallsvariablen X und Y .
4. Falls $\text{Cov}(X, Y) = 0$, dann heißen die Zufallsvariablen X und Y *unkorreliert*.

Satz 4.7 (*Eigenschaften der Varianz und der Kovarianz*)

Seien X, Y zwei Zufallsvariablen mit $E(X^2) < \infty, E(Y^2) < \infty$. Dann gelten folgende Eigenschaften:

$$1. \text{Cov}(X, Y) = E(XY) - EX \cdot EY, \quad \text{Var } X = E(X^2) - (EX)^2. \quad (4.1)$$

$$2. \text{Cov}(aX + b, cY + d) = ac \cdot \text{Cov}(X, Y), \quad \forall a, b, c, d \in \mathbb{R}, \quad (4.2)$$

$$3. \text{Var}(aX + b) = a^2 \text{Var}(X), \quad \forall a, b \in \mathbb{R}. \quad (4.3)$$

3. $\text{Var } X \geq 0$. Es gilt $\text{Var } X = 0$ genau dann, wenn $X = EX$ fast sicher.
4. $\text{Var}(X + Y) = \text{Var } X + \text{Var } Y + 2\text{Cov}(X, Y)$.
5. Falls X und Y unabhängig sind, dann sind sie unkorreliert, also gilt $\text{Cov}(X, Y) = 0$.

Beweis 1. Beweisen wir die Formel $\text{Cov}(X, Y) = E(XY) - EX \cdot EY$.

Die Darstellung (4.1) für die Varianz ergibt sich aus dieser Formel für $X = Y$.

$$\begin{aligned} \text{Cov}(X, Y) &= E(X - EX)(Y - EY) \\ &= E(XY - XEY - YEY + EX \cdot EY) \\ &= E(XY) - EX \cdot EY - EY \cdot EX + EX \cdot EY \\ &= E(XY) - EX \cdot EY. \end{aligned}$$

2. Beweisen wir die Formel (4.2). Die Formel (4.3) folgt aus (4.2) für $X = Y$, $a = c$ und $b = d$. Es gilt

$$\begin{aligned}\text{Cov}(aX + b, cY + d) &= \mathbb{E}((aX + b - aEX - b)(cY + d - cEY - d)) \\ &= \mathbb{E}(ac(X - EX)(Y - EY)) \\ &= ac \text{Cov}(X, Y), \quad \forall a, b, c, d \in \mathbb{R}.\end{aligned}$$

3. Es gilt offensichtlich

$$\text{Var } X = \mathbb{E}(X - EX)^2 \geq 0, \text{ da } (X - EX)^2 \geq 0 \quad \forall \omega \in \Omega.$$

Falls $X = EX$ fast sicher, dann gilt $(X - EX)^2 = 0$ fast sicher und somit $\mathbb{E}(X - EX)^2 = 0 \implies \text{Var } X = 0$.

Falls umgekehrt $\text{Var } X = 0$, dann bedeutet es $\mathbb{E}(X - EX)^2 = 0$ für $(X - EX)^2 \geq 0$. Damit folgt nach Satz 4.3, 8) $(X - EX)^2 = 0$ fast sicher $\implies X = EX$ fast sicher.

$$\begin{aligned}4. \text{ Var } (X + Y) &= \mathbb{E}(X + Y)^2 - (\mathbb{E}(X + Y))^2 \\ &= \mathbb{E}(X^2 + 2XY + Y^2) - (EX + EY)^2 \\ &= \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}Y^2 - (EX)^2 - 2EX \cdot EY - (EY)^2 \\ &= \underbrace{\mathbb{E}(X^2) - (EX)^2}_{\text{Var } X} + \underbrace{\mathbb{E}(Y^2) - (EY)^2}_{\text{Var } Y} + \underbrace{2(\mathbb{E}(XY) - EX \cdot EY)}_{\text{Cov } (X, Y)} \\ &= \text{Var } X + \text{Var } Y + 2\text{Cov}(X, Y).\end{aligned}$$

5. Falls X und Y unabhängig sind, dann gilt nach dem Satz 4.3, 7)
 $\mathbb{E}(XY) = EX \cdot EY$ und somit $\text{Cov}(X, Y) = \mathbb{E}(XY) - EX \cdot EY = 0$.

□

Folgerung 4.8

1. Es gilt $\text{Var } a = 0 \quad \forall a \in \mathbb{R}$.
2. Falls $\text{Var } X = 0$, dann ist $X = \text{const}$ fast sicher.
3. Für Zufallsvariablen X_1, \dots, X_n mit $\mathbb{E}X_i^2 < \infty$, $i = 1, \dots, n$ gilt

$$\text{Var} \left(\sum_{i=1}^n X_i \right) = \sum_{i=1}^n \text{Var } X_i + 2 \sum_{i < j} \text{Cov}(X_i, X_j).$$

4. Falls X_1, \dots, X_n paarweise unkorreliert sind, dann gilt

$$\text{Var} \left(\sum_{i=1}^n X_i \right) = \sum_{i=1}^n \text{Var } X_i.$$

Dies gilt insbesondere dann, wenn die Zufallsvariablen X_1, \dots, X_n paarweise unabhängig sind.

Beispiel 4.9

1. Poisson-Verteilung:

Sei $X \sim \text{Poisson}(\lambda)$, $\lambda > 0$. Zeigen wir, dass $\text{Var } X = \lambda$.
Es ist uns bereits bekannt, dass $\text{E}X = \lambda$. Somit gilt

$$\begin{aligned}
\text{Var } X &= \text{E}(X^2) - \lambda^2 = \sum_{k=0}^{\infty} k^2 \underbrace{e^{-\lambda} \frac{\lambda^k}{k!}}_{=P(X=k)} - \lambda^2 \\
&= e^{-\lambda} \sum_{k=1}^{\infty} (k(k-1) + k) \frac{\lambda^k}{k!} - \lambda^2 \\
&= e^{-\lambda} \sum_{k=1}^{\infty} k(k-1) \frac{\lambda^k}{k!} + \underbrace{\sum_{k=1}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!}}_{=\text{E}X=\lambda} - \lambda^2 \\
&= e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2} \cdot \lambda^2}{(k-2)!} + \lambda - \lambda^2 \\
&\stackrel{m=k-2}{=} \lambda^2 \underbrace{\sum_{m=0}^{\infty} e^{-\lambda} \cdot \frac{\lambda^m}{m!}}_{=1} + \lambda - \lambda^2 = \lambda.
\end{aligned}$$

$$\boxed{\Rightarrow \text{Var } X = \lambda.}$$

2. Normalverteilung:

Sei $X \sim N(\mu, \sigma^2)$. Zeigen wir, dass $\text{Var } X = \sigma^2$. Wie wir wissen, gilt $\text{E}X = \mu$ und somit

$$\begin{aligned}
\text{Var } X &= \text{E}(X - \mu)^2 = \int_{\mathbb{R}} (x - \mu)^2 \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}_{=f_X(x)} dx \\
&\stackrel{y=\frac{x-\mu}{\sigma}}{=} \frac{1}{\sqrt{2\pi}\sigma^2} \int_{\mathbb{R}} y^2 e^{-\frac{y^2}{2}} dy \\
&= \frac{1}{\sqrt{2\pi}\sigma^2} \int_{-\infty}^{\infty} y e^{-\frac{y^2}{2}} d\left(\frac{y^2}{2}\right) = \frac{-1}{\sqrt{2\pi}\sigma^2} \int_{-\infty}^{\infty} y d\left(e^{-\frac{y^2}{2}}\right) \\
&= \frac{1}{\sqrt{2\pi}\sigma^2} \left(-ye^{-\frac{y^2}{2}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma^2} (-0 + \sqrt{2\pi}) = \sigma^2.
\end{aligned}$$

4.3 Kovarianz und Korrelationskoeffizient

Definition 4.10 Seien X und Y zwei Zufallsvariablen mit $0 < \text{Var } X, \text{Var } Y < \infty$. Die Größe

$$\varrho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var } X} \sqrt{\text{Var } Y}}$$

heißt *Korrelationskoeffizient* von X und Y .

$\varrho(X, Y)$ ist ein Maß für die lineare Abhängigkeit der Zufallsvariablen X und Y .

Satz 4.11 (*Eigenschaften des Korrelationskoeffizienten*):

Seien X und Y zwei Zufallsvariablen mit $0 < \text{Var } X, \text{Var } Y < \infty$. Dann gilt

1. $|\varrho(X, Y)| \leq 1$.
2. $\varrho(X, Y) = 0$ genau dann, wenn X und Y unkorreliert sind. Eine hinreichende Bedingung dafür ist die Unabhängigkeit von X und Y .
3. $|\varrho(X, Y)| = 1$ genau dann, wenn X und Y fast sicher linear abhängig sind, d.h., $\exists a \neq 0, b \in \mathbb{R} : P(Y = aX + b) = 1$.

Beweis Setzen wir

$$\bar{X} = \frac{X - \text{E}X}{\sqrt{\text{Var } X}}, \quad \bar{Y} = \frac{Y - \text{E}Y}{\sqrt{\text{Var } Y}}.$$

Diese Konstruktion führt zu den sogenannten *standardisierten Zufallsvariablen* \bar{X} und \bar{Y} , für die

$$\begin{aligned} \text{E}\bar{X} &= 0, \quad \text{Var } \bar{X} = 1, \quad \text{Cov}(\bar{X}, \bar{Y}) = \text{E}(\bar{X}\bar{Y}) = \varrho(X, Y) \\ \text{E}\bar{Y} &= 0, \quad \text{Var } \bar{Y} = 1. \end{aligned}$$

1. Es gilt

$$\begin{aligned} 0 &\leq \text{Var}(\bar{X} \pm \bar{Y}) = \text{E}(\bar{X} \pm \bar{Y})^2 = \underbrace{\text{E}(\bar{X})^2}_{\text{Var } \bar{X}=1} \pm 2\text{E}(\bar{X} \cdot \bar{Y}) + \underbrace{\text{E}(\bar{Y})^2}_{\text{Var } \bar{Y}=1} \\ &= 2 \pm 2\varrho(X, Y) \implies 1 \pm \varrho(X, Y) \geq 0 \implies |\varrho(X, Y)| \leq 1. \end{aligned}$$

2. Folgt aus der Definition 4.10 und dem Satz 4.7, 5).

3. “ \Leftarrow ” Falls $Y = aX + b$ fast sicher, $a \neq 0, b \in \mathbb{R}$, dann gilt Folgendes:
Bezeichnen wir $\text{E}X = \mu$ und $\text{Var } X = \sigma^2$. Dann ist $\text{E}Y = a\mu + b$, $\text{Var } Y = a^2 \cdot \sigma^2$ und somit

$$\begin{aligned} \varrho(X, Y) &= \text{E}(\bar{X}\bar{Y}) = \text{E}\left(\frac{X - \mu}{\sigma} \cdot \frac{aX + b - a\mu - b}{|a| \cdot \sigma}\right) \\ &= \text{E}\left(\underbrace{\left(\frac{X - \mu}{\sigma}\right)^2}_{\bar{X}^2} \cdot \text{sgn } a\right) = \text{sgn } a \cdot \underbrace{\text{E}(\bar{X}^2)}_{\text{Var } \bar{X}=1} = \text{sgn } a = \pm 1. \end{aligned}$$

“ \Rightarrow ” Sei $|\varrho(X, Y)| = 1$. O.B.d.A. betrachten wir den Fall $\varrho(X, Y) = 1$. Aus 1) gilt $\text{Var}(\bar{X} - \bar{Y}) = 2 - 2\varrho(X, Y) = 0 \implies \bar{X} - \bar{Y} = \text{const}$ fast sicher aus dem Satz 4.7, 3). Somit sind X und Y linear abhängig.

Für den Fall $\varrho(X, Y) = -1$ betrachten wir analog

$$\text{Var}(\bar{X} + \bar{Y}) = 2 + 2\varrho(X, Y) = 0.$$

□

4.4 Höhere und gemischte Momente

Außer des Erwartungswertes, der Varianz und der Kovarianz gibt es eine Reihe von weiteren Charakteristiken von Zufallsvariablen, die für uns von Interesse sind.

Definition 4.12 Seien X, X_1, \dots, X_n Zufallsvariablen auf dem Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) .

1. Der Ausdruck $\mu_k = E(X^k)$, $k \in \mathbb{N}$ heißt *k-tes Moment* der Zufallsvariablen X .
2. $\bar{\mu}_k = E(X - EX)^k$, $k \in \mathbb{N}$ heißt *k-tes zentriertes Moment* der Zufallsvariablen X .
3. $E(X_1^{k_1} \cdot \dots \cdot X_n^{k_n})$, $k_1, \dots, k_n \in \mathbb{N}$ heißt *gemischtes Moment* der Zufallsvariablen X_1, \dots, X_n der Ordnung $k = k_1 + \dots + k_n$.
4. $E[(X_1 - EX_1)^{k_1} \cdot \dots \cdot (X_n - EX_n)^{k_n}]$ heißt *zentriertes gemischtes Moment* der Zufallsvariablen X_1, \dots, X_n der Ordnung $k = k_1 + \dots + k_n$.

Anmerkung:

- (a) Die angegebenen Momente müssen nicht unbedingt existieren, beispielsweise existiert EX^k , $k \in \mathbb{N}$ für $X \sim \text{Cauchy}(0, 1)$ nicht.
- (b) Offensichtlich ist $\text{Var } X$ das zweite zentrierte Moment von X , genauso wie $\text{Cov}(X, Y)$ das zweite zentrierte gemischte Moment von X und Y ist. Dabei haben Momente dritter und vierter Ordnung eine besondere Bedeutung:

Definition 4.13 1. Der Quotient

$$\gamma_1 = \text{Sch}(X) = \frac{\bar{\mu}_3}{\sqrt{(\bar{\mu}_2)^3}} = \frac{E(X - EX)^3}{\sqrt{(\text{Var } X)^3}} = E(\bar{X}^3)$$

heißt *Schiefe* oder *Symmetriekoeffizient* der Verteilung von X . Falls $\gamma_1 > 0$, dann ist die Verteilung von X rechtsschief bzw. linkssteil (für

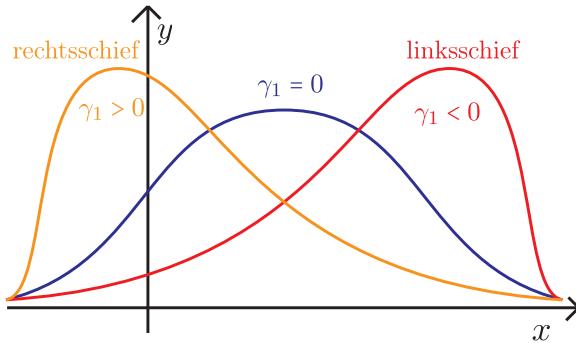


Abbildung 4.1: Veranschaulichung der Schiefe einer Verteilung an Hand der Grafik ihrer Dichte

$\gamma_1 < 0$ linksschief bzw. rechtssteil) vgl. hierzu Abbildung 4.1. Es ist ein Maß für die Symmetrie der Verteilung.

2. Der Ausdruck

$$\gamma_2 = \frac{\bar{\mu}_4}{\bar{\mu}_2^2} - 3 = \frac{E(X - EX)^4}{(Var X)^2} - 3 = E(\bar{X}^4) - 3$$

heißt *Wölbung (Exzess)* der Verteilung von X . Es ist ein Maß für die “Spitzigkeit” der Verteilung:

- $\gamma_2 > 0$ – Verteilung steilgipflig
- $\gamma_2 < 0$ – Verteilung flachgipflig, vgl. Abb. 4.2.

Die beiden Kerngrößen messen Abweichungen der Verteilung von X

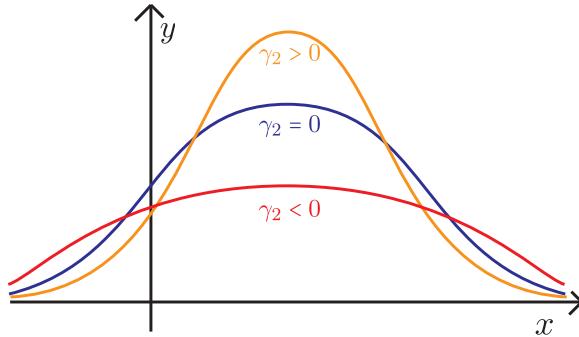


Abbildung 4.2: Veranschaulichung der Wölbung einer Verteilung an Hand der Grafik ihrer Dichte

von einer Gaußschen $N(\mu, \sigma^2)$ -Verteilung, für die $\gamma_1 = \gamma_2 = 0$.

Übungsaufgabe 4.14 Beweisen Sie, dass $\gamma_1 = \gamma_2 = 0$ für $X \sim N(\mu, \sigma^2)$.

4.5 Entropie

In der Physik wird die *Entropie* als ein logarithmisches Maß für das Chaos bzw. Ordnung, Diversität oder Vielfalt der energetischen Ebenen eines thermodynamischen Systems verstanden. In der Wahrscheinlichkeitstheorie ist es schlicht eine weitere Erwartung–basierte Charakteristik einer Zufallsvariablen X , die die Größe des Wertebereichs C von X bzw. ihre Konzentration innerhalb von C wiedergibt.

Nachdem C. Shannon (1943-1948) den von L. Boltzmann in den 1870er Jahren eingeführten statistischen Begriff der Entropie in der Informationstheorie anwendete, wurde dieser in den Arbeiten [135, 136] von A. Rényi weiter verallgemeinert:

Definition 4.15 Sei die Zufallsvariable $X : \Omega \rightarrow \mathbb{R}$ diskret verteilt mit Wertebereich C .

1. Die *Shannon*¹–*Entropie* von X wird eingeführt als

$$H(X) := - \sum_{x \in C} \Pr(X = x) \log \Pr(X = x).$$

2. Die *Rényi*²–*Entropie* der Ordnung $\alpha > 0$, $\alpha \neq 1$ von X wird definiert durch

$$H_\alpha(X) := (1 - \alpha)^{-1} \log \left(\sum_{x \in C} (\Pr(X = x))^\alpha \right).$$

Definition 4.16 Die Zufallsvariable $X : \Omega \rightarrow \mathbb{R}$ sei absolut stetig verteilt mit Dichte f_X .

1. Die *differentielle Entropie* von X nennt man die Größe

$$h(X) := - \int_{\mathbb{R}} f_X(y) \log f_X(y) dy.$$

2. Die *differentielle Rényi–Entropie* der Ordnung $\alpha > 0$, $\alpha \neq 1$ von X ist entsprechend gleich

$$h_\alpha(X) := (1 - \alpha)^{-1} \log \int_{\mathbb{R}} f_X^\alpha(y) dy.$$

Auf Grund des folgenden Grenzwertverhaltens für $\alpha \rightarrow 1$ werden wir die Bezeichnungen $H_1 = H$, $h_1 = h$ verwenden:

¹Benannt nach dem amerikanischen Mathematiker Claude E. Shannon (1916-2001), dem Vater der stochastischen Informationstheorie.

²Benannt nach ihrem Urheber ungarischen Mathematiker Alfréd Rényi (1921-1970).

Lemma 4.17 (Eigenschaften der Entropie) Sei X eine Zufallsvariable mit einer diskreten bzw. absolut stetigen Verteilung. Für beliebiges $\alpha > 0$ gelten folgende Eigenschaften:

1. Asymptotik bzgl. α : $H(X) = \lim_{\alpha \rightarrow 1} H_\alpha(X)$ bzw. $h(X) = \lim_{\alpha \rightarrow 1} h_\alpha(X)$.
2. Verschiebungsinvarianz: $H_\alpha(X+b) = H_\alpha(X)$ bzw. $h_\alpha(X+b) = h_\alpha(X)$ für beliebiges $b \in \mathbb{R}$.
3. Skalierung: $H_\alpha(aX) = H_\alpha(X)$, $h_\alpha(aX) = h_\alpha(X) + \log |a|$ für beliebiges $a \neq 0$.

Beweis 1. Beide Aussagen folgen unter Verwendung von der L'Hospital-Regel, indem man den Zähler in H_α , h_α und den Nenner $1 - \alpha$ bzgl. α differenziert und dann $\alpha \rightarrow 1$ streben lässt.

2. Für diskrete Zufallsvariablen X bewirkt die Addition einer Konstanten b lediglich die Verschiebung ihres Wertebereichs ($C+b$), was durch die Substitution $x \mapsto x-b$ in der Definition 4.15 behoben wird. Für absolut stetig verteilte Zufallsvariablen X folgt die Aussage direkt aus Korollar ??.
3. Beweisen wir die Aussagen für $\alpha \neq 1$. Der Fall $\alpha = 1$ folgt daraus mit Hilfe des Grenzwertübergangs für $\alpha \rightarrow 1$, siehe Punkt 1. Die Behauptung für H_α wird analog wie im Punkt 2 bewiesen mit Substitution $x \mapsto x/a$ in der Summe. Nach dem Korollar ?? gilt für h_α , dass

$$\begin{aligned} h_\alpha(aX) &= (1 - \alpha)^{-1} \log \int_{\mathbb{R}} |a|^{-\alpha} f_X^\alpha(y/a) dy = (1 - \alpha)^{-1} \log |a|^{1-\alpha} \\ &\quad + (1 - \alpha)^{-1} \log \int_{\mathbb{R}} f_X^\alpha(y/a) d(y/a) = \log |a| + h_\alpha(X). \end{aligned}$$

□

Bemerkung 4.18 1. Es gilt

$$\begin{aligned} H(X) &= -\mathbb{E} \log p_X(X), \quad H_\alpha(X) = (1 - \alpha)^{-1} \log \mathbb{E} p_X^{\alpha-1}(X), \\ h(X) &= -\mathbb{E} \log f_X(X), \quad h_\alpha(X) = (1 - \alpha)^{-1} \log \mathbb{E} f_X^{\alpha-1}(X), \end{aligned}$$

wobei p_X bzw. f_X die (Zähl)Dichte der diskret bzw. absolut stetig verteilten Zufallsvariablen X ist. Bei ersten dieser Gleichungen wird in der Informationstheorie als die erwartete Information interpretiert, gewonnen aus der Beobachtung von X .

2. Die Entropien H_α bzw. h_α , $\alpha > 0$, können auf ähnlichem Wege auch für Zufallsvektoren (H_α sogar für beliebige Zufallselemente) eingeführt werden, da ihre Definition lediglich von den (Zähl)Dichten Gebrauch macht.

3. Sei $\alpha > 0$ beliebig. Obwohl $H_\alpha(X) \geq 0$ für alle diskret verteilten Zufallsvariablen X , kann $h_\alpha(X)$ auch negative Werte annehmen, vgl. Beispiel 4.19.
4. In der Informationstheorie ist es üblich, den natürlichen Logarithmus (zur Basis e) in den Definitionen 4.15, 4.16 durch den \log_2 zu ersetzen, was mit der binären Kodierung bei der Informationsübertragung zusammenhängt.

Beispiel 4.19 Die Entropie H_α bzw. h_α von zwei wichtigen Verteilungen sei hier gegeben, wobei $\alpha > 0$ beliebig ist:

1. Gleichverteilung: Für $X \sim \mathcal{U}\{x_1, \dots, x_n\}$, $n \in \mathbb{N}$, und $Y \sim \mathcal{U}[0, M]$, $M > 0$, gilt $H_\alpha(X) = \log n$, $h_\alpha(Y) = \log M$. Dabei ist die Entropie gleich Null für eine konstante X und tendiert gegen $+\infty$ mit unbegrenzt wachsender Anzahl n der Zustände des Systems. Für $M \rightarrow +0$ gilt dagegen $h_\alpha(Y) \rightarrow -\infty$.
2. Normalverteilung: Für $X \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$, gilt

$$h_\alpha(X) = \log \sigma + \frac{1}{2} \log 2\pi + \begin{cases} \frac{1}{2}, & \alpha = 1, \\ \frac{\log \alpha}{2(\alpha-1)}, & \alpha > 0, \alpha \neq 1. \end{cases}$$

Wir schlussfolgern $\lim_{\sigma \rightarrow +0} h_\alpha(X) = -\infty$, $\lim_{\sigma \rightarrow +\infty} h_\alpha(X) = +\infty$.

Übungsaufgabe 4.20 Zeigen Sie, dass die Entropie einer Zufallsvariablen $X \sim Exp(\lambda)$, $\lambda > 0$, durch $h_\alpha(X) = \frac{\lambda^{\alpha-1}}{(1-\alpha)\alpha}$, $\alpha > 0$, $\alpha \neq 1$, sowie $h(X) = 1 - \log \lambda$ gegeben ist.

Die statistische Schätzung der Entropie wird in den Arbeiten [96, 17, 155, 31, 118, 105, 104, 125] ausführlich behandelt.

4.6 Ungleichungen

Satz 4.21 (Ungleichung von Markow):

Sei X eine Zufallsvariable mit $E|X|^r < \infty$ für ein $r \geq 1$. Dann gilt

$$P(|X| \geq \varepsilon) \leq \frac{E|X|^r}{\varepsilon^r} \quad \forall \varepsilon > 0.$$

Beweis Es gilt

$$\begin{aligned} E|X|^r &= \underbrace{E(|X|^r \cdot I(|X| \leq \varepsilon))}_{\geq 0} + E(|X|^r \cdot I(|X| > \varepsilon)) \\ &\geq E(\varepsilon^r \cdot I(|X| > \varepsilon)) = \varepsilon^r \cdot P(|X| > \varepsilon), \end{aligned}$$

daher $P(|X| > \varepsilon) \leq \frac{E|X|^r}{\varepsilon^r}$. \square

Folgerung 4.22 (*Ungleichung von Tschebyschew*).

1. Sei X eine Zufallsvariable mit $\text{E}X^2 < \infty$ und $\varepsilon > 0$. Dann gilt

$$P(|X - \text{E}X| \geq \varepsilon) \leq \frac{\text{Var } X}{\varepsilon^2}.$$

2. Sei $\text{E}e^{sX} < \infty$ für ein $s > 0$. Dann gilt

$$P(X \geq \varepsilon) \leq \frac{\text{E}e^{\lambda X}}{e^{\lambda \varepsilon}}, \forall \varepsilon > 0 \quad \forall 0 \leq \lambda \leq s.$$

Beweis Benutze die Markow–Ungleichung für die Zufallsvariable

1. $Y = X - \text{E}X$ und $r = 2$ und
2. $Y = e^{\lambda X} \geq 0$, $\varepsilon = e^{\lambda \varepsilon_0}$, $r = 1$.

□

Beispiel 4.23 Der Durchmesser der Mondscheibe wird aus den Bildern der Astrophotographie folgendermaßen bestimmt: bei jeder Aufnahme der Mondscheibe wird ihr Durchmesser X_i gemessen. Nach n Messungen wird der Durchmesser als $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ aus allen Beobachtungen geschätzt. Sei $\mu = \text{E}X_i$ der wahre (unbekannte) Wert des Monddurchmessers. Wie viele Beobachtungen n müssen durchgeführt werden, damit die Schätzung \bar{X}_n weniger als um 0,1 vom Wert μ mit Wahrscheinlichkeit von mindestens 0,99 abweicht? Mit anderen Worten, finde n : $P(|\bar{X}_n - \mu| \leq 0,1) \geq 0,99$. Diese Bedingung ist äquivalent zu $P(|\bar{X}_n - \mu| > 0,1) \leq 1 - 0,99 = 0,01$. Sei $\text{Var } X_i = \sigma^2 > 0$. Dann gilt

$$\text{Var } \bar{X}_n = \frac{1}{n^2} \cdot \sum_{i=1}^n \text{Var } X_i = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n},$$

wobei vorausgesetzt wird, dass alle Messungen X_i unabhängig voneinander durchgeführt werden. Somit gilt nach der Ungleichung von Tschebyschew

$$P(|\bar{X}_n - \mu| > 0,1) \leq \frac{\text{Var } \bar{X}_n}{0,1^2} = \frac{\sigma^2}{n \cdot 0,01},$$

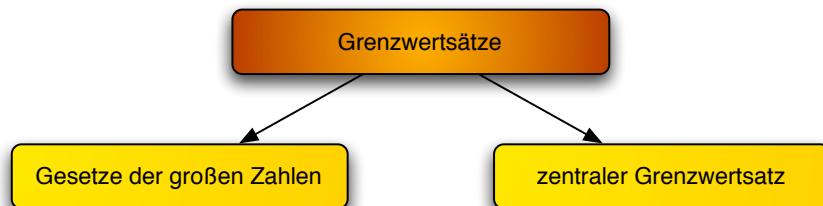
woraus folgt, dass

$$n \geq \frac{\sigma^2}{(0,01)^2} = 10^4 \cdot \sigma^2.$$

Für $\sigma = 1$ braucht man z.B. mindestens 10000 Messungen! Diese Zahl zeigt, wie ungenau die Schranke in der Ungleichung von Tschebyschew ist. Eine viel genauere Antwort ($n \geq 670$) kann man mit Hilfe des zentralen Grenzwertsatzes bekommen. Dies wird allerdings erst im Kapitel 5 behandelt.

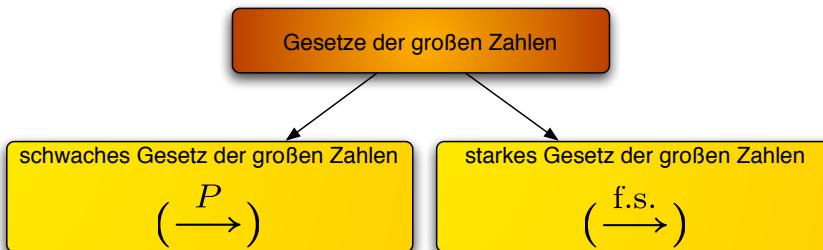
Kapitel 5

Grenzwertsätze



In diesem Kapitel betrachten wir Aussagen der Wahrscheinlichkeitsrechnung, die Näherungsformeln von großer anwendungsbezogener Bedeutung liefern. Dies wird an mehreren Beispielen erläutert.

5.1 Gesetze der großen Zahlen



Ein typisches Gesetz der großen Zahlen besitzt die Form

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{n \rightarrow \infty} EX_0, \quad (5.1)$$

wobei $\{X_n\}_{n \in \mathbb{N}}$ unabhängige identisch verteilte Zufallsvariablen mit $X_n \xrightarrow{d} X_0$, $E|X_0| < \infty$ sind.

Die Konvergenz in (5.1) wird entweder in Wahrscheinlichkeit oder fast sicher verstanden.

Definition 5.1 Sei $\{X_n\}_{n \in \mathbb{N}}$ eine Folge von Zufallsvariablen definiert auf dem Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) . Sei X eine weitere Zufallsvariable auf (Ω, \mathcal{F}, P) . Man sagt, die Folge $\{X_n\}$ konvergiert gegen X für $n \rightarrow \infty$

1. *fast sicher oder mit Wahrscheinlichkeit 1* ($X_n \xrightarrow[n \rightarrow \infty]{\text{f.s.}} X$), falls

$$P(\{\omega \in \Omega : X_n(\omega) \xrightarrow[n \rightarrow \infty]{} X(\omega)\}) = 1.$$

2. *in Wahrscheinlichkeit oder stochastisch* ($X_n \xrightarrow[n \rightarrow \infty]{P} X$), falls

$$\forall \varepsilon > 0 \quad P(|X_n - X| > \varepsilon) \xrightarrow[n \rightarrow \infty]{} 0.$$

Wenn \xrightarrow{P} gemeint ist, spricht man von dem *schwachen Gesetz der großen Zahlen*. Falls $\xrightarrow{\text{f.s.}}$ gemeint ist, heißt die Aussage (5.1) *starkes Gesetz der großen Zahlen*.

Im Folgenden verwenden wir die Bezeichnungen $S_n = \sum_{i=1}^n X_i$, $\bar{X}_n = \frac{S_n}{n}$, $\forall n \in \mathbb{N}$ für eine Folge $\{X_n\}_{n \in \mathbb{N}}$ von Zufallsvariablen auf dem Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) .

5.1.1 Schwaches und Starkes Gesetz der großen Zahlen

Satz 5.2 (Markow)

Sei $\{X_n\}_{n \in \mathbb{N}}$ eine Folge von Zufallsvariablen auf dem Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) mit $EX_i^2 < \infty \forall i$. Falls

$$\text{Var } \bar{X}_n \xrightarrow[n \rightarrow \infty]{} 0, \tag{5.2}$$

dann gilt

$$\bar{X}_n - \frac{1}{n} \sum_{i=1}^n EX_i \xrightarrow[n \rightarrow \infty]{P} 0.$$

Folgerung 5.3 Seien die Zufallsvariablen $\{X_n\}_{n \in \mathbb{N}}$ im Satz 5.2 unabhängig. Dann gilt Folgendes:

1. Die Bedingung $\text{Var } \bar{X}_n \xrightarrow[n \rightarrow \infty]{} 0$ bekommt die Form

$$\frac{1}{n^2} \sum_{i=1}^n \text{Var } X_i \xrightarrow[n \rightarrow \infty]{} 0.$$

2. Falls $\text{Var } X_n \leq c = \text{const} \quad \forall n \in \mathbb{N}$, dann gilt die Bedingung (5.2) und somit die Aussage des Satzes 5.2 (Satz von Tschebyschew).
3. Insbesondere ist die Bedingung $\text{Var } X_n \leq c = \text{const} \quad \forall n \in \mathbb{N}$ erfüllt, falls $\{X_n\}_{n \in \mathbb{N}}$ unabhängige identisch verteilte Zufallsvariablen mit

$$\mathbb{E}X_n = \mu, \text{Var } X_n = \sigma^2 < \infty$$

sind. Dann nimmt das schwache Gesetz der großen Zahlen die klassische Form

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{P} \mu$$

an.

Die Existenz der zweiten Momente ist für das schwache Gesetz der großen Zahlen nicht entscheidend. So kann man mit Hilfe der charakteristischen Funktionen folgenden Satz beweisen:

Satz 5.4 (*Schwaches Gesetz der großen Zahlen von Kchintschin*)

Sei $\{X_n\}_{n \in \mathbb{N}}$ eine Folge von stochastisch unabhängigen, integrierbaren Zufallsvariablen, $n \in \mathbb{N}$, mit demselben Erwartungswert $\mathbb{E}X_n = \mu < \infty$. Dann gilt

$$\overline{X}_n \xrightarrow[n \rightarrow \infty]{P} \mu.$$

Satz 5.5 (*Starkes Gesetz der großen Zahlen von Kolmogorow*)

Seien $\{X_n\}_{n \in \mathbb{N}}$ unabhängige identisch verteilte Zufallsvariablen. Es gilt $\overline{X}_n \xrightarrow[n \rightarrow \infty]{\text{f.s.}} \mu$ genau dann, wenn $\exists \mathbb{E}X_n = \mu < \infty$.

5.1.2 Anwendung der Gesetze der großen Zahlen

1. *Monte–Carlo–Methoden zur numerischen Integration*

Sei $g : [0, 1]^d \rightarrow \mathbb{R}$ eine beliebige stetige Funktion. Wie kann man mit Hilfe der Gesetze der großen Zahlen

$$\int_{[0,1]^d} g(x) dx = \int_0^1 \dots \int_0^1 g(x_1, \dots, x_d) dx_1 \dots dx_d$$

numerisch berechnen?

Der Algorithmus ist wie folgt:

- Generiere eine Folge von Realisierungen von unabhängigen identisch verteilten Zufallsvariablen

$$X_1, \dots, X_n \text{ mit } X_i \sim [0, 1]^d, i = 1, \dots, n.$$

- Setze

$$\int_{[0,1]^d} g(x) dx \approx \frac{1}{n} \sum_{i=1}^n g(X_i) \quad (5.3)$$

für große n . Dieser Vorgang ist berechtigt, denn nach dem Satz 5.5 gilt

$$\frac{1}{n} \sum_{i=1}^n g(X_i) \xrightarrow{n \rightarrow \infty} \mathbb{E}g(X_1) = \int_{[0,1]^d} g(x) dx$$

und somit gilt (5.3) für ausreichend große n .

Bemerkung:

Dieselbe Methode kann durch geeignete Transformation vom Integrationsgebiet $G \subset \mathbb{R}^d$ und andere Wahl von Zufallsvariablen X_i auf die Berechnung von $\int_G g(x) dx$ erweitert werden, G kompakte Teilmenge von \mathbb{R}^d . So genügt es nur $X_i \sim U(G)$, $i = 1, \dots, n$ zu betrachten.

2. Numerische Berechnung der Zahl π :

Wie kann π mit Hilfe eines Rechners beliebig genau berechnet werden? Dazu wird das starke Gesetz der großen Zahlen wie folgt verwendet:

- Generiere Realisierungen von unabhängig und identisch verteilten Zufallvektoren $X_1, \dots, X_n \in \mathbb{R}^2$ mit $X_i \sim U[-1, 1]^2$, $i = 1, \dots, n$.
- Es gilt

$$\pi \approx \frac{4}{n} \sum_{i=1}^n I(|X_i| \leq 1) \quad (5.4)$$

für große n .

In der Tat, nach dem Satz 5.5 gilt

$$\frac{1}{n} \sum_{i=1}^n I(|X_i| \leq 1) \xrightarrow{n \rightarrow \infty} \mathbb{E}I(|X_1| \leq 1) = P(|X_1| \leq 1) = \frac{|B_1(0)|}{|[-1, 1]^2|} = \frac{\pi}{2^2} = \frac{\pi}{4}.$$

Somit ist die Verwendung der Berechnungsformel (5.4) berechtigt für große n .

5.2 Zentraler Grenzwertsatz

Für die Gesetze der großen Zahlen wurde die Normierung $\frac{1}{n}$ der Summe $S_n = \sum_{i=1}^n X_i$ gewählt, um $\frac{S_n}{n} \xrightarrow{n \rightarrow \infty} \mathbb{E}X_1$ zu beweisen. Falls jedoch eine andere Normierung gewählt wird, so sind andere Grenzwertaussagen möglich. Im Fall der Normierung $\frac{1}{\sqrt{n}}$ spricht man von zentralen Grenzwertsätzen: unter gewissen Voraussetzungen gilt also

$$\frac{S_n - n\mathbb{E}X_1}{\sqrt{n \text{Var } X_1}} \xrightarrow{n \rightarrow \infty} Y \sim N(0, 1).$$

5.2.1 Klassischer zentraler Grenzwertsatz

Satz 5.6 Sei $\{X_n\}_{n \in \mathbb{N}}$ eine Folge von unabhängigen identisch verteilten Zufallsvariablen auf dem Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) mit $EX_i = \mu$, $\text{Var } X_i = \sigma^2$, wobei $0 < \sigma^2 < \infty$. Dann gilt

$$P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq x\right) \xrightarrow{n \rightarrow \infty} \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy \quad \forall x \in \mathbb{R},$$

wobei $\Phi(x)$ die Verteilungsfunktion einer $N(0, 1)$ -verteilten Zufallsvariablen ist.

Folgerung 5.7 Unter den Voraussetzungen des Satzes 5.6 gilt zusätzlich

1.

$$P\left(\frac{S_n - n\mu}{\sqrt{n} \cdot \sigma} < x\right) \xrightarrow{n \rightarrow \infty} \Phi(x) \quad \forall x \in \mathbb{R},$$

2.

$$P\left(a \leq \frac{S_n - n\mu}{\sqrt{n} \cdot \sigma} \leq b\right) \xrightarrow{n \rightarrow \infty} \Phi(b) - \Phi(a) \quad \forall a, b \in \mathbb{R}, a \leq b.$$

Beispiel 5.8 1. Satz von de Moivre–Laplace

Falls $X_n \sim \text{Bernoulli}(p)$, $n \in \mathbb{N}$ unabhängig sind und $p \in (0, 1)$, dann genügt die Folge $\{X_n\}_{n \in \mathbb{N}}$ mit $EX_n = p$, $\text{Var } X_n = p(1 - p)$ den Voraussetzungen des Satzes 5.6. Das Ergebnis

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1)$$

wurde mit einfachen Mitteln als erster zentraler Grenzwertsatz von Abraham de Moivre (1667-1754) bewiesen und trägt daher seinen Namen. Es kann folgendermaßen interpretiert werden:

Falls die Anzahl n der Experimente groß wird, so wird die Binomialverteilung von $S_n \sim \text{Bin}(n, p)$, das die Anzahl der Erfolge in n Experimenten darstellt, approximiert durch

$$\begin{aligned} P(a \leq S_n \leq b) &= P\left(\frac{a - np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{b - np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{b - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - np}{\sqrt{np(1-p)}}\right), \end{aligned}$$

$\forall a, b \in \mathbb{R}$, $a \leq b$, wobei $p \in (0, 1)$ als die Erfolgswahrscheinlichkeit in einem Experiment interpretiert wird. So kann z.B. S_n als die Anzahl

von Kopf in n Würfen einer fairen Münze ($p = \frac{1}{2}$) betrachtet werden. Hier gilt also

$$P(a \leq S_n \leq b) \underset{n\text{-groß}}{\approx} \Phi\left(\frac{2b-n}{\sqrt{n}}\right) - \Phi\left(\frac{2a-n}{\sqrt{n}}\right), \quad a < b.$$

2. Berechnen wir die Anzahl der notwendigen Messungen des Monddurchmessers im Beispiel 4.23 mit Hilfe des zentralen Grenzwertsatzes:
finde $n \in \mathbb{N}$ mit

$$P(|\bar{X}_n - \mu| \leq 0,1) > 0,99,$$

oder äquivalent dazu

$$P(|\bar{X}_n - \mu| > 0,1) \leq 0,01,$$

wobei $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ ist. Es gilt

$$\begin{aligned} P(|\bar{X}_n - \mu| \leq 0,01) &= P\left(-0,1 \leq \frac{S_n - n\mu}{n} \leq 0,1\right) \\ &= P\left(-0,1 \frac{\sqrt{n}}{\sigma} \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq 0,1 \frac{\sqrt{n}}{\sigma}\right) \\ &\underset{n \text{ groß}}{\approx} \Phi\left(\frac{0,1\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{0,1\sqrt{n}}{\sigma}\right) = 2\Phi\left(\frac{0,1\sqrt{n}}{\sigma}\right) - 1, \end{aligned}$$

weil $N(0,1)$ eine symmetrische Verteilung ist und somit

$$\Phi(x) = 1 - \Phi(-x) \quad \forall x \in \mathbb{R}$$

gilt.

Es muss also $2\Phi\left(\frac{0,1\sqrt{n}}{\sigma}\right) - 1 > 0,99$ erfüllt sein. Dies ist äquivalent zu

$$\Phi\left(\frac{0,1\sqrt{n}}{\sigma}\right) > \frac{1,99}{2} = 0,995 \iff \frac{0,1\sqrt{n}}{\sigma} > \Phi^{-1}(0,995)$$

oder

$$\begin{aligned} n &> \frac{\sigma^2}{(0,1)^2} \left(\Phi^{-1}(0,995)^2\right) = \frac{\sigma^2 (\Phi^{-1}(0,995))^2}{0,01} \\ &= \frac{\sigma^2 (2,58)^2}{0,01} = \sigma^2 \cdot 665,64. \end{aligned}$$

Für $\sigma^2 = 1$ ergibt sich die Antwort

$$n \geq 666$$

5.2.2 Konvergenzgeschwindigkeit im zentralen Grenzwertsatz

In diesem Abschnitt möchten wir die *Schnelligkeit der Konvergenz im zentralen Grenzwertsatz* untersuchen. Damit aber diese Fragestellung überhaupt sinnvoll erscheint, muss die Konvergenz im zentralen Grenzwertsatz gleichmäßig sein:

$$\sup_{x \in \mathbb{R}} \left| P\left(\frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\sigma} \leq x\right) - \Phi(x) \right| \xrightarrow{n \rightarrow \infty} 0.$$

Das ist tatsächlich der Fall, wie aus der Stetigkeit von $\Phi(x)$ und dem folgenden Lemma 5.9 hervorgeht.

Lemma 5.9 Seien $\{F_n\}_{n=1}^\infty$, F Verteilungsfunktionen, sodass $F(x)$ stetig ist $\forall x \in \mathbb{R}$ und $F_n(x) \xrightarrow{n \rightarrow \infty} F(x) \forall x \in \mathbb{R}$. Dann ist die Konvergenz von F_n zu F gleichmäßig:

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{n \rightarrow \infty} 0.$$

Satz 5.10 (Berry–Esséen)

Sei $\{X_n\}_{n \in \mathbb{N}}$ eine Folge von unabhängigen identisch verteilten Zufallsvariablen mit $\mathbb{E}X_n = \mu$, $\text{Var } X_n = \sigma^2 > 0$, $\mathbb{E}|X_n|^3 < \infty$. Sei

$$F_n(x) = P\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \leq x\right), \quad x \in \mathbb{R}, n \in \mathbb{N}.$$

Dann gilt

$$\sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \leq \frac{c \cdot \mathbb{E}|X_1 - \mu|^3}{\sigma^3 \sqrt{n}},$$

wobei c eine Konstante ist, $\frac{1}{\sqrt{2\pi}} \leq c < 0,4785$, $\frac{1}{\sqrt{2\pi}} \approx 0,39894$.

5.2.3 Grenzwertsatz von Lindeberg

Im klassischen zentralen Grenzwertsatz wurden Folgen von unabhängigen und identisch verteilten Zufallsvariablen betrachtet. In diesem Abschnitt lassen wir die Voraussetzung der identischen Verteiltheit fallen und formulieren einen allgemeineren Grenzwertsatz in der Form von Lindeberg.

Satz 5.11 (Lindeberg)

Sei $\{X_n\}_{n \in \mathbb{N}}$ eine Folge von unabhängigen Zufallsvariablen mit $\mathbb{E}X_n = \mu_n$, $0 < \sigma_n^2 = \text{Var } X_n < \infty \forall n$. Sei $S_n = \sum_{i=1}^n X_i$, $D_n^2 = \sum_{i=1}^n \sigma_i^2$. Falls

$$\frac{1}{D_n^2} \sum_{k=1}^n \mathbb{E} \left((X_k - \mu_k)^2 \cdot I(|X_k - \mu_k| > \varepsilon D_n) \right) \xrightarrow{n \rightarrow \infty} 0, \quad (5.5)$$

dann gilt

$$\frac{S_n - \sum_{i=1}^n \mu_i}{D_n} \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1).$$

Folgerung 5.12 Sei $\{X_n\}_{n \in \mathbb{N}}$ eine Folge von unabhängigen Zufallsvariablen mit $|X_n| \leq c < \infty \forall n \in \mathbb{N}$ und $D_n \xrightarrow{n \rightarrow \infty} \infty$. Dann gilt der zentrale Grenzwertsatz in der Form des Satzes 5.11.

Beweis Wir müssen die Gültigkeit der Lindeberg–Bedingung (5.5) prüfen: aus der Ungleichung von Tschebyschew folgt

$$\begin{aligned} \mathbb{E} \left((X_k - \mu_k)^2 \cdot I(|X_k - \mu_k| > \varepsilon D_n) \right) &\underset{|X_k| \leq c, |\mathbb{E} X_k| \leq c}{\leq} (2c)^2 P(|X_k - \mu_k| > \varepsilon D_n) \\ &\leq 4c^2 \frac{\sigma_k^2}{\varepsilon^2 D_n^2} \xrightarrow{n \rightarrow \infty} 0, \end{aligned}$$

für $1 \leq k \leq n$ und somit

$$\frac{1}{D_n^2} \sum_{k=1}^n \mathbb{E} \left((X_k - \mu_k)^2 \cdot I(|X_k - \mu_k| > \varepsilon D_n) \right) \underset{\substack{=D_n^2 \\ \sum_{k=1}^n \sigma_k^2}}{\leq} 4c^2 \frac{\sum_{k=1}^n \sigma_k^2}{\varepsilon^2 D_n^4} = \frac{4c^2}{\varepsilon^2 D_n^2} \xrightarrow{n \rightarrow \infty} 0,$$

weil $D_n^2 \xrightarrow{n \rightarrow \infty} \infty$. □

Kapitel 6

Monte–Carlo–Simulation von Zufallsvariablen

Man kann ein Objekt der Wahrscheinlichkeitstheorie (z.B., ein Zufallselement) erst tief verstehen, wenn man es auf dem Rechner nachbilden (d.h., *simulieren*) kann. Der Ausdruck *Monte–Carlo–Simulation* wurde erstmals 1949 in der Arbeit [110] geprägt. Seine Urheber amerikanische Mathematiker *Nicholas Metropolis* und *Stanislaw M. Ulam*¹ wollten dabei verdeutlichen, dass Simulationsvorgänge dem Roulette–Spiel (wofür der Stadtteil Monte–Carlo im Fürstentum Monaco bekannt ist) in einer gewissen Hinsicht ähnlich sind. Die Basis solcher Simulationen bilden sogenannte *pseudozufällige Zahlen*, die als unabhängige Realisierungen einer auf $(0, 1)$ gleichverteilten Zufallsvariable interpretiert werden. Mitte des 20. Jahrhunderts wurden für die Erzeugung solcher Zahlenfolgen noch physikalische Geräte benutzt, die natürliche Fluktuationen (d.h., *Rauschen*) von z.B. elektrischer Spannung gemessen haben, und bei Überschreitung eines vorgegebenen Spannungsniveaus Eins, und sonst Null als binäre Zufallszahlen generiert haben. Diese binären Zahlenfolgen wurden als eine Binärdarstellung einer zufälligen Zahl auf dem Intervall $(0, 1)$ interpretiert. So generierte Zahlenfolgen wurden in Tabellen zufälliger Zahlen gespeichert. Seit der Entwicklung mächtiger Computer gelten solche direkten Methoden als antiquiert. Sie wurden durch rechnergestützte Generatoren von Pseudozufallszahlen ersetzt.

6.1 Pseudozufallszahlen

Es mag überraschend erscheinen, dass auf einem Rechner, der deterministisch arbeitet und so keine Zufälle zulässt, zufällige Zahlen simuliert werden

¹Die Monte–Carlo Methoden wurden während des 2. Weltkrieges von Metropolis sowie Ulam zusammen mit den amerikanischen Physikern *Enrico Fermi* und *John von Neumann* in Los Alamos, USA im Rahmen des berüchtigten *Manhattan-Projekts* entwickelt. In dem Projekt ging es um die Erschaffung der ersten Atombombe.

können. In Wirklichkeit liefert der Rechner eine periodische Folge von deterministischen Zahlen, wobei die Periode jedoch ausreichend groß ist. Somit wiederholen sich die Werte der *pseudozufälligen* Zahlen selbst bei relativ langen Simulationsstudien nicht. Man sagt, dass eine Folge von pseudozufälligen Zahlen eine Zufallsvariable X *simuliert*, wenn diese Folge annähernd dieselben statistischen Eigenschaften aufweist wie eine Stichprobe von unabhängigen Realisierungen von X . Dies kann mit Hilfe statistischer Tests auf Gleichverteiltheit wie z.B. dem Kolmogorov–Smirnov–Test, χ^2 –Test, usw. nachgewiesen werden.

Wie kann man eine $\mathcal{U}(0, 1)$ –verteilte Zufallsvariable simulieren? Auf dem Rechner wird diese absolut stetige Verteilung durch eine diskrete Gleichverteilung ersetzt, die diese ausreichend gut approximiert. Wenn wir z.B. pseudozufällige Zahlen mit einer Genauigkeit bis auf d Dezimalstellen brauchen, so kann der Wertebereich dieser diskreten Gleichverteilung als

$$\{k/10^d : k = 1, \dots, 10^d - 1\}$$

gewählt werden. Jeder Wert $k/10^d$ wird mit Wahrscheinlichkeit

$$p_k = \frac{1}{10^d - 1}$$

angenommen.

Die Klasse der Methoden, die solche Verteilungen simulieren können, heißt *Generatoren pseudozufälliger Zahlen*. Sie unterscheiden sich in ihren Eigenschaften und in ihrer Komplexität. Die meisten modernen Generatoren arbeiten iterativ, also liefern eine Zahlenfolge $x_k = G(x_{k-1})$, $k \in \mathbb{N}$, wobei $G : (0, 1) \rightarrow (0, 1)$ eine Abbildung ist und der Anfangswert $u_0 \in (0, 1)$ fixiert wird. Es ist klar, dass alle Punkte $(x_k, G(x_k)) \in [0, 1]^2$ auf der Grafik von G liegen, somit soll G gewählt werden, sodass ihre Grafik möglichst dicht das ganze Einheitsquadrat $[0, 1]^2$ füllt. Erst dann haben die Punkte $(x_k, G(x_k))$ eine Chance, pseudogleichverteilt auf $[0, 1]^2$ auszusehen. Als ein natürlicher Kandidat für eine solche Funktion gilt $G(x) = \lfloor ax \rfloor$, $x \in (0, 1)$, wobei $\lfloor ax \rfloor$ der nicht ganzzahlige Teil von ax und $a > 0$ eine sehr große Zahl sind.

Hier werden wir zwei einfache Generatoren dieser Art kennen lernen. Weitere Generatoren können den Büchern [102, 117, 59, 139, 160, 70] entnommen werden.

1. *Residuenmethode*²: Seien $a, n \in \mathbb{N}$, wobei n und $(n - 1)/2$ Primzahlen sind. Es gelte zusätzlich $a^{(n-1)/2} \equiv -1 \pmod{n}$. Definiere die Folge

$$u_k = au_{k-1} \pmod{n}, \quad k \in \mathbb{N}, \tag{6.1}$$

wobei der Anfangswert $u_0 \in \{1, \dots, n - 1\}$ der *Keim* der Folge (engl. *seed*) heißt. Somit werden $x_k = u_k/n$ als unabhängige Realisierungen

²Sie heißt auf Englisch *residual method* oder *congruential method*. Ihre Idee wurde von D.H. Lehmer (1949) vorgeschlagen.

der Zufallsvariable $X \sim \mathcal{U}(0, 1)$ interpretiert. Man kann beweisen, dass die Folge $\{u_k\}$ für beliebiges u_0 die Periode $n - 1$ hat. Somit kann x_k höchstens $n - 1$ Werte annehmen. Beispielsweise erfüllen $a = 1000$ und $n = 2001179$ die obigen Voraussetzungen. Alternativ kann man $a = 5^{17}$, $n = 2^{42}$ mit $u_0 = 1$ benutzen, selbst wenn a und n die obigen Voraussetzungen nicht erfüllen. Die Periode einer solchen Folge wird gleich 2^{40} sein.

Wie genau ist der Generator (6.1)? Sei U_0 eine auf $\{1, \dots, n-1\}$ gleichverteilte Zufallsvariable. Somit erzeugt die Relation (6.1) die Folge $U_k \equiv aU_{k-1} \pmod{n}$ von Zufallsvariablen, auf deren Basis eine neue Folge $X_k = U_k/n$ konstruiert wird, die mit $X \sim \mathcal{U}(0, 1)$ verglichen werden soll. Man kann zeigen, dass alle X_k identisch verteilt sind mit Mittelwert

$$\mathbb{E} X_k = \mathbb{E} X = 1/2$$

und Varianz

$$\text{Var } X_k = \frac{n-2}{12n} \rightarrow \frac{1}{12} = \text{Var } X,$$

falls $n \rightarrow \infty$. Die Zufallsvariablen X_k sind offensichtlich nicht unabhängig voneinander. Man kann zeigen, dass der Korrelationskoeffizient

$$\text{Corr}(X_k, X_{k+1}) = \frac{\text{Cov}(X_k, X_{k+1})}{\sqrt{\text{Var } X_k \text{Var } X_{k+1}}} \approx 1/a$$

und somit nicht Null ist.

2. Seien $a = 10^m + 1$ für $m \in \mathbb{N}$, $b \in \mathbb{N}$ nicht teilbar durch 2 oder 5, und sei $n = 10^d$ für $d \in \mathbb{N}$. Definiere die Folge

$$u_k = au_{k-1} + b \pmod{n}, \quad k \in \mathbb{N} \quad (6.2)$$

für einen Keimwert u_0 . Die Periode der Folge (6.2) ist $n - 1$. Setze $x_k = u_k/n$, $k \in \mathbb{Z}_+$.

Um bei unterschiedlichen Simulationsvorgängen verschiedene Folgen (6.1)–(6.2) zu bekommen, soll deren Keimwert u_0 so oft wie möglich geändert werden. Eine Ausnahme aus dieser Regel bilden Berechnungen, bei denen dieselbe Folge von Pseudozufallszahlen verwendet werden soll.

Im Folgenden nehmen wir an, dass wir ausreichend gute unabhängige Realisierungen der Zufallsvariablen $U \sim \mathcal{U}(0, 1)$ generieren können. Um andere Zufallsvariablen auf Basis einer Gleichverteilung simulieren zu können, gibt es eine Reihe von Methoden wie z.B. die *Inversionsmethode* oder die *Akzeptanz- und Verwerfungsmethode*.

6.2 Inversionsmethode

Lemma 6.1 Sei X eine Zufallsvariable mit Verteilungsfunktion F_X und Quantilfunktion F_X^{-1} .

1. Falls F_X stetig ist, so ist die Zufallsvariable $Y = F_X(X)$ gleichverteilt auf dem Intervall $[0, 1]$
2. Sei $U \sim \mathcal{U}[0, 1]$ eine Zufallsvariable. Dann ist die Zufallsvariable $Z = F_X^{-1}(U)$ genau so verteilt wie X .

Die Aussage 2 von Lemma 6.1 ermöglicht die Simulation von Zufallsvariablen X auf dem Rechner, falls ihre Quantilfunktion F_X^{-1} explizit berechnet werden kann. Insbesondere ist es der Fall, wenn die Verteilungsfunktion F_X strikt monoton steigend ist, und damit die Quantilfunktion F_X^{-1} mit der herkömmlichen Inversen von F_X übereinstimmt. Zur Simulation erzeugt man dann eine *Pseudozufallszahl* u , die ein Zufallsgenerator des Rechners hergibt, und die (näherungsweise) einer Realisierung der Zufallsvariablen $U \sim \mathcal{U}(0, 1)$ entspricht. Der Ausdruck $F_X^{-1}(u)$ liefert dann eine Realisierung von X .

Die Inversionsmethode gehört zur Klasse der sog. *Transformationsmethoden* zur Simulation von Zufallsvariablen. Diese bauen auf den Zusammenhängen der Art $X \stackrel{d}{=} T(Y)$ auf, wobei man den Zufallsvektor $Y = (Y_1, Y_2, \dots, Y_m)$, $m \in \mathbb{N}$ simulieren kann, und die messbare Transformation $T : A \rightarrow \mathbb{R}$ findet, $A \in \mathcal{B}_{\mathbb{R}^m}$, so dass man aus Y die gesuchte Zufallsvariable X bekommt.

Beispiel 6.2 1. *Exponentialverteilung*: Sei $X \sim Exp(\lambda)$ für ein $\lambda > 0$.

Ihre Quantilfunktion ist gegeben durch $F_X^{-1}(y) = -\lambda^{-1} \log(1 - y)$, $y \in [0, 1]$. Damit kann X per Inversionsmethode durch

$$X \stackrel{d}{=} -\lambda^{-1} \log(1 - U) \stackrel{d}{=} -\lambda^{-1} \log U, \quad U \sim \mathcal{U}(0, 1)$$

simuliert werden, weil hier $1 - U \stackrel{d}{=} U$ gilt. Da $P(U = 1) = P(U = 0) = 0$, sind Realisierungen von X f.s. endlich.

2. *Bernoulli-Verteilung*: Um $X \sim Ber(p)$, $p \in (0, 1)$ zu simulieren, kann die Inversionsmethode wie folgt verwendet werden. Es gilt

$$F_X(x) = \begin{cases} 1, & x \geq 1, \\ 1 - p, & x \in [0, 1), \\ 0, & x < 0 \end{cases}$$

und somit $F_X^{-1}(y) = I(y > 1 - p)$. Daher simuliere

$$X = I(U > 1 - p) \stackrel{d}{=} I(1 - U < p) \stackrel{d}{=} I(U \leq p)$$

für $U \sim \mathcal{U}(0, 1)$. Die Simulation weiterer diskret verteilten Zufallsvariablen mittels Inversionsmethode wird in Abschnitt 6.5 besprochen.

3. *Pareto–Verteilung:* Mit Hilfe von Beispiel ?? ?? lässt sich leicht zeigen, dass für $X \sim Par(\alpha, \mu)$, $\alpha, \mu > 0$ gilt $X \stackrel{d}{=} \mu U^{-1/\alpha}$, wobei $U \sim \mathcal{U}(0, 1)$.

Da Quantilfunktionen nicht immer analytisch gegeben sind (wie etwa im Fall einer Normalverteilung), sind somit natürliche Grenzen für die Verwendung der Inversionsmethode gesetzt. Deshalb gibt es weitere Simulationsmethoden für Zufallsvariablen, die auf anderen Ideen basieren, wie z.B. die *Akzeptanz– und Verwerfungsmethode*, die wir gleich beschreiben werden.

6.3 Akzeptanz– und Verwerfungsmethode

Sei $X : \Omega \rightarrow \mathbb{R}$ eine Zufallsvariable auf dem Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) , die absolut stetig verteilt ist mit Dichte f_X . Betrachten wir die Klasse der auf \mathbb{R} integrierbaren Funktionen $f \geq 0$, die proportional zu f_X sind: $f(x) = cf_X(x)$, $x \in \mathbb{R}$ für ein $c > 0$. Somit gilt

$$f_X(x) = \frac{f(x)}{\int_{\mathbb{R}} f(y) dy}, \quad x \in \mathbb{R}. \quad (6.3)$$

Nun wählen wir eine konkrete Funktion f aus dieser Klasse aus.

Nehmen wir an, dass eine auf \mathbb{R} Lebesgue–integrierbare Funktion $g \geq 0$ mit den folgenden Eigenschaften existiert:

- $g(x) \geq f(x)$ für alle $x \in \mathbb{R}$,
- Wir können eine Zufallsvariable Y , die absolut stetig verteilt ist mit Dichte

$$g_Y(x) = \frac{g(x)}{\int_{\mathbb{R}} g(y) dy}, \quad x \in \mathbb{R}, \quad (6.4)$$

simulieren.

Die *Akzeptanz– und Verwerfungsmethode* besteht aus den folgenden Schritten:

1. Simuliere die Zufallsvariablen Y und $U \sim \mathcal{U}(0, 1)$ unabhängig voneinander.
2. Falls $Ug(Y) \leq f(Y)$, liefere Y .
3. Andernfalls, gehe zu 1.

Diese Schritte sind in Abbildung 6.1 veranschaulicht.

Im Folgenden bedeutet die Bezeichnung $Y | A$ die Einschränkung einer Zufallsvariablen $Y : \Omega \rightarrow \mathbb{R}$ auf das Ereignis $A \in \mathcal{F}$, oder, genauer gesagt, $Y(\omega)$ für $\omega \in A$.

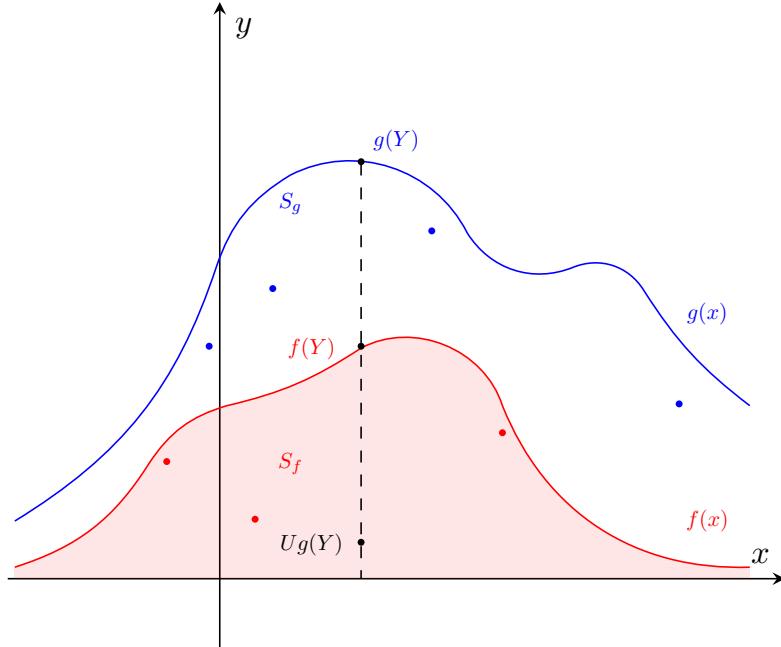


Abbildung 6.1: Akzeptanz– und Verwerfungsmethode.

Satz 6.3 Seien die Zufallsvariablen X und Y wie oben eingeführt. Die Zufallsvariable

$$\tilde{X} = Y \mid \{Ug(Y) \leq f(Y)\} \quad (6.5)$$

hat dieselbe Verteilung wie X .

Um diesen Satz zu beweisen, wird eine Reihe Hilfsergebnisse gebraucht:

Lemma 6.4 Sei B eine beschränkte Borelmenge in \mathbb{R}^d , $d \geq 1$, mit positivem Volumen $|B| > 0$. Sei $Z \sim \mathcal{U}(B)$ ein d -dimensionaler Zufallsvektor $Z = (Z_1, \dots, Z_d)$. Falls $B_0 \subset B$, $|B_0| > 0$ eine Borelmenge ist, so ist die bedingte Verteilung des Vektors $Z \mid \{Z \in B_0\}$ eine Gleichverteilung $\mathcal{U}(B_0)$ auf B_0 .

Beweis Für eine beliebige Borelmenge $A \subset \mathbb{R}^d$ gilt

$$P(Z \in A \mid Z \in B_0) = \frac{P(Z \in A \cap B_0)}{P(Z \in B_0)} = \frac{|A \cap B_0 \cap B|/|B|}{|B_0 \cap B|/|B|} = \frac{|A \cap B_0|}{|B_0|}.$$

□

Für eine beliebige nichtnegative Funktion $g : \mathbb{R} \rightarrow \mathbb{R}_+$ bezeichnen wir durch $S_g = \{(x, y) \in \mathbb{R}^2 : 0 \leq y \leq g(x)\}$ ihren *Untergraphen*.

Lemma 6.5 Sei g eine nichtnegative auf \mathbb{R} Lebesgue-integrierbare Funktion und $U \sim \mathcal{U}(0, 1)$ eine Zufallsvariable. Dann gilt Folgendes:

1. Falls der zweidimensionale Zufallsvektor $(X_1, X_2) \sim \mathcal{U}(S_g)$, dann ist seine erste Koordinate X_1 eine absolut stetig verteilte Zufallsvariable mit der Dichte

$$g_{X_1}(x) = \frac{g(x)}{\int_{\mathbb{R}} g(y) dy}, \quad x \in \mathbb{R}. \quad (6.6)$$

2. Falls die Zufallsvariable X_1 absolut stetig verteilt ist mit Dichte (6.6), dann gilt

$$(X_1, Ug(X_1)) \sim \mathcal{U}(S_g).$$

Beweis

1. Für eine beliebige Borelmenge $B \subset \mathbb{R}$ gilt

$$\begin{aligned} P(X_1 \in B) &= P((X_1, X_2) \in B \times \mathbb{R}) = \frac{|(B \times \mathbb{R}) \cap S_g|}{|S_g|} \\ &= \iint_{(B \times \mathbb{R}) \cap S_g} \frac{1}{|S_g|} dy dx = \frac{\int_B \int_0^{g(x)} dy dx}{\int_{\mathbb{R}} g(x) dx} = \frac{\int_B g(x) dx}{\int_{\mathbb{R}} g(x) dx}. \end{aligned}$$

Somit ist Punkt 1 bewiesen.

2. Es gilt für eine beliebige Borelmenge B aus \mathbb{R}^2

$$\begin{aligned} P((X_1, Ug(X_1)) \in B) &= \mathbb{E}\left(\mathbb{E} I_{\{(X_1, Ug(X_1)) \in B\}} | X_1\right) \\ &= \int_{\mathbb{R}} \mathbb{E} I_{\{(x, Ug(x)) \in B \cap S_g\}} g_{X_1}(x) dx \\ &= \int_{\mathbb{R}} P(Ug(x) \in B \cap S_g \cap L_x) g_{X_1}(x) dx \\ &= \int_{\mathbb{R}} P\left(U \in \frac{1}{g(x)}(B \cap S_g \cap L_x)\right) g_{X_1}(x) dx \\ &= \int_{\mathbb{R}} \left| \frac{1}{g(x)}(B \cap S_g \cap L_x) \right| \frac{g(x)}{|S_g|} dx \\ &= \int_{\mathbb{R}} \frac{|B \cap S_g \cap L_x|}{|S_g|} dx = \frac{|B \cap S_g|}{|S_g|}, \end{aligned}$$

wobei L_x die Achse $\{(x, y) : y \in \mathbb{R}\}$ bezeichnet und in der letzten Gleichung der Satz von Fubini benutzt wurde; vgl. Abbildung 6.2. Somit ist der Zufallsvektor $(X_1, Ug(X_1))$ gleichverteilt auf dem Gebiet S_g .

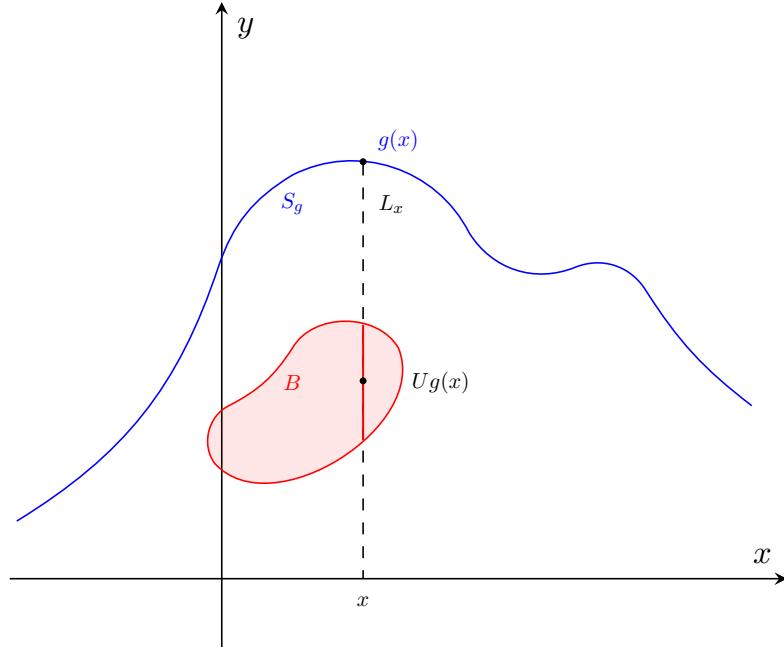


Abbildung 6.2: Illustration zum Beweis des Lemmas 6.5.

□

Beweis des Theorems 6.3

Warum funktioniert nun die Akzeptanz- und Verwerfungsmethode? Sei $g(x) \geq f(x)$, $x \in \mathbb{R}$. Die Zufallsvariable Y sei absolut stetig verteilt mit Dichte (6.4), und sei $U \sim \mathcal{U}(0, 1)$. Nach Lemma 6.5 2 gilt $(Y, Ug(Y)) \sim \mathcal{U}(S_g)$. Falls aber $(Y, Ug(Y)) \in S_f \subset S_g$, so gilt $(Y, Ug(Y)) \sim \mathcal{U}(S_f)$; vgl. Lemma 6.4. So mit hat nach Lemma 6.5 1 die Zufallsvariable Y die Dichte (6.3), und die Relation (6.5) ist bewiesen. □

Bemerkung 6.6 Intuition Akzeptanz- und Verwerfungsmethode Der geometrische Sinn der obigen Vorgehensweise ist folgender: Falls ein zufälliger Punkt mit Koordinaten, die gleichverteilt auf S_g sind, unter den Graph von f fällt, wird seine x -Koordinate als Ergebnis des Algorithmus ausgegeben. Alle Punkte, die oberhalb des Graphen von f fallen, werden verworfen; vgl. Abbildung 6.1.

In jedem Schritt der Akzeptanz- und Verwerfungsmethode wird der Vorschlag Y akzeptiert, falls $Ug(Y) \leq f(Y)$. Sonst wird ein neuer unabhängiger Wert von Y generiert. Sei M die Anzahl solcher Schritte bis zur ersten Akzeptanz von Y . Die Größe M ist offensichtlich geometrisch verteilt mit Parameter

$$p = P(Ug(Y) \leq f(Y)).$$

Somit ist die Erfolgswahrscheinlichkeit gleich

$$p = \int_{\mathbb{R}} P\left(U \leq \frac{f(y)}{g(y)}\right) g_Y(y) dy = \int_{\mathbb{R}} \frac{f(y)}{g(y)} \frac{g(y)}{|S_g|} dy = \frac{|S_f|}{|S_g|}.$$

Folglich ist die mittlere Anzahl der notwendigen Simulationsschritte gleich

$$\mathbb{E} M = \frac{1}{p} = \frac{|S_g|}{|S_f|} = \frac{\int_{\mathbb{R}} g(x) dx}{\int_{\mathbb{R}} f(x) dx} > 1. \quad (6.7)$$

Das heißt, je besser die obere Schranke g für f ist, desto schneller ist im Mittel die Simulation. Die Varianz von M ist gleich

$$\text{Var } M = \frac{1}{p^2} - \frac{1}{p} = \frac{1}{p} \left(\frac{1}{p} - 1 \right) = \frac{|S_g|}{|S_f|} \left(\frac{|S_g|}{|S_f|} - 1 \right).$$

Ein Beispiel von Verwendung der Akzeptanz– und Verwerfungsmethode zur Simulation der Normalverteilung geben wir in Abschnitt 6.4.1.

Übungsaufgabe 6.7 (Gamma–Verteilung)

Wie lautet der Akzeptanz– und Verwerfungsalgorithmus zur Simulation von $X \sim \Gamma(1, a)$, $a > 0$, $a \neq 1$? Verwenden Sie dabei die Funktion

$$g(x) = \begin{cases} x^{a-1} I_{(0 \leq x < 1)} + e^{-x} I_{(x \geq 1)}, & a < 1, \\ \frac{x^{\lambda-1}}{(x^\lambda + a^\lambda)^2} I_{(x \geq 0)}, & a > 1, \end{cases}$$

mit $\lambda = \sqrt{2a - 1}$. Geben Sie die mittlere Anzahl und die Varianz der notwendigen Simulationsschritte an.

6.4 Simulation der Normalverteilung

Wie bereits erwähnt wurde, gibt es keinen geschlossenen Ausdruck für die Quantilfunktion der Standardnormalverteilung, was die Verwendung der Inversionsmethode erschwert. Deshalb werden in diesem Abschnitt andere Simulationsmöglichkeiten für $N(0, 1)$ aufgezeigt, z.B. mit Hilfe der Akzeptanz– und Verwerfungsmethode oder der Box–Muller–Transformation. Diese Algorithmen, im Gegensatz zu den Approximationsansätzen basierend auf dem zentralen Grenzwertsatz (Theorem 5.6), sind genau, also liefern exakte Realisierungen von $N(0, 1)$ –verteilter Zufallsvariable. Eine Realisierung von einer Zufallsvariablen $Z \sim N(\mu, \sigma^2)$ bekommt man durch die Relation $Z = \mu + \sigma X$, wobei X eine Realisierung von $N(0, 1)$ darstellt.

6.4.1 Akzeptanz– und Verwerfungsmethode für $N(0, 1)$

Sei die Zufallsvariable X standardnormalverteilt mit Dichte

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}.$$

Es gilt $f_X(x) \leq g(x) = \sqrt{\frac{e}{2\pi}} e^{-|x|}$, $x \in \mathbb{R}$, weil aus

$$x^2 - 2|x| + 1 = (x \pm 1)^2 \geq 0$$

die Ungleichung $e^{-x^2/2} \leq e^{1/2-|x|}$ folgt. Sei die Zufallsvariable Y absolut stetig verteilt mit einer Dichte, die proportional zu $g(x)$ ist. Dies bedeutet, dass $|Y| \sim Exp(1)$ und daher $|Y| \stackrel{d}{=} -\log V$ für eine Zufallsvariable $V \sim \mathcal{U}(0, 1)$; vgl. Beispiel 6.2 1. Für ein $U \sim \mathcal{U}(0, 1)$ sieht das Akzeptanz–Kriterium $Ug(Y) \leq f_X(Y)$ folgendermaßen aus:

$$U \frac{1}{\sqrt{2\pi}} e^{1/2-|Y|} \leq \frac{1}{\sqrt{2\pi}} e^{-|Y|^2/2},$$

was sich umschreiben lässt wie

$$U e^{1/2-(-\log V)} \leq e^{-(\log^2 V)/2},$$

wobei die Zufallsvariablen U und V stochastisch unabhängig sind. Beide Seiten der Ungleichung logarithmierend, bekommt man

$$\log U + \log V + \frac{1}{2} \leq -\frac{1}{2} \log^2 V,$$

oder, äquivalent dazu,

$$2 \log U \leq -(\log V + 1)^2. \tag{6.8}$$

Falls die Bedingung (6.8) erfüllt ist, wird der Wert von $-\log V$ als eine Realisierung von $|X|$ akzeptiert. Da die Zufallsvariable $|X|$ symmetrisch ist, kann das Vorzeichen von X stochastisch unabhängig von U und V mit Wahrscheinlichkeit 1/2 gewählt werden. Mit anderen Worten, es gilt $X \stackrel{d}{=} (2B - 1)\log V$, wobei $B \sim Ber(1/2)$ eine von U, V stochastisch unabhängige Zufallsvariable ist. Offensichtlich nimmt $2B - 1$ Werte ± 1 mit Wahrscheinlichkeit 1/2 an. Die Simulation von B als $I(2W \leq 1)$ für $W \sim \mathcal{U}(0, 1)$ ist in Beispiel 6.2 2 bereits beschrieben worden.

Zusammengefasst, sieht der Algorithmus zur Simulation von $X \sim N(0, 1)$ wie folgt aus:

Bemerkung 6.8 Simulation von $N(0, 1)$ mit Akzeptanz– und Verwerfungsmethode

1. Simuliere $U, V, W \sim \mathcal{U}(0, 1)$ stochastisch unabhängig voneinander.

2. Falls $2 \log U \leq -(\log V + 1)^2$ gilt, liefere $X = (2I(2W \leq 1) - 1) \log V$.
3. Sonst gehe zu Schritt 1.

Dieser Algorithmus ist ziemlich schnell, denn die mittlere Anzahl M der Simulationsschritte ist nach (6.7) gleich

$$\mathbb{E} M = \frac{|S_g|}{|S_{f_X}|} = \frac{\sqrt{\frac{e}{2\pi}} \int_{\mathbb{R}} e^{-|x|} dx}{1} = \sqrt{\frac{2e}{\pi}} \int_0^\infty e^{-x} dx = \sqrt{\frac{2e}{\pi}} \approx 1.315$$

mit der Varianz $\text{Var } M \approx 1.315(1.315 - 1) = 0.414225$.

6.4.2 Box–Muller–Transformation

Eine Alternativmethode zur Simulation von der Standardnormalverteilung, auch *Polarmethode* genannt (vgl. z.B. [139, Sektion 5.3]), liefert die sog. *Box–Muller–Transformation*

$$X \stackrel{d}{=} \sqrt{-2 \log U} \cos(2\pi V) \stackrel{d}{=} \sqrt{-2 \log U} \sin(2\pi V),$$

wobei $U, V \sim \mathcal{U}(0, 1)$ stochastisch unabhängig sind. Sie basiert auf folgendem Lemma:

Lemma 6.9

1. Seien (R, Θ) die Polarkoordinaten des Zufallsvektors (X_1, X_2) , wobei X_1 und X_2 zwei $N(0, 1)$ –verteilte, stochastisch unabhängige Zufallsvariablen sind. Dann sind R und Θ stochastisch unabhängige Zufallsvariablen, und es gilt $R^2 \sim \text{Exp}(1/2)$, $\Theta \sim \mathcal{U}(0, 2\pi)$.
2. Falls $R^2 \sim \text{Exp}(1/2)$ und $\Theta \sim \mathcal{U}(0, 2\pi)$ stochastisch unabhängige Zufallsvariablen sind, dann sind die Zufallsvariablen $X_1 = R \cos \Theta$ und $X_2 = R \sin \Theta$ standardnormalverteilt und stochastisch unabhängig.

Beweis 1. Seien $(r, \varphi) = (r(x, y), \varphi(x, y))$ die Polarkoordinaten des Vektors $(x, y) \in \mathbb{R}^2$. Für beliebige $z \geq 0$ und $\alpha \in [0, 2\pi)$ gilt

$$\begin{aligned} P(R^2 \leq z, \Theta \leq \alpha) &= \frac{1}{(\sqrt{2\pi})^2} \iint_{\substack{x^2+y^2 \leq z \\ \varphi(x,y) \leq \alpha}} e^{-\frac{x^2+y^2}{2}} dx dy \\ &= \frac{1}{2\pi} \int_0^\alpha \int_0^{\sqrt{z}} e^{-\frac{r^2}{2}} r dr d\varphi \stackrel{(t=r^2)}{=} \int_0^\alpha \frac{d\varphi}{2\pi} \int_0^z \frac{1}{2} e^{-\frac{t}{2}} dt. \end{aligned}$$

Somit ist die Dichte von (R^2, Θ) ein Produkt von zwei Komponenten, die die Dichten der $\text{Exp}(1/2)$ – und $\mathcal{U}(0, 2\pi)$ –Verteilungen sind. Nach Theorem 3.30 2 sind R^2, Θ stochastisch unabhängig mit vorgegebenen Dichten.

2. Für beliebige $x, y \in \mathbb{R}$ gilt

$$\begin{aligned}
P(X_1 \leq x, X_2 \leq y) &= P(\sqrt{R^2} \cos \Theta \leq x, \sqrt{R^2} \sin \Theta \leq y) \\
&= \frac{1}{2\pi} \int_0^{2\pi} \int_0^\infty I(\sqrt{t} \cos \varphi \leq x, \sqrt{t} \sin \varphi \leq y) \frac{1}{2} e^{-\frac{t}{2}} dt d\varphi \\
&\stackrel{(t=r^2)}{=} \frac{1}{2\pi} \int_0^{2\pi} \int_0^\infty I(r \cos \varphi \leq x, r \sin \varphi \leq y) r e^{-\frac{r^2}{2}} dr d\varphi \\
&\stackrel{\left(\begin{array}{l} x_1 = r \cos \varphi \\ x_2 = r \sin \varphi \end{array} \right)}{=} \frac{1}{2\pi} \int_0^\infty \int_0^\infty I(x_1 \leq x, x_2 \leq y) e^{-\frac{x_1^2 + x_2^2}{2}} dx_1 dx_2 \\
&= \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{x_1^2}{2}} dx_1 \cdot \frac{1}{\sqrt{2\pi}} \int_0^y e^{-\frac{x_2^2}{2}} dx_2.
\end{aligned}$$

Somit sind Zufallsvariablen X_1, X_2 stochastisch unabhängig und standardnormalverteilt.

□

Bemerkung 6.10 (Rayleigh–Verteilung) Die oben eingeführte Zufallsvariable $R = \sqrt{X_1^2 + X_2^2}$ ist *Rayleigh–verteilt* mit Parameter $\sigma = 1$. Generell wird die Dichte der *Rayleigh–Verteilung mit Parameter $\sigma > 0$* (wir schreiben abkürzend $RL(\sigma)$) durch den Ausdruck

$$f(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} I(x \geq 0)$$

definiert. Dadurch bekommt die Tailfunktion von $R_\sigma \sim RL(\sigma)$ folgende simple Form:

$$\bar{F}_{R_\sigma}(x) = I(x < 0) + e^{-\frac{x^2}{2\sigma^2}} I(x \geq 0).$$

Die $RL(\sigma)$ -Verteilung ist nach dem britischen Physiker und Nobel–Preisträger *John W. Strutt, 3. Baron Rayleigh* (1842–1919) benannt. Sie dient beispielsweise zur Modellierung von 10-minutigen Mittelwerten der Windgeschwindigkeiten.

Per Definition der χ^2 –Verteilung gilt

$$R^2 = X_1^2 + X_2^2 \sim \chi_2^2 = \Gamma(1/2, 1) = Exp(1/2)$$

nach Anmerkung ???. Außerdem ist die Relation $R_\sigma \stackrel{d}{=} \sigma R$ für jedes $\sigma > 0$ offensichtlich, $R_1 \stackrel{d}{=} R$. Damit gilt

$$R_\sigma \stackrel{d}{=} \sqrt{Z_1^2 + Z_2^2},$$

wobei $Z_1, Z_2 \sim N(0, \sigma^2)$ stochastisch unabhängige Zufallsvariablen sind.

Nach Lemma 6.9 2 kann $X \sim N(0, 1)$ als $R \cos \Theta$ oder $R \sin \Theta$ simuliert werden, wobei $R \stackrel{d}{=} \sqrt{-2 \log U}$ und $\Theta \stackrel{d}{=} 2\pi V$ für $U, V \sim \mathcal{U}(0, 1)$. Dadurch erhalten wir folgenden direkten (d.h., nicht-iterativen) Algorithmus zur Simulation von $N(0, 1)$:

Bemerkung 6.11 Simulation von $N(0, 1)$ mit Polarmethode

1. Simuliere stochastisch unabhängige Zufallsvariablen $U, V \sim \mathcal{U}(0, 1)$.
2. Liefere $X = \sqrt{-2 \log U} \cos(2\pi V)$ oder $X = \sqrt{-2 \log U} \sin(2\pi V)$.

Die Zufallsvariable $Z \sim N(\mu, \sigma^2)$ kann als $\mu + \sigma X$ für beliebige $\mu \in \mathbb{R}, \sigma > 0$ simuliert werden.

6.4.3 Abgeschnittene Normalverteilung

Sei $Z \sim N(0, 1)$. Die *abgeschnittene Normalverteilung* kann als Verteilung der Zufallsvariable $X = Z | \{Z \geq a\}$, $a \in \mathbb{R}$, eingeführt werden. Die Dichte der abgeschnittenen Normalverteilung ist gleich

$$f_X(x) = \frac{e^{-\frac{x^2}{2}} I(x \geq a)}{\int_a^\infty e^{-\frac{y^2}{2}} dy}.$$

Oft interessiert man sich für den Fall $a \geq 0$, weil dadurch nichtnegative Kenngrößen (wie z.B. Risiken, Geschwindigkeiten, Geldbeträge) modelliert werden können.

Falls $a \geq 1$, kann X unter Verwendung von der Akzeptanz- und Verwerfungsmethode simuliert werden, indem die Ungleichung $e^{-\frac{x^2}{2}} \leq x e^{-\frac{x^2}{2}}$, $x \geq 1$ zum Einsatz kommt. Seien $f(x) = e^{-\frac{x^2}{2}} I(x \geq a)$ und $g(x) = x e^{-\frac{x^2}{2}} I(x \geq 0)$. Es gilt $f(x) \leq g(x)$ für alle x , wobei g die Dichte der $RL(1)$ -Verteilung darstellt. Die Zufallsvariable $Y \sim RL(1)$ kann laut Anmerkung 6.10 per Inversionsmethode durch $Y \stackrel{d}{=} \sqrt{-2 \log V}$ für $V \sim \mathcal{U}(0, 1)$ simuliert werden. Die Akzeptanz-Regel $Ug(Y) \leq f(Y)$, $U \sim \mathcal{U}(0, 1)$ lautet hier

$$U \sqrt{-2 \log V} \leq I(\sqrt{-2 \log V} \geq a),$$

oder, in Alternativdarstellung,

$$U^2 \leq \frac{I(V \leq e^{-a^2/2})}{-2 \log V}. \quad (6.9)$$

Führen wir nun die einzelnen Schritte vom Simulationsalgorithmus aus Abschnitt 6.3 zusammen:

Bemerkung 6.12 Simulation der abgeschnittenen Normalverteilung mit Akzeptanz– und Verwerfungsmethode

1. Simuliere $U, V \sim \mathcal{U}(0, 1)$ stochastisch unabhängig von einander.
2. Falls (6.9) gilt, liefere $X = \sqrt{-2 \log V}$.
3. Sonst gehe zu Schritt 1.

Die mittlere Anzahl der notwendigen Simulationsschritte ist hier gleich

$$\mathbb{E} M = \frac{\int_0^\infty y e^{-\frac{y^2}{2}} dy}{\int_a^\infty e^{-\frac{y^2}{2}} dy} = \left(\int_a^\infty e^{-\frac{y^2}{2}} dy \right)^{-1}, \quad a \geq 1.$$

6.5 Simulation von diskret verteilten Zufallsvariablen

Sei X eine diskret verteilte Zufallsvariable, die Werte $x_k \in \mathbb{R}$, $k \in \mathbb{N}$ mit Wahrscheinlichkeiten $p_k = P(X = x_k)$ annimmt. Es gilt $\sum_{k=1}^\infty p_k = 1$. Sei $P_k = p_1 + \dots + p_k$ für alle k . Wie simuliert man X ?

In Beispiel 6.2.2 wurde eine Bernoulli-verteilte Zufallsvariable per Inversionsmethode (vgl. Abschnitt 6.2) simuliert. Diese Methode kann auch zur Simulation von allgemeinen diskret verteilten Zufallsvariablen X wie folgt benutzt werden:

Bemerkung 6.13 Simulation von diskreten Verteilungen mit Inversionsmethode

1. Simuliere eine Zufallsvariable $U \sim \mathcal{U}(0, 1)$.

2. Liefere

$$\tilde{X} = \begin{cases} x_1, & U < P_1, \\ x_2, & P_1 \leq U < P_2, \\ \vdots, & \vdots \\ x_k, & P_{k-1} \leq U < P_k, \\ \vdots & \vdots \end{cases} \quad (6.10)$$

als eine Realisierung von X .

Es gilt offensichtlich $\tilde{X} \stackrel{d}{=} X$.

Diese Methode kann effizient nur für relativ kleine endliche Wertebereiche $C = \{x_1, \dots, x_n\}$ von X (d.h., falls $p_k = 0$ für $k > n$) angewandt werden, weil für große n zu viele Fälle unterschieden werden müssen. Als Ausweg aus dieser Situation können *Markov-Ketten-Monte-Carlo-Methoden* genannt werden, die wegen ihrer Komplexität nicht in diesem Einführungstext behandelt werden; siehe z.B. Bücher [6, 102, 59, 70, 140].

Beispiel 6.14 (Binomialverteilung)

Um eine Zufallsvariable $X \sim Bin(n, p)$ zu simulieren, kann die Darstellung

$$X \stackrel{d}{=} X_1 + \dots + X_n$$

benutzt werden, wobei X_k , $k = 1, \dots, n$ stochastisch unabhängige $Ber(p)$ -verteilte Zufallsvariablen sind. Aus Beispiel 6.2 2 folgt, dass

$$X \stackrel{d}{=} I(U_1 \leq p) + \dots + I(U_n \leq p)$$

für stochastisch unabhängige Zufallsvariablen $U_1, \dots, U_n \sim \mathcal{U}(0, 1)$. Allerdings ist dieser Ansatz für große n offensichtlich ineffizient. In diesem Fall kann die Methode (6.10) benutzt werden, wobei die Wahrscheinlichkeiten $p_k = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, \dots, n$ rekursiv durch die Formel

$$p_k = \frac{n-k+1}{k} \frac{p}{1-p} \cdot p_{k-1}, \quad k = 1, \dots, n, \quad p_0 = (1-p)^n$$

berechnet werden, die einen Spezialfall der sog. *Panjer-Rekursion* darstellt, vgl. z.B. [122, 159], [138, Theorem 4.3.1]. Für kleine $\mathbb{E} X = np$ werden die Vergleiche in (6.10) in natürlicher Reihenfolge durchgeführt. Für große np ist es effizienter mit $P_{[np]}$ anzufangen. Falls n groß und $p \ll 0.25$ klein ist, kann man von der Poisson-Approximation in Theorem 3.14 ?? Gebrauch machen, und zwar gilt $X \approx \tilde{X} \sim Poisson(np)$, siehe Beispiel 6.16 2. Hier ist allerdings mit einem Approximationsfehler zu rechnen, vgl. Anmerkung ?? ??.

Alternativ kann die Akzeptanz- und Verwerfungsmethode verwendet werden. Sei hierfür $X : \Omega \rightarrow C$ eine diskret verteilte Zufallsvariable mit Zähl-dichte $\{p_k\}_{k \in \mathbb{N}}$ und (evtl. unendlichem) Wertebereich $C = \{x_1, \dots, x_n, \dots\}$. Für eine stetig verteilte Zufallsvariable $Z : \Omega \rightarrow \mathbb{R}_+$ mit der stückweise konstanten Dichte

$$f_Z(x) = p_{\lfloor x \rfloor} I(x \geq 1)$$

gilt offensichtlich die Gleichung $X \stackrel{d}{=} x_{\lfloor Z \rfloor}$. Die Zufallsvariable Z kann somit behilflich sein, die Nummer $k \in \mathbb{N}$ des Zustandes $x_k \in C$ von X zu simulieren, was zusammenfassend folgendermassen festzuhalten ist:

Bemerkung 6.15 (Simulation von diskreten Verteilungen mit Akzeptanz- und Verwerfungsmethode)

1. Simuliere die Zufallsvariable Z mit Hilfe der Akzeptanz– und Verwerfungsmethode aus Abschnitt 6.3.
2. Liefere $x_{\lfloor Z \rfloor}$ als eine Realisierung von X .

Es gibt eine Reihe von *ad hoc* Algorithmen für die Simulation von diskret verteilten Zufallsvariablen X , die auf speziellen Eigenschaften der Verteilung von X beruhen. Manche von ihnen werden hier exemplarisch aufgeführt.

Beispiel 6.16 (Ad hoc Methoden)

1. *Geometrische Verteilung:*

Für die Zufallsvariable $X \sim Geo(p)$, $0 < p < 1$, gilt die Gleichung

$$X \stackrel{d}{=} \lfloor \log_{1-p} U \rfloor + 1, \quad U \sim \mathcal{U}(0, 1),$$

die eine geeignete Transformation von Pseudozufallszahl U zur Simulation von X liefert. Um diese Darstellung zu beweisen, schreibt man

$$\begin{aligned} P(\lfloor \log_{1-p} U \rfloor + 1 = k) &= P(k - 1 \leq \log_{1-p} U < k) \\ &= P((1 - p)^k < U \leq (1 - p)^{k-1}) \\ &= (1 - p)^{k-1} - (1 - p)^k = p(1 - p)^{k-1}, \quad k \in \mathbb{N}. \end{aligned}$$

2. *Poisson–Verteilung:*

Sei $X \sim Poisson(\lambda)$. Es gilt die Darstellung

$$X = \inf\{k \in \mathbb{Z}_+ : \sum_{j=1}^{k+1} Y_j > \lambda\} \tag{6.11}$$

für stochastisch unabhängige Zufallsvariablen $Y_j \sim Exp(1)$. Die Inversionsmethode ergibt $Y_j \stackrel{d}{=} -\log U_j$ für $U_j \sim \mathcal{U}(0, 1)$, $j \in \mathbb{N}$. Die Bedingung in (6.11) kann somit als $-\sum_{j=1}^{k+1} \log U_j > \lambda$ oder $\prod_{j=1}^{k+1} U_j < e^{-\lambda}$ umgeschrieben werden. Fassen wir diese Überlegungen in einem Algorithmus zusammen:

- (a) Setze $k = 0$, $T = 1$.
- (b) Simuliere $U \sim \mathcal{U}(0, 1)$ und bilde $T = UT$.
- (c) Falls $T \geq e^{-\lambda}$, setze $k = k + 1$ und gehe zu Schritt 2b.
- (d) Andernfalls, liefere k als Realisierung von X .

Da $\mathbb{E} X = \lambda$, ist die mittlere Anzahl M der Simulationsschritte hier gleich $\lambda + 1$. Somit wird dieser Algorithmus für große λ nicht mehr effizient.

Für große λ kann man die Inversionsmethode (6.10) anwenden, wobei die Zähldichte $p_k = e^{-\lambda} \frac{\lambda^k}{k!}$, $k \in \mathbb{Z}_+$ durch die Panjer-Rekursion

$$p_k = \frac{\lambda}{k} p_{k-1}, \quad k \in \mathbb{N}, \quad p_0 = e^{-\lambda}$$

schnell berechnet werden können, vgl. z.B. [138, Theorem 4.3.1]. Um die Suche in (6.10) für große λ zu optimieren, wird wegen $E N = \lambda$ zunächst U mit $P_{\lfloor \lambda \rfloor}$ verglichen. Falls $U < P_{\lfloor \lambda \rfloor}$ wird weiterhin geprüft, ob $U < P_{\lfloor \lambda \rfloor - 1}$ gilt, usw. Man setzt dann $X = \min\{k : U < P_k\}$. Analog geht man im Falle $U \geq P_{\lfloor \lambda \rfloor}$ vor.

Lemma 6.17 Für große λ , die mittlere Anzahl der Vergleiche ist dabei ungefähr gleich $1 + 0.798\sqrt{\lambda}$.

Beweis Nach dem zentralen Grenzwertsatz (Beispiel ??) gilt

$$\frac{X - \lambda}{\sqrt{\lambda}} \xrightarrow{d} Y \sim N(0, 1), \quad \lambda \rightarrow +\infty. \quad (6.12)$$

Die mittlere Anzahl der Vergleiche ist gleich

$$1 + \mathbb{E}|X - \lambda| = 1 + \sqrt{\lambda} \mathbb{E} \frac{|X - \lambda|}{\sqrt{\lambda}}$$

und kann mit Hilfe von (6.12) als

$$1 + \sqrt{\lambda} \mathbb{E}|Y| \approx 1 + 0.798\sqrt{\lambda}$$

approximiert werden, wobei $\mathbb{E}|Y|$ leicht durch die Erwartung der $RL(1)$ -Verteilung zu berechnen ist. \square

Aus (6.12) folgt die Relation $X \approx \lambda + \sqrt{\lambda}Y$, die für $\lambda \geq 100$ als Approximation

$$X \approx \lfloor \lambda + 0.5 + \sqrt{\lambda}Y \rfloor, \quad Y \sim N(0, 1)$$

nutzbar ist. Es gibt auch genauere approximative Simulationsmethoden für die Poisson-Verteilung wie etwa die *Anscombe*- und *Peizer und Pratt*-Approximationen; vgl. [34, S. 35–37, 142].

6.6 Monte-Carlo- und Quasi-Monte-Carlo-Integration

In Abschnitt ?? wurde bereits die Hauptidee der Integralberechnung mit Monte-Carlo-Methode erläutert. Hier geben wir weitere Finessen davon...

Siehe mehr dazu in den Büchern [172, 6, 157, 102, 139, 160, 59, 70, 140].

Kapitel 7

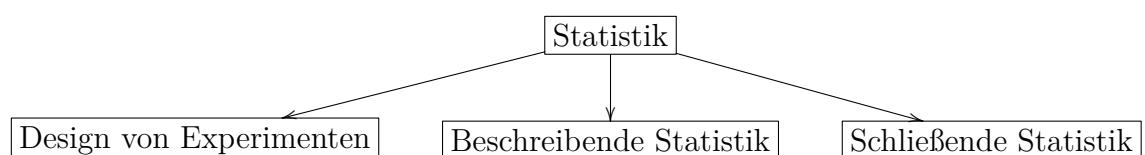
Beschreibende Statistik

7.1 Typische Fragestellungen, Aufgaben und Ziele der Statistik

Im alltäglichen Sprachgebrauch versteht man unter „Statistik“ eine Darstellung von Ergebnissen des Zusammenzählens von Daten und Fakten jeglicher Art, wie z.B. ökonomischen Kenngrößen, politischen Umfragen, Daten der Marktforschung, klinischen Studien in der Biologie und Medizin, usw.

Die *mathematische Statistik* jedoch kann viel mehr. Sie arbeitet mit *Daten-Stichproben*, die nach einem bestimmten Zufallsmechanismus aus der *Grundgesamtheit* aller Daten, die in Folge von Beobachtung, Experimenten (reale Daten) oder Computersimulation (synthetische Daten) erhoben wurden. Dabei beschäftigt sich die mathematische Statistik mit folgenden Fragestellungen:

1. Wie sollen die Daten gewonnen werden? (Design von Experimenten)
2. Wie sollen (insbesondere riesengroße) Datensätze beschrieben werden, um die Gesetzmäßigkeiten und Strukturen in ihnen entdecken zu können? (Beschreibende (deskriptive) und explorative Statistik)
3. Welche Schlüsse kann man aus den Daten ziehen? (Schließende oder induktive Statistik)



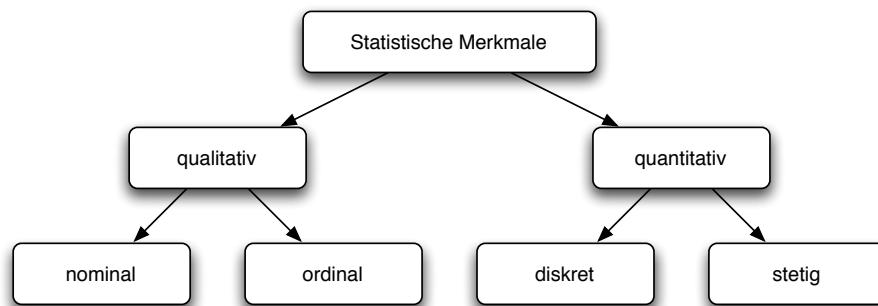
In dieser einführenden Vorlesung werden wir Teile der beschreibenden und schließenden Statistik kennenlernen, wobei die Datenerhebung aus Platzgründen ausgelassen wird. Die *Arbeitsweise eines Statistikers* sieht folgendermaßen aus:

1. *Datenerhebung*
2. *Visualisierung und beschreibende Datenanalyse*
3. *Datenbereinigung* (z.B. Erkennung fehlerhafter Messungen, Ausreißern, usw.)
4. *Explorative Datenanalyse* (Suche nach Gesetzmäßigkeiten)
5. *Modellierung der Daten* mit Methoden der Stochastik
6. *Modellanpassung* (Schätzung der Modellparameter)
7. *Modellvalidierung* (wie gut war die Modellanpassung?)
8. *Schließende Datenanalyse*:
 - Konstruktion von *Vertrauensintervallen* (Konfidenzintervallen) für Modellparameter und deren Funktionen,
 - Tests statistischer Hypothesen,
 - Vorhersage von Zielgrößen (z.B. auf Basis modellbezogener Computersimulation).

Uns werden in diesem Vorlesungsskript vor allem die Arbeitspunkte 2), 4)–6) und 8) beschäftigen.

7.2 Statistische Merkmale und ihre Typen

Die Daten, die zur statistischen Analyse vorliegen, können eine oder mehrere interessierende Größen (die auch *Variablen* oder *Merkmale* genannt werden) umfassen. Ihre Werte werden *Merkmalsausprägungen* genannt. In dem nachfolgenden Diagramm werden mögliche Typen der statistischen Merkmale gegeben.



Diese Typen entstehen in Folge der Klassifikation von Wertebereichen (Skalen) der Merkmale. Dennoch ist diese Einteilung nicht vollständig und kann bei Bedarf erweitert werden. Man unterscheidet *qualitative* und *quantitative* Merkmale. *Quantitative Merkmale* lassen sich inhaltlich gut durch Zahlen darstellen (z.B. Kredithöhe in €, Körpergewicht und Körpergröße, Blutdruck usw.). Sie können *diskrete* oder *stetige* Wertebereiche haben, wobei diskrete Merkmale isolierte Werte annehmen können (z.B. Anzahl der Schäden eines Versicherers pro Jahr). Stetige Wertebereiche hingegen sind überabzählbar. Dennoch liegen in der Praxis stetige Merkmale in gerundeter Form vor (z.B. Körpergröße auf cm gerundet, Geldbeträge auf € gerundet usw.).

Im Gegensatz zu den quantitativen Merkmalen sind die Inhalte der *qualitativen Merkmale*, wie z.B. Blutgruppe (0, A, B und AB) oder Familienstand (ledig, verheiratet, verwitwet), nicht sinnvoll durch Zahlen darzustellen. Sie können zwar formell mit Zahlen kodiert werden (z.B. bei Blutgruppen 0 = 0, A = 1, B = 2, AB = 3), aber solche Kodierungen stellen keinen inhaltlichen Zusammenhang zwischen Ausprägungen und Zahlen-Codes dar sondern dienen lediglich der besseren Identifikation der Merkmale auf einem Rechner. Es ist insbesondere unsinnig, Mittelwerte und ähnliches von solchen Codes zu bilden.

Ein qualitatives Merkmal mit nur 2 Ausprägungen (z.B. männlich / weiblich, Raucher / Nichtraucher) heißt *alternativ*. Ein qualitatives Merkmal kann *ordinal* (wenn sich eine natürliche lineare Ordnung in den Merkmalsausprägungen finden lässt, wie z.B. gut / mittel / schlecht bei Qualitätsbewertung in Umfragen oder sehr gut / gut / befriedigend / ausreichend / mangelhaft / ungenügend bei Schulnoten) oder *nominal* (wenn eine solche Ordnung nicht vorhanden ist) sein. Beispiele von nominalen Merkmalen sind Fahrzeugmarken in der KFZ-Versicherung (z.B. BMW, Peugeot, Volvo, usw.) oder Führerscheinklassen (A, B, C, ...). Datenmerkmale können auch mehrdimensionale Ausprägungen haben. In dieser Vorlesung behandeln wir jedoch hauptsächlich eindimensionale Merkmale.

7.3 Statistische Daten und Stichproben

Aus den obigen Beispielen wird klar, dass ein Statistiker mit Datensätzen der Form (x_1, \dots, x_n) arbeitet, wobei die Einzeleinträge x_i aus einer Grundgesamtheit $G \subset \mathbb{R}^k$ stammen, die hypothetisch unendlich groß ist. Der vorliegende Datensatz (x_1, \dots, x_n) wird auch (*konkrete*) *Stichprobe* von Umfang n genannt. Die Menge B aller potentiell möglichen Stichproben bezeichnen wir als *Stichprobenraum* und setzen zur Vereinfachung der Notation $B = \mathbb{R}^{kn}$. In diesem Skript werden wir meistens die univariate statistische Analyse (also $k = 1$, ein eindimensionales Merkmal) betreiben. In der beschreibenden Statistik arbeitet man mit Stichproben (x_1, \dots, x_n) und ihren

Funktionen, um diese Daten visualisieren zu können. Für die Aufgabe der schließenden Statistik jedoch reicht diese Datenebene nicht mehr aus. Daher wird die zweite Ebene der Betrachtung eingeführt, die sogenannte *Modellebene*. Dabei wird angenommen, dass die konkrete Stichprobe (x_1, \dots, x_n) eine *Realisierung* eines stochastischen Modells (X_1, \dots, X_n) darstellt, wobei X_1, \dots, X_n (meistens unabhängige identisch verteilte) Zufallsvariablen auf einem (nicht näher spezifizierten) Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) sind. Diese Zufallsvariablen $X_i, i = 1, \dots, n$ können als konsequente Beobachtungen eines Merkmals interpretiert werden.

Der Vektor (X_1, \dots, X_n) wird dabei *Zufallsstichprobe* genannt. Man setzt weiter voraus, dass $EX_i^2 < \infty \forall i = 1, \dots, n$, damit man von der Varianz $\text{Var } X_i$ der Einzeleinträge sprechen kann. Es wird außerdem angenommen, dass ein $\omega \in \Omega$ existiert, sodass $X_i(\omega) = x_i \forall i = 1, \dots, n$. Sei F die Verteilungsfunktion der Zufallsvariablen X_i . Eine der wichtigsten Aufgaben der Statistik ist die Bestimmung von F (man sagt, „Schätzung von F “) aus den konkreten Daten (x_1, \dots, x_n) . Dabei können auch Momente von F und ihre Funktionen (Erwartungswert, Varianz, Schiefe, usw.) von Interesse sein.

7.4 Stichprobenfunktionen

Um die obigen Aufgaben erfüllen zu können, braucht man gewisse Funktionen $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^m, m \in \mathbb{N}$ auf dem Stichprobenraum, die diese Stichprobe bewerten.

Definition 7.1 Eine Borel-messbare Abbildung $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ heißt *Stichprobenfunktion*. Wenn man auf der Modellebene mit einer Zufallsstichprobe (X_1, \dots, X_n) arbeitet, so heißt die Zufallsvariable

$$\varphi(X_1, \dots, X_n)$$

eine *Statistik*. In der Schätztheorie spricht man dabei von *Schätzern* und bei statistischen Tests wird $\varphi(X_1, \dots, X_n)$ *Teststatistik* genannt.

Beispiele für Stichprobenfunktionen sind unter anderen das *Stichprobenmittel*

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i,$$

die *Stichprobenvarianz*

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

und die *Ordnungsstatistiken*

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)},$$

die entstehen, wenn man eine Stichprobe, die aus quantitativen Merkmalen besteht, linear ordnet ($x_{(1)} = \min_{i=1,\dots,n} x_i, \dots, x_{(n)} = \max_{i=1,\dots,n} x_i$). Weitere Beispiele und ihre Charakteristiken werden in Kapitel 7 gegeben.

Sei eine konkrete Stichprobe (x_1, \dots, x_n) , $x_i \in \mathbb{R}$ gegeben, wobei die x_i als Realisierungen der Zufallsvariablen $X_i \stackrel{d}{=} X$ mit Verteilungsfunktion F interpretiert werden können.

7.5 Verteilungen und ihre Darstellungen

In diesem Abschnitt werden wir Methoden zur statistischen Beschreibung und grafischen Darstellung der (unbekannten) Verteilung F betrachten.

7.5.1 Häufigkeiten und Diagramme

Falls das quantitative Merkmal X eine endliche Anzahl von Ausprägungen $\{a_1, \dots, a_k\}$, $a_1 < a_2 < \dots < a_k$, besitzt, also

$$P(X \in \{a_1, \dots, a_k\}) = 1,$$

dann kann eine Schätzung der Zähldichte $p_i = P(X = a_i)$ von X aus den Daten (x_1, \dots, x_n) grafisch dargestellt werden. Ähnliche Darstellungen sind für die Dichte $f(x)$ von absolut stetigen Merkmalen X möglich, wobei ihr Wertebereich C sich in k Klassen aufteilen lässt: $(c_{i-1}, c_i]$, $i = 1, \dots, k$, wobei $c_0 = -\infty$, $c_1 < \dots < c_{k-1}$, $c_k = \infty$ ist. Dann kann die Zähldichte $p_i = P(X \in (c_{i-1}, c_i])$ gegeben durch

$$p_i = \int_{c_{i-1}}^{c_i} f(x) dx_i, \quad i = 0, \dots, k$$

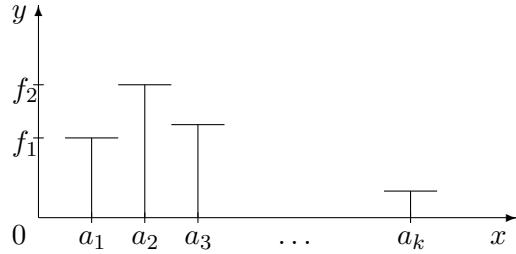
betrachtet werden.

Definition 7.2

1. Die *absolute Häufigkeit* von Merkmalsausprägung a_i bzw. Klasse $(c_{i-1}, c_i]$, $i = 1, \dots, k$ ist $n_i = \#\{x_j, j = 1, \dots, n : x_j = a_i\}$ bzw. $n_i = \#\{x_j, j = 1, \dots, n : x_j \in (c_{i-1}, c_i]\}$.
2. Die *relative Häufigkeit* von Merkmalsausprägung a_i bzw. Klasse $(c_{i-1}, c_i]$ ist $f_i = n_i/n$, $i = 1, \dots, k$.

Es gilt offensichtlich $n = \sum_{i=1}^k n_i$, $0 \leq f_i \leq 1$, $\sum_{i=1}^k f_i = 1$. Die absoluten und relativen Häufigkeiten werden oft in Häufigkeitstabellen zusammengefasst. Zu ihrer Visualisierung dienen so genannte *Diagramme*. *Histogramme* werden gebildet, indem man die Paare (a_i, f_i) (bzw. $(1/2(c_1 + x_{(1)}), f_1)$, $(1/2(c_{i-1} + c_i), f_i)$, $i = 2, \dots, k-1$, $(1/2(c_{k-1} + x_{(n)}), f_k)$) im absolut stetigen Fall, wobei hier die Bezeichnung $a_i = 1/2(c_{i-1} + c_i)$ verwendet wird und $x_{(1)} < c_1$, $x_{(n)} > c_{k-1}$ angenommen wird.) auf der Koordinatenebene (x, y) folgendermaßen aufträgt:

- *Stabdiagramm:* f_i wird als Höhe des senkrechten Strichs über a_i dargestellt:



- *Säulendiagramm:* genauso wie ein Stabdiagramm, nur werden Striche durch Säulen der Form $(c_{i-1}, c_i] \times f_i$ ersetzt, wobei im diskreten Fall die Aufteilung der reellen Achse $-\infty = c_0 < c_1 < c_2 < \dots < c_{k-1} < c_k = \infty$ in Intervalle beliebig vorgenommen werden kann.

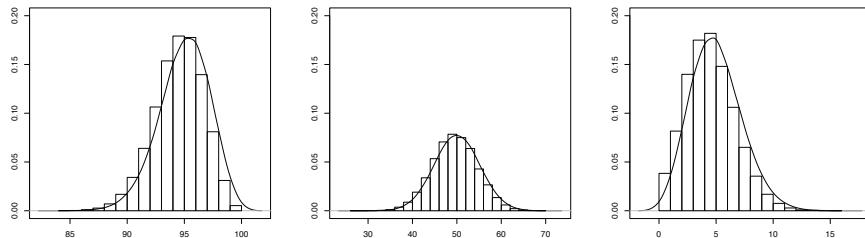
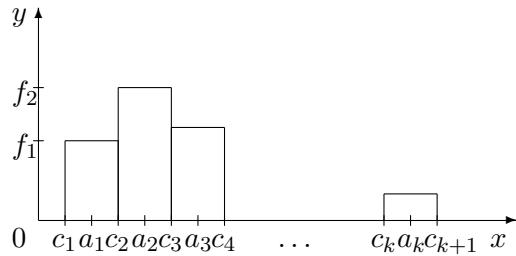


Abbildung 7.1: Das Histogramm der Daten mit einer rechtssteilen (links-schiefen), symmetrischen und linkssteilen (rechtsschiefen) Verteilung und ihre Dichte.

Bemerkung 7.3 Die in Abschnitt 7.5.1 betrachteten Methoden dienen der Visualisierung von (Zähl-) Dichten der Verteilung eines beobachteten Merkmals X . Aus dem Histogramm kann z.B. die Interpretation der Form der Dichte abgelesen werden:

Ist die zugrundeliegende Verteilung F_X symmetrisch bzw. linkssteil (rechts-schif) oder rechtssteil (linksschif) (vgl. Abb. 7.1) oder ist sie unimodal (d.h.

eingipflig), bimodal (d.h. mit 2 Gipfeln) oder multimodal (also mit mehreren Gipfeln) (vgl. Abb. 7.2).

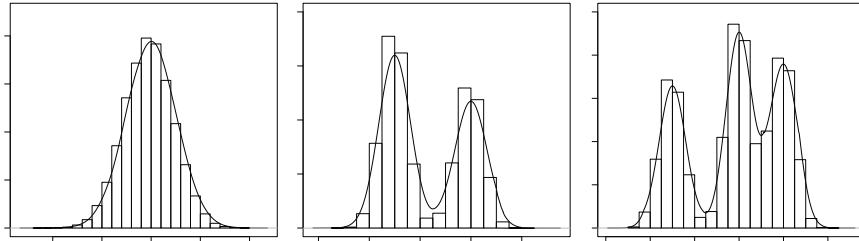


Abbildung 7.2: Histogramm der Daten mit der Dichte einer unimodalen, bimodalen und multimodalen Verteilung

7.5.2 Empirische Verteilungsfunktion

Es sei eine konkrete Stichprobe (x_1, \dots, x_n) gegeben, die eine Realisierung des statistischen Modells (X_1, \dots, X_n) ist, wobei X_1, \dots, X_n unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion $F_X : X_i \stackrel{d}{=} X \sim F_X$ sind. Wie kann die unbekannte Verteilungsfunktion F_X aus den Daten (x_1, \dots, x_n) rekonstruiert (die Statistiker sagen „geschätzt“) werden? Dies ist mit Hilfe der sogenannten empirischen Verteilungsfunktion möglich:

Definition 7.4

1. Die Funktion $\hat{F}_n(x) = \#\{x_i : x_i \leq x, i = 1, \dots, n\}/n, \quad \forall x \in \mathbb{R}$ heißt *empirische Verteilungsfunktion der konkreten Stichprobe* (x_1, \dots, x_n) . Dabei gilt $\hat{F}_n : \mathbb{R}^{n+1} \rightarrow [0, 1]$, weil $\hat{F}_n(x) = \varphi(x_1, \dots, x_n, x)$.
2. Die mit $x \in \mathbb{R}$ indizierte Zufallsvariable $\hat{F}_n : \Omega \times \mathbb{R} \rightarrow [0, 1]$ heißt *empirische Verteilungsfunktion der Zufallsstichprobe* (X_1, \dots, X_n) , wenn

$$\hat{F}_n(x, \omega) = \hat{F}_n(x) = \frac{1}{n} \#\{X_i, i = 1, \dots, n : X_i(\omega) \leq x\}, \quad x \in \mathbb{R}.$$

Äquivalent zur Definition 7.4 kann man

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x), \quad x \in \mathbb{R}$$

schreiben, wobei

$$I(x \in A) = \begin{cases} 1, & x \in A \\ 0, & \text{sonst.} \end{cases}$$

Es gilt

$$\hat{F}_n(x) = \begin{cases} 1, & x \geq x_{(n)}, \\ \frac{i}{n}, & x_{(i)} \leq x < x_{(i+1)}, \quad i = 1, \dots, n-1, \\ 0, & x < x_{(1)}. \end{cases}$$

für $x_{(1)} < x_{(2)} < \dots < x_{(n)}$.

Dabei ist die Höhe des Sprungs an Stelle $x_{(i)}$ gleich der relativen Häufigkeit f_i des Wertes $x_{(i)}$. Falls $x_{(i)} = x_{(i+1)}$ für ein $i \in \{1, \dots, n\}$, so tritt der Wert i/n nicht auf. In Abbildung 7.3 sieht man, dass $\hat{F}_n(x)$ eine rechtss-

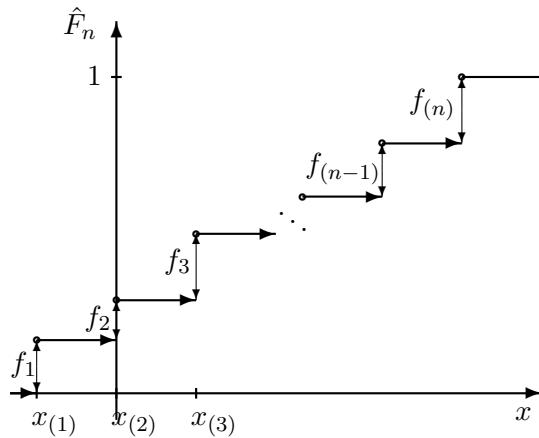
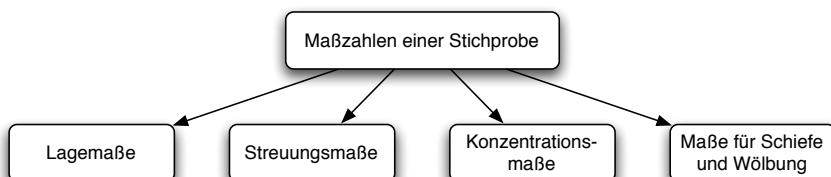


Abbildung 7.3: Eine typische empirische Verteilungsfunktion

tetige monoton nichtfallende Treppenfunktion ist, für die $\hat{F}_n(x) \xrightarrow{x \rightarrow -\infty} 0$, $\hat{F}_n(x) \xrightarrow{x \rightarrow \infty} 1$ gilt.

Übungsaufgabe 7.5 Zeigen Sie, dass $\hat{F}_n(x)$ eine Verteilungsfunktion ist.

7.6 Beschreibung von Verteilungen



Es sei eine konkrete Stichprobe (x_1, \dots, x_n) gegeben. Im Folgenden werden Kennzahlen (die sogenannten Maße) dieser Stichprobe betrachtet, welche die wesentlichen Aspekte der der Stichprobe zugrundeliegenden Verteilung wiedergeben:

1. Wo liegen die Werte x_i (Mittel, Ordnungsstatistiken, Quantile)? \Rightarrow Lagemaße
2. Wie stark streuen die Werte x_i (Varianz) \Rightarrow Streuungsmaße
3. Wie stark sind die Werte x_i in gewissen Bereichen von \mathbb{R} konzentriert \Rightarrow Konzentrationsmaße
4. Wie schief bzw. gewölbt ist die Verteilung von X \Rightarrow Maße für Schiefe und Wölbung

7.6.1 Lagemaße

Man unterscheidet folgende wichtige Lagemaße:

- Mittelwerte: Stichprobenmittel (arithmetisch), geometrisches und harmonisches Mittel, gewichtetes Mittel, getrimmtes Mittel
- Ordnungsstatistiken und Quantile, insbesondere Median und Quartile
- Modus

Betrachten wir sie der Reihe nach:

1. *Mittelwertbildung*: Seit der Antike kennt man mindestens 3 Arten der *Mittelberechnung* von n Zahlen (x_1, \dots, x_n) :

- *arithmetisch*: $\bar{x}_n = 1/n \sum_{i=1}^n x_i$, $\forall x_1, \dots, x_n \in \mathbb{R}$,
- *geometrisch*: $x_n^g = \sqrt[n]{x_1 \cdot \dots \cdot x_n}$, $x_1, \dots, x_n > 0$,
- *harmonisch*: $x_n^h = \left(1/n \sum_{i=1}^n x_i^{-1}\right)^{-1}$, $x_1, \dots, x_n \neq 0$.

- (a) Das *arithmetische Mittel* wird in der Statistik am meisten benutzt, weil es keine Voraussetzungen über den Wertebereich von x_1, \dots, x_n braucht. Es wird auch *Stichprobenmittel* genannt. Offensichtlich ist \bar{x}_n ein Spezialfall des sogenannten gewichteten Mittels $x_n^w = \sum_{i=1}^n w_i x_i$, wobei für die Gewichte $w_i \geq 0 \quad \forall i = 1, \dots, n$ und $\sum_{i=1}^n w_i = 1$ gilt. Als eine natürliche Gewichtewahl kommt $w_i = 1/n$, $\forall i = 1, \dots, n$ bei einer konkreten Stichprobe (x_1, \dots, x_n) in Frage. Die Summe aller Abweichungen von \bar{x}_n ist Null, denn $\sum_{i=1}^n (x_i - \bar{x}_n) = n\bar{x}_n - n\bar{x}_n = 0$, d.h. \bar{x}_n stellt geometrisch den Schwerpunkt der Werte x_i dar, falls jedem Punkt eine Einheitsmasse zugeordnet wird. Wenn es in der Stichprobe große Ausreißer gibt, so beeinflussen sie das Stichprobenmittel entscheidend und erschweren so die objektive Datenanalyse. Deshalb verwendet man oft die robuste Version des arithmetischen

Mittels, das sogenannte *getrimmte Mittel*:

$$\tilde{x}_n^{(k)} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)},$$

bei dessen Berechnung die k kleinsten und k größten Ausreißer ausgelassen werden, wobei $k \ll n/2$.

- (b) Das *geometrische Mittel* wird hauptsächlich bei der Beobachtung von Wachstums- und Zinsfaktoren verwendet. Sei $x_i = B_i/B_{i-1}$, $i = 1, \dots, n$ der Wachstumsfaktor des Merkmals B_i , das in den Jahren $i = 1, \dots, n$ beobachtet wurde (z.B. Inflationsfaktor). Dann ist $B_n = B_0 \cdot x_1 \cdot \dots \cdot x_n$ und somit wäre der Zins im Jahre n

$$B_n^g = B_0 \cdot x_1 \cdot \dots \cdot x_n = B_0 \cdot (x_n^g)^n.$$

Für das geometrische Mittel gilt

$$\log x_n^g = \frac{1}{n} \sum_{i=1}^n \log x_i \leq \log \left(\frac{1}{n} \sum_{i=1}^n x_i \right)$$

wegen der Konkavität des Logarithmus, d.h. $\log x_n^g = \overline{\log x_n} \leq \log \bar{x}_n$ und somit $x_n^g \leq \bar{x}_n$, wobei $x_n^g = \bar{x}_n$ genau dann, wenn $x_1 = \dots = x_n$.

- (c) Das *harmonische Mittel* wird bei der Ermittlung von z.B. durchschnittlicher Geschwindigkeiten gebraucht.

Beispiel 7.6 Seien x_i Geschwindigkeiten mit denen Bauteile eine Produktionslinie der Länge l durchlaufen. Die gesamte Bearbeitungszeit ist $l/x_1 + \dots + l/x_n$ und die Durchschnittslaufgeschwindigkeit

$$\frac{l + \dots + l}{l/x_1 + \dots + l/x_n} = x_n^h.$$

Es gilt $x_{(1)} \leq x_n^h \leq x_n^g \leq \bar{x}_n \leq x_{(n)}$ für $x_i > 0$, $i = 1, \dots, n$.

Übungsaufgabe 7.7 Beweisen Sie diese Relation per Induktion bzgl. n .

2. Ordnungsstatistiken und Quantile

Definition 7.8 Die *Ordnungsstatistiken* $x_{(i)}$, $i = 1, \dots, n$ der Stichprobe (x_1, \dots, x_n) sind durch die messbare Permutation $\varphi(x_1, \dots, x_n)$ gegeben, so dass

$$x_{(i)} = \min \{x_j : \#\{k : x_k \leq x_j\} \geq i\}, \quad \forall i = 1, \dots, n.$$

Somit gilt $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. Dieselbe Definition kann auch auf der Modellebene gegeben werden.

Definition 7.9

- (a) Sei nun X die Zufallsvariable, die das Merkmal modelliert. Sei F_X ihre Verteilungsfunktion. Die verallgemeinerte Inverse von F_X , definiert durch

$$F_X^{-1}(y) = \inf \{x : F_X(x) \geq y\}, \quad y \in [0, 1],$$

heißt *Quantilfunktion* von F_X bzw. X . Es gilt $F_X^{-1} : [0, 1] \rightarrow \mathbb{R} \cup \{\pm\infty\}$. Die Zahl $F_X^{-1}(\alpha)$, $\alpha \in [0, 1]$ wird α -Quantil von F_X genannt.

- (b) • $F_X^{-1}(0, 25)$ heißt *unteres Quartil*,
• $F_X^{-1}(0, 75)$ heißt *oberes Quartil*,
• $F_X^{-1}(0, 5)$ heißt der *Median* der Verteilung von X .

Zwischen Ordnungsstatistiken und Quantilen besteht ein enger Zusammenhang. So bedeutet $F_X^{-1}(\alpha)$, $\alpha \in (0, 1)$, dass ca. $\alpha \cdot 100\%$ aller Merkmalsausprägungen in der Stichprobe (x_1, \dots, x_n) unter $F_X^{-1}(\alpha)$ und ca. $(1 - \alpha) \cdot 100\%$ über $F_X^{-1}(\alpha)$ liegen (im absolut stetigen Fall). Insbesondere gilt $F_X^{-1}(\alpha) \approx x_{([n\alpha])}$, deshalb werden Ordnungsstatistiken auch *empirische Quantile* genannt. Dabei ist x_α definiert als

$$x_\alpha = \begin{cases} x_{([n\alpha]+1)}, & n\alpha \notin \mathbb{N} \\ 1/2(x_{([n\alpha])} + x_{([n\alpha]+1)}), & n\alpha \in \mathbb{N} \end{cases}.$$

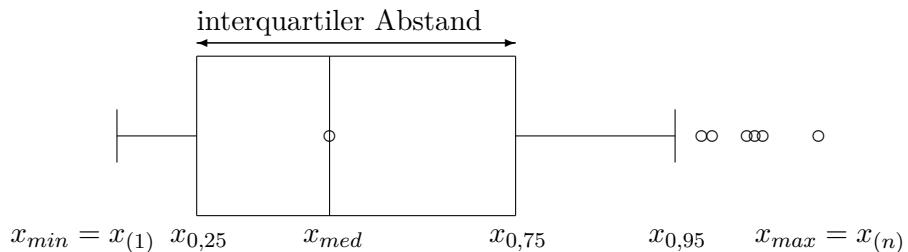
Dies ist die allgemeine Definition des *empirischen α -Quantils*.

Der *empirische Median* ist

$$x_{med} = \begin{cases} x_{(\frac{n+1}{2})}, & n \text{ ungerade} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), & n \text{ gerade.} \end{cases}$$

Somit sind mindestens 50% aller Stichprobenwerte kleiner gleich und 50% größer gleich x_{med} . Der Median ist ein Lagemaß, das ein robuster Ersatz für den Mittelwert darstellt, denn er ist bzgl. Ausreißern in der Stichprobe nicht sensibel.

Die oben genannten Statistiken werden in einem *Box-Plot* zusammengefasst und grafisch dargestellt:



Manchmal werden $x_{(1)}$ und $x_{(n)}$ durch $x_{0,05}$ und $x_{0,95}$ ersetzt. Die restlichen Werte werden darüber hinaus als Einzelpunkte auf der x -Achse abgebildet. Dann liegt ein sogenannter *modifizierter Box-Plot* vor.

3. *Modus:* Sei (x_1, \dots, x_n) eine Stichprobe, die aus n unabhängigen Realisierungen des Merkmals X besteht. Sei $(p(x))$ $f(x)$ die (Zähl-) Dichte von X , wobei die Verteilung von X unimodal ist.

Definition 7.10

- (a) Der Wert $x_{mod} = \operatorname{argmax} f(x)$ ($\operatorname{argmax} p(x)$) wird der *Modus der Verteilung von X* genannt (vgl. Abb. 7.4).
- (b) Empirisch wird \hat{x}_{mod} als $\frac{c_{m-1}+c_m}{2}$ für $m = \operatorname{argmax} f_i$ definiert, also als die Mitte des Intervalls mit der größten Häufigkeit des Vorkommens in der Stichprobe, falls dieser eindeutig bestimmbar ist.

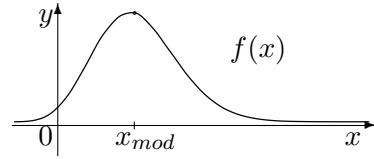


Abbildung 7.4: Veranschaulichung des Modus

Den Mittelwert \bar{x}_n , Median x_{med} und Modus x_{mod} kann man auch wie folgt definieren:

$$\bar{x}_n = \operatorname{argmin}_{x \in \mathbb{R}} \sum_{i=1}^n (x_i - x)^2$$

$$x_{med} = \operatorname{argmin}_{x \in \mathbb{R}} \sum_{i=1}^n |x_i - x|$$

$$\hat{x}_{mod} = \frac{c_{m-1}+c_m}{2}, \quad \text{wobei } m = \operatorname{argmin}_{j=1, \dots, n} \sum_{i=1}^n I(x_i \notin (c_{j-1}, c_j])$$

Übungsaufgabe 7.11 Zeigen Sie die Äquivalenz der oben genannten Definitionen des Mittelwerts \bar{x}_n , Medians x_{med} und des Modus x_{mod} zu den bekannten Definitionen.

Die Größen \bar{x}_n , x_{med} und \hat{x}_{mod} können auch zur Beschreibung der Symmetrie einer unimodalen Verteilung F_X von Daten (x_1, \dots, x_n) verwendet werden, da

- bei symmetrischen Verteilung F_X gilt $\bar{x}_n \approx x_{med} \approx \hat{x}_{mod}$
- bei linkssteilen Verteilung F_X gilt $\hat{x}_{mod} < x_{med} < \bar{x}_n$
- bei rechtssteilen Verteilung F_X gilt $\bar{x}_n < x_{med} < \hat{x}_{mod}$.

7.6.2 Streuungsmaße

Bekannte Streuungsmaße einer konkreten Stichprobe (x_1, \dots, x_n) sind die folgenden Größen:

- *Spannweite* $x_{(n)} - x_{(1)}$,
- *empirische Varianz* $\bar{s}_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$,
- *Stichprobenvarianz* $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{n}{n-1} \bar{s}_n^2$,
- *empirische Standardabweichungen* $\bar{s}_n = \sqrt{\bar{s}_n^2}$, $s_n = \sqrt{s_n^2}$,
- *empirischer Variationskoeffizient* $\gamma_n = s_n / \bar{x}_n$, falls $\bar{x}_n > 0$.

Die Spannweite zeigt die *maximale Streuung* in den Daten, wobei sich die empirische Varianz mit der *mittleren quadratischen Abweichung* vom Stichprobenmittel auseinandersetzt. Hier sind einige Eigenschaften von \bar{s}_n^2 (bzw. s_n^2 , da sie sich nur durch einen Faktor unterscheiden):

Lemma 7.12

1. Für jedes $b \in \mathbb{R}$ gilt

$$\sum_{i=1}^n (x_i - b)^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(\bar{x}_n - b)^2$$

und somit für $b = 0$

$$\bar{s}_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i^2 - \bar{x}_n^2) \quad \text{bzw.} \quad s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i^2 - \bar{x}_n^2) .$$

2. *Transformationsregel:*

Falls die Daten (x_1, \dots, x_n) linear transformiert werden, d.h. $y_i = ax_i + b$, $a \neq 0$, $b \in \mathbb{R}$, dann gilt

$$\bar{s}_{n,y}^2 = a^2 \bar{s}_{n,x}^2 \quad \text{bzw.} \quad \bar{s}_{n,y} = |a| \bar{s}_{n,x} ,$$

wobei

$$\bar{s}_{n,y}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_n)^2 , \quad \bar{s}_{n,x}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

Beweis 1. Es gilt:

$$\begin{aligned}
 \sum_{i=1}^n (x_i - b)^2 &= \sum_{i=1}^n (x_i - \bar{x}_n + \bar{x}_n - b)^2 \\
 &= \sum_{i=1}^n (x_i - \bar{x}_n)^2 + 2 \sum_{i=1}^n (x_i - \bar{x}_n) \cdot (\bar{x}_n - b) + \sum_{i=1}^n (\bar{x}_n - b)^2 \\
 &= \sum_{i=1}^n (x_i - \bar{x}_n)^2 + 2(\bar{x}_n - b) \cdot \underbrace{\sum_{i=1}^n (x_i - \bar{x}_n)}_{=0} + n(\bar{x}_n - b)^2, \quad \forall b \in \mathbb{R}.
 \end{aligned}$$

2. Es gilt:

$$\bar{s}_{n,y}^2 = \frac{1}{n} \sum_{i=1}^n (ax_i + b - a\bar{x}_n - b)^2 = \frac{a^2}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = a^2 \bar{s}_{n,x}^2.$$

□

Der Skalierungsunterschied zwischen \bar{s}_n^2 und s_n^2 ist den Eigenschaften der *Erwartungstreue* von s_n^2 zu verdanken, die später im Laufe dieser Vorlesung behandelt wird, und besagt, dass für eine Zufallsstichprobe (X_1, \dots, X_n) mit X_i unabhängig identisch verteilt, $X_i \sim X$, $\text{Var } X = \sigma^2 \in (0, \infty)$ gilt $E s_n^2 = \sigma^2$, wobei $E \bar{s}_n^2 = \frac{n}{n-1} \sigma^2 \xrightarrow{n \rightarrow \infty} \sigma^2$. Das heißt, während bei der Verwendung von s_n^2 zur Schätzung von σ^2 kein Fehler „im Mittel“ gemacht wird, ist diese Aussage für \bar{s}_n^2 nur asymptotisch (für große Datenmengen n) richtig.

Aufgrund von $\sum_{i=1}^n (x_i - \bar{x}_n) = 0$ ist z.B. $x_n - \bar{x}_n$ durch $x_i - \bar{x}_n$, $i = 1, \dots, n-1$ bestimmt. Somit verringert sich die *Anzahl der Freiheitsgrade* in der Summe $\sum_{i=1}^n (x_i - \bar{x}_n)^2$ um 1 und somit scheint die Normierung $\frac{1}{n-1}$ plausibel zu sein.

Die *Standardabweichungen* \bar{s}_n und s_n werden verwendet, damit man dieselben Einheiten (und nicht ihre Quadrate, also z.B. Euro und nicht Euro²) erhält. Für normalverteilte Stichproben ($X \sim N(\mu, \sigma^2)$) liefert \bar{s}_n auch die „ k -Sigma-Regel“ (vgl. Vorlesung WR), die besagt, dass in den Intervallen

$$\begin{aligned}
 [\bar{x}_n - \bar{s}_n, \bar{x}_n + \bar{s}_n] &\quad \text{ca.} \quad 68\%, \\
 [\bar{x}_n - 2\bar{s}_n, \bar{x}_n + 2\bar{s}_n] &\quad \text{ca.} \quad 95\%, \\
 [\bar{x}_n - 3\bar{s}_n, \bar{x}_n + 3\bar{s}_n] &\quad \text{ca.} \quad 99\%
 \end{aligned}$$

aller Daten liegen.

Der Vorteil vom *empirischen Variationskoeffizienten* ist, dass er *maßstabsunabhängig* ist und somit den Vergleich von Streuungseigenschaften unterschiedlicher Stichproben zulässt.

7.6.3 Maße für Schiefe und Wölbung

Im Vorlesungsskript WR, Abschnitt 4.5 S. 99 wurden folgende Maße für Schiefe bzw. Wölbung der Verteilung einer Zufallsvariable X eingeführt:
Schiefe oder Symmetriekoeffizient:

$$\gamma_1 = \frac{\mu'_3}{\sigma^3} = E(\tilde{X}^3),$$

wobei

$$\mu'_k = E(X - EX)^k, \quad \sigma^2 = \mu'_2 = \text{Var } X, \quad \tilde{X} = \frac{X - EX}{\sigma}.$$

Wölbung (Exzess):

$$\gamma_2 = \frac{\mu'_4}{\sigma^4} - 3 = E(\tilde{X}^4) - 3,$$

vorausgesetzt, dass $E(X^4) < \infty$. Für ihre Bedeutung und Interpretation siehe die oben genannten Seiten des WR-Vorlesungsskriptes. Falls nun das Merkmal X statistisch in einer Stichprobe (x_1, \dots, x_n) beobachtet wird, wie können γ_1 und γ_2 aus diesen Daten geschätzt und interpretiert werden?

Als Schätzer für das k -te zentrierte Moment $\mu'_k = E(X - EX)^k$, $k \in \mathbb{N}$ schlagen wir

$$\hat{\mu}'_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^k$$

vor, die Varianz σ^2 wird durch

$$\hat{s}_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

geschätzt. Somit bekommt man den Momentenkoeffizient an der Schiefe (engl. „skewness“)

$$\hat{\gamma}_1 = \frac{\hat{\mu}'_3}{\hat{s}_n^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \right)^{3/2}}.$$

Falls die Verteilung von X linksschief ist, überwiegen positive Abweichungen im Zähler und somit gilt $\hat{\gamma}_1 > 0$ für linksschiefe Verteilungen. Analog gilt $\hat{\gamma}_1 \approx 0$ für symmetrische und $\hat{\gamma}_1 < 0$ für rechtsschiefe Verteilungen.

Das *Wölbungsmaß von Fisher* (engl. „kurtosis“) ist gegeben durch

$$\hat{\gamma}_2 = \frac{\hat{\mu}'_4}{\hat{s}_n^4} - 3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \right)^2} - 3.$$

Falls $\hat{\gamma}_2 > 0$ so ist die Verteilung von X steilgipflig, für $\hat{\gamma}_2 < 0$ ist sie flachgipflig. Falls $X \sim N(\mu, \sigma^2)$, so gilt $\hat{\gamma}_2 \approx 0$. Die Ursache dafür ist, dass

die steilgipfligen Verteilungen schwerere Tails haben als die flachgipfligen. Als Maß dient dabei die Normalverteilung, für die $\gamma_1 = \gamma_2 = 0$ und somit $\hat{\gamma}_1 \approx 0$, $\hat{\gamma}_2 \approx 0$. So definiert, sind $\hat{\gamma}_1$ und $\hat{\gamma}_2$ nicht resistent gegenüber Ausreisern. Eine robuste Variante von $\hat{\gamma}_1$ ist beispielsweise durch den sogenannten *Quantilskoeffizienten der Schiefe*

$$\hat{\gamma}_q(\alpha) = \frac{(x_{1-\alpha} - x_{med}) - (x_{med} - x_\alpha)}{x_{1-\alpha} - x_\alpha}, \quad \alpha \in (0, 1/2)$$

gegeben.

Für $\alpha = 0,25$ erhält man den Quantilskoeffizienten. $\hat{\gamma}_q(\alpha)$ misst den Unterschied zwischen der Entfernung des α - und $(1 - \alpha)$ -Quantils zum Median. Bei linkssteilen (bzw. rechtssteilen) Verteilungen liegt das (untere) x_α -Quantil näher an (bzw. weiter entfernt von) dem Median. Somit gilt

- $\hat{\gamma}_q(\alpha) > 0$ für linkssteile Verteilungen,
- $\hat{\gamma}_q(\alpha) < 0$ für rechtssteile Verteilungen,
- $\hat{\gamma}_q(\alpha) = 0$ für symmetrische Verteilungen.

Durch das zusätzliche Normieren (Nenner) gilt $-1 \leq \hat{\gamma}_q(\alpha) \leq 1$.

7.7 Quantilplots (Quantil-Grafiken)

Nach der ersten beschreibenden Analyse eines Datensatzes (x_1, \dots, x_n) soll überlegt werden, mit welcher Verteilung diese Stichprobe modelliert werden kann. Hier sind die sogenannten *Quantilplots* behilflich, da sie grafisch zeigen, wie gut die Daten (x_1, \dots, x_n) mit dem Verteilungsgesetz G übereinstimmen, wobei G die Verteilungsfunktion einer hypothetischen Verteilung ist.

Sei X eine Zufallsvariable mit (unbekannter) Verteilungsfunktion F_X . Auf Basis der Daten (X_1, \dots, X_n) , X_i unabhängig identisch verteilt und $X_i \stackrel{d}{=} X$ möchte man prüfen, ob $F_X = G$ für eine bekannte Verteilungsfunktion G gilt. Die Methode der *Quantil-Grafiken* besteht darin, dass man die entsprechenden Quantil-Funktionen \hat{F}_n^{-1} und G^{-1} von \hat{F}_n und G grafisch vergleicht. Hierzu

- plotte man $G^{-1}(k/n)$ gegen $\hat{F}_n^{-1}(k/n) = X_{(k)}$, $k = 1, \dots, n$.
- Falls die Punktwolke

$$\left\{ \left(G^{-1}(k/n), X_{(k)} \right), \quad k = 1, \dots, n \right\}$$

näherungsweise auf einer Geraden $y = ax + b$ liegt, so sagt man, dass $F_X(x) \approx G\left(\frac{x-a}{b}\right)$, $x \in \mathbb{R}$.

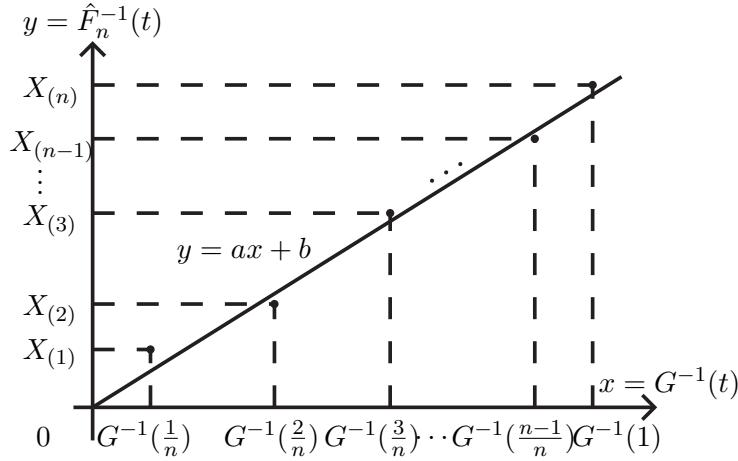


Abbildung 7.5: Quantil-Grafik

Diese empirische Vergleichsmethode beruht auf folgenden Überlegungen:

- Man ersetzt die unbekannte Funktion F_X durch die aus den Daten berechenbare Funktion \hat{F}_n . Dabei macht man einen Fehler, der allerdings asymptotisch (für $n \rightarrow \infty$) klein ist. Dies folgt aus dem Satz 8.42 (WR-Skript) von Gliwenko-Cantelli, der besagt, dass

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_X(x)| \xrightarrow{n \rightarrow \infty} 0.$$

Der Vergleich der entsprechenden Quantil-Funktionen wird durch folgendes Ergebnis bestärkt: Falls $EX < \infty$, dann gilt

$$\sup_{t \in [0,1]} \left| \int_0^t (\hat{F}_n^{-1}(y) - F_X^{-1}(y)) dy \right| \xrightarrow[n \rightarrow \infty]{\text{f.s.}} 0.$$

Somit setzt man bei der Verwendung der Quantil-Grafiken voraus, dass der Stichprobenumfang n ausreichend groß ist, um $\hat{F}_n^{-1} \approx F_X^{-1}$ zu gewährleisten.

- Man setzt zusätzlich voraus, dass die Gleichungen

$$\begin{aligned} y &= ax + b, \\ y &= F_X^{-1}(t), \\ x &= G^{-1}(t) \end{aligned}$$

für alle t (und nicht nur näherungsweise für $t = k/n$, $k = 1, \dots, n$) gelten. Daraus folgt, dass $G(x) = t = F_X(y) = F_X(ax + b)$ für alle x , oder $F_X(y) = G\left(\frac{y-b}{a}\right)$ für alle y , weil $x = \frac{y-b}{a}$ ist.

Aus praktischer Sicht ist es besser, Paare $\left(G^{-1}\left(\frac{k}{n+1}\right), X_{(k)}\right)$, $k = 1, \dots, n$ zu plotten. Dadurch wird vermieden, dass $G^{-1}(n/n) = G^{-1}(1) = \infty$ vorkommt, wie es zum Beispiel bei einer Verteilung G der Fall ist, bei der $F(x) < 1$ gilt für alle $x \in \mathbb{R}$. Tatsächlich gilt für $k = n$, dass $\frac{n}{n+1} < 1$ und somit $G^{-1}\left(\frac{n}{n+1}\right) < \infty$.

Beispiel 7.13 (Exponential-Verteilung, $G(x) = (1 - e^{-\lambda x}) \cdot I(x \geq 0)$) Es gilt $G^{-1}(y) = -1/\lambda \log(1 - y)$, $y \in (0, 1)$. So wird man beim Quantil-Plot Paare

$$\left(-\frac{1}{\lambda} \log\left(1 - \frac{k}{n+1}\right), X_{(k)}\right), \quad k = 1, \dots, n$$

zeichnen, wobei der Faktor $1/\lambda$ für die Linearität unwesentlich ist und weggelassen werden kann.

Beispiel 7.14 (Normalverteilung, $G(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$, $x \in \mathbb{R}$) Leider ist die analytische Berechnung von Φ^{-1} mit einer geschlossenen Formel nicht möglich. Aus diesem Grund wird $\Phi^{-1}\left(\frac{k}{n+1}\right)$ numerisch berechnet und in Tabellen oder statistischen Software-Paketen (wie z.B. R) abgelegt. Um die empirische Verteilung der Daten mit der Normalverteilung zu vergleichen, trägt man Punkte mit Koordinaten

$$\left(\Phi^{-1}\left(\frac{k}{n+1}\right), X_{(k)}\right), \quad k = 1, \dots, n$$

auf der Ebene auf und prüft, ob sie eine Gerade bilden (vgl. Abb. 7.6).

Bemerkung 7.15 Falls $\bar{x}_n = 0$ und die Verteilung F_X linkssteil ist, so sind die Quantile von F_X kleiner als die von Φ . Somit ist der Normal-Quantilplot konkav. Falls $\bar{x}_n = 0$ und F_X rechtssteil ist, so wird der Normal-Quantilplot konvex sein.

Beispiel 7.16 (Haftpflichtversicherung (Belgien, 1992)) In Abbildung 7.7 sind Ordnungsstatistiken der Stichprobe von $n = 227$ Schadenhöhen der Industrie-Unfälle in Belgien im Jahr 1992 (Haftpflichtversicherung) gegen Quantile von Exponential-, Pareto-, Standardnormal- und Weibull-Verteilungen geplottet. Im Bereich von Kleinschäden zeigen die Exponential- und Pareto-Verteilungen eine gute Übereinstimmung mit den Daten. Die Verteilung von mittelgroßen Schäden kann am besten durch die Lognormal- und Weibull-Verteilungen modelliert werden. Für Großschäden erweist sich die Weibull-Verteilung als geeignet.

Beispiel 7.17 (Rendite der BMW-Aktie) In Abbildung 7.8 ist der Quantilplot für Renditen der BMW-Aktie beispielhaft zu sehen.

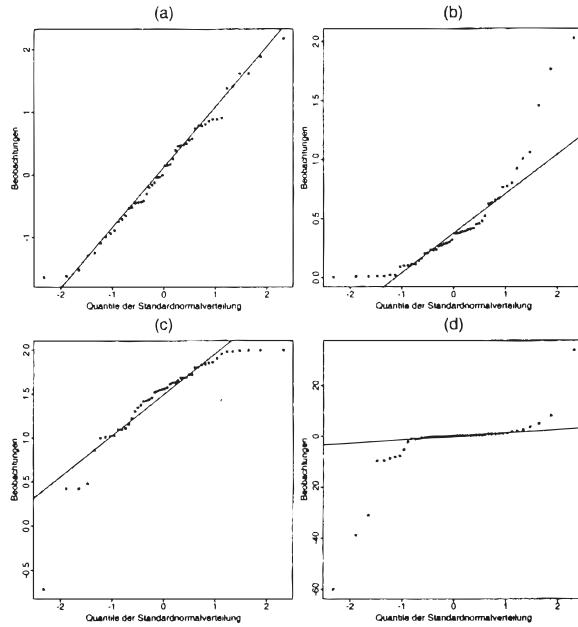


Abbildung 7.6: QQ-Plot einer Normalverteilung (a), einer linkssteilen Verteilung (b), einer rechtssteilen Verteilung (c) und einer symmetrischen, aber stark gekrümmten Verteilung (d)

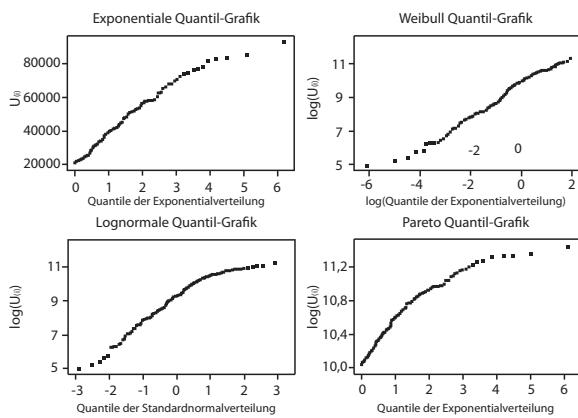


Abbildung 7.7: Ordnungsstatistiken einer Stichprobe von Schadenhöhen der Industrie-Unfälle in Belgien im Jahr 1992

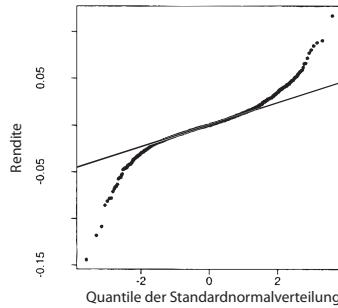


Abbildung 7.8: Quantilplot der Rendite der BMW-Aktie

7.8 Dichteschätzung

Sei eine Stichprobe (x_1, \dots, x_n) von unabhängigen Realisierungen eines absolut stetig verteilten Merkmals X mit Dichte f_X gegeben. Mit Hilfe der in Abschnitt ?? eingeführten Histogramme lässt sich f_X grafisch durch eine Treppenfunktion \hat{f}_X darstellen. Dabei gibt es zwei entscheidende Nachteile der Histogrammdarstellung:

1. Willkür in der Wahl der Klasseneinteilung $[c_{i-1}, c_i]$,
2. Eine (möglicherweise) stetige Funktion f_X wird durch eine Treppenfunktion \hat{f}_X ersetzt.

In diesem Abschnitt werden wir versuchen, diese Nachteile zu beseitigen, indem wir eine Klasse von Kerndichtenschätzern einführen, die (je nach Wahl des Kerns) auch zu stetigen Schätzern \hat{f}_X führen.

Definition 7.18 Der Kern $K(x)$ wird definiert als eine nicht-negative messbare Funktion auf \mathbb{R} mit der Eigenschaft $\int_{\mathbb{R}} K(x) dx = 1$.

Definition 7.19 Der *Kerndichteschätzer* der Dichte f_X aus den Daten (x_1, \dots, x_n) mit Kernfunktion $K(x)$ ist gegeben durch

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right), \quad x \in \mathbb{R},$$

wobei $h > 0$ die sogenannte *Bandbreite* ist.

Beispiele für Kerne:

1. *Rechteckskern:*

$$K(x) = 1/2 \cdot I(x \in [-1, 1]).$$

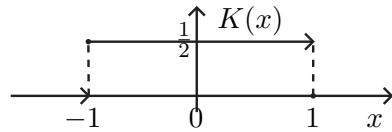
Dabei ist

$$\frac{1}{h} K\left(\frac{x - x_i}{h}\right) = \begin{cases} 1/(2h), & x_i - h \leq x < x_i + h, \\ 0, & \text{sonst,} \end{cases}$$

und somit

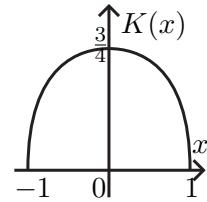
$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^k K\left(\frac{x - x_i}{h}\right) = \frac{\#\{x_i \in [x-h, x+h)\}}{2nh},$$

das auch *gleitendes Histogramm* genannt wird. Dieser Dichteschätzer ist (noch) nicht stetig, was durch die (besonders einfache rechteckige unstetige) Form des Kerns erklärt wird.



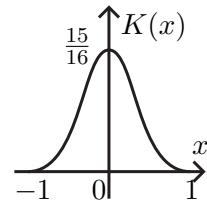
2. *Epanechnikov-Kern:*

$$K(x) = \begin{cases} 3/4(1-x^2), & x \in [-1, 1] \\ 0, & \text{sonst.} \end{cases}$$



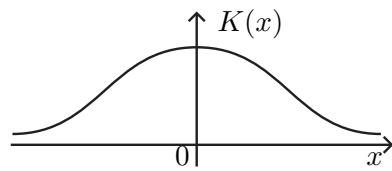
3. *Bisquare-Kern:*

$$K(x) = \frac{15}{16} ((1-x^2)^2 \cdot I(x \in [-1, 1])) .$$



4. *Gauss-Kern:*

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}.$$



Dabei ist die Wahl der Bandbreite h entscheidend für die Qualität der Schätzung. Je größer $h > 0$, desto glatter wird \hat{f}_X sein und desto mehr „Details“ werden „herausgemittelt“. Für kleinere h wird \hat{f}_X rauer. Dabei können aber auch Details auftreten, die rein stochastischer Natur sind und keine Gesetzmäßigkeiten zeigen. Mit der adäquaten Wahl von h beschäftigen sich viele wissenschaftliche Arbeiten, die empirische Faustregeln, aber auch kompliziertere Optimierungsmethoden dafür vorschlagen. Insgesamt ist das Problem der optimalen Dichteschätzung in der Statistik immer noch offen.

7.9 Beschreibung und Exploration von bivariaten Datensätzen

Im Gegensatz zu der Datenlage in den Abschnitten 7.5 bis 7.8 betrachten wir im Folgenden Datensätze bestehend aus 2 Stichproben (x_1, \dots, x_n) und (y_1, \dots, y_n) , die als Realisierungen von stochastischen Stichproben (X_1, \dots, X_n) und (Y_1, \dots, Y_n) aufgefasst werden, wobei X_1, \dots, X_n unabhängige identisch verteilte Zufallsvariablen mit $X_i \stackrel{d}{=} X \sim F_X$, Y_1, \dots, Y_n unabhängige identisch verteilte Zufallsvariablen mit $Y_i \stackrel{d}{=} Y \sim F_Y$ sind. Wir betrachten hier ausschließlich quantitative Merkmale X und Y . Es wird ein Zusammenhang zwischen X und Y vermutet, der an Hand von (konkreten) Stichproben (x_1, \dots, x_n) und (y_1, \dots, y_n) näher untersucht werden soll. Mit anderen Worten, wir interessieren uns für die Eigenschaften der bivariaten Verteilung $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$ des Zufallsvektors $(X, Y)^T$.

7.9.1 Zusammenhangsmaße

Jetzt wird uns die Frage beschäftigen, in welchem Maße die Merkmale X und Y voneinander abhängig sind. Um die $\text{Cov}(X, Y) = E(X - EX)(Y - EY)$ aus den Daten zu schätzen, setzt man die sogenannte *empirische Kovarianz*

$$S_{xy}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)$$

ein. Dabei ist S_{xy}^2 jedoch von den Skalen von X und Y abhängig.

- Um eine skaleninvariantes Zusammenhangsmaß zu bekommen, betrachtet man die empirische Variante des Korrelationskoeffizienten

$$\varrho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var } X} \cdot \sqrt{\text{Var } Y}},$$

den sogenannten *Bravais-Pearson-Korrelationskoeffizienten*

$$\varrho_{xy} = \frac{S_{xy}^2}{\sqrt{S_{xx}^2 \cdot S_{yy}^2}},$$

wobei

$$S_{xx}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2, \quad S_{yy}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_n)^2$$

die Stichprobenvarianzen der Stichproben (x_1, \dots, x_n) und (y_1, \dots, y_n) sind. Dabei erbt ρ_{xy} alle Eigenschaften des Korrelationskoeffizienten $\rho(X, Y)$:

- (a) $|\rho_{xy}| \leq 1$
- (b) $\rho_{xy} = \pm 1$, falls ein linearer Zusammenhang in den Daten $(x_i, y_i)_{i=1, \dots, n}$ vorliegt, d.h. alle Punkte (x_i, y_i) , $i = 1, \dots, n$ liegen auf einer Geraden mit positivem (bei $\rho_{xy} = 1$) bzw. negativem (bei $\rho_{xy} = -1$) Anstieg.
- (c) Wenn $|\rho_{xy}|$ klein ist ($\rho_{xy} \approx 0$), so sind die Datensätze unkorreliert. Dabei wird oft folgende grobe Einteilung vorgenommen:
Merkmale X und Y sind
 - „schwach korreliert“, falls $|\rho_{xy}| < 0.5$,
 - „stark korreliert“, falls $|\rho_{xy}| \geq 0.8$.

Ansonsten liegt ein mittlerer Zusammenhang zwischen X und Y vor.

Lemma 7.20 Für ρ_{xy} gilt die alternative rechengünstige Darstellung

$$\rho_{xy} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x}_n \bar{y}_n}{\sqrt{(\sum_{i=1}^n x_i^2 - n \bar{x}_n^2)(\sum_{i=1}^n y_i^2 - n \bar{y}_n^2)}}. \quad (7.1)$$

2. Spearmans Korrelationskoeffizient

Einen alternativen Korrelationskoeffizienten erhält man, wenn man die Stichprobenwerte x_i bzw. y_i in ρ_{xy} durch ihre Ränge $rg(x_i)$ bzw. $rg(y_i)$ ersetzt, die als Position dieser Werte in den ansteigend geordneten Stichproben zu verstehen sind:

$rg(x_i) = j$, falls $x_i = x_{(j)}$ für ein $j \in \{1, \dots, n\}$, $\forall i = 1, \dots, n$. Es bedeutet, dass $rg(x_{(i)}) = i$ $\forall i = 1, \dots, n$, falls $x_i \neq x_j$ für $i \neq j$.

Falls die Stichprobe (x_1, \dots, x_n) k identische Werte x_i (die sogenannten *Bindungen*) enthält, so wird diesen Werten der sogenannte Durchschnittsrang $rg(x_i)$ zugewiesen, der als arithmetisches Mittel der k in Frage kommenden Ränge errechnet wird. Zum Beispiel findet folgende Zuordnung statt:

x_i	(3, 1, 7, 5, 3, 3)
$rg(x_i)$	(a, 1, 6, 5, a, a)

wobei der Durchschnittsrang a von Stichprobeneintrag 3 gleich $a = \frac{1}{3}(2 + 3 + 4) = 3$ ist.

Somit wird der sogenannte *Spearmans Korrelationskoeffizient* (Rangkorrelationskoeffizient) der Stichproben

$$(x_1, \dots, x_n) \quad \text{und} \quad (y_1, \dots, y_n)$$

als der *Bravais-Pearson-Koeffizient* der Stichproben ihrer Ränge

$$(\text{rg}(x_1), \dots, \text{rg}(x_n)) \quad \text{und} \quad (\text{rg}(y_1), \dots, \text{rg}(y_n))$$

definiert:

$$\varrho_{sp} = \frac{\sum_{i=1}^n (\text{rg}(x_i) - \bar{\text{rg}}_x)(\text{rg}(y_i) - \bar{\text{rg}}_y)}{\sqrt{\sum_{i=1}^n (\text{rg}(x_i) - \bar{\text{rg}}_x)^2 \sum_{i=1}^n (\text{rg}(y_i) - \bar{\text{rg}}_y)^2}},$$

wobei

$$\begin{aligned}\bar{\text{rg}}_x &= \frac{1}{n} \sum_{i=1}^n \text{rg}(x_i) = \frac{1}{n} \sum_{i=1}^n \text{rg}(x_{(i)}) = \frac{1}{n} \sum_{i=1}^n i = \frac{n(n+1)}{2n} = \frac{n+1}{2}, \\ \bar{\text{rg}}_y &= \frac{1}{n} \sum_{i=1}^n \text{rg}(y_i) = \frac{n+1}{2}.\end{aligned}$$

Dieselbe Darstellung $\bar{\text{rg}}_y$ gilt auch, wenn Bindungen vorhanden sind.

Dieser Koeffizient misst monotone Zusammenhänge in den Daten. Aus den Eigenschaften der Bravais-Pearson-Koeffizienten folgt $|\varrho_{sp}| \leq 1$. Betrachten wir die Fälle $\varrho_{sp} = \pm 1$ gesondert:

- $\varrho_{sp} = 1$ bedeutet, dass die Punkte $(\text{rg}(x_i), \text{rg}(y_i))$, $i = 1, \dots, n$ auf einer Geraden mit positiver Steigung liegen. Da aber $\text{rg}(x_i), \text{rg}(y_i) \in \mathbb{N}$, kann diese Steigung nur 1 sein. Es bedeutet, dass dem kleinsten Wert in der Stichprobe (x_1, \dots, x_n) der kleinste Wert in (y_1, \dots, y_n) entspricht, usw., d.h., für wachsende x_i wachsen auch die y_i streng monoton: $x_i < x_j \implies y_i < y_j \quad \forall i \neq j$.
- Analog gilt dann für $\varrho_{sp} = -1$, dass $x_i < x_j \implies y_i > y_j \quad \forall i \neq j$.

Dies kann folgendermaßen zusammengefaßt werden:

- $\varrho_{sp} > 0$: gleichsinniger monotoner Zusammenhang (x_i groß \iff y_i groß)
- $\varrho_{sp} < 0$: gegensinniger monotoner Zusammenhang (x_i groß \iff y_i klein)
- $\varrho_{sp} \approx 0$: kein monotoner Zusammenhang.

Da der Spearmans Korrelationskoeffizient nur Ränge von x_i und y_i betrachtet, eignet er sich auch für ordinale (und nicht nur quantitative) Daten.

Lemma 7.21 Falls die Stichproben (x_1, \dots, x_n) und (y_1, \dots, y_n) keine Bindung enthalten ($x_i \neq x_j, y_i \neq y_j \forall i \neq j$), dann gilt

$$\varrho_{sp} = 1 - \frac{6}{(n^2 - 1)n} \sum_{i=1}^n d_i^2,$$

wobei $d_i = \text{rg}(x_i) - \text{rg}(y_i) \quad \forall i = 1, \dots, n$.

Beweis Als Übungsaufgabe. □

Satz 7.22 (Invarianzeigenschaften)

1. Wenn die Merkmale X und Y linear transformiert werden:

$$\begin{aligned} f(X) &= a_x X + b_x, \quad a_x \neq 0, b_x \in \mathbb{R}, \\ g(Y) &= a_y Y + b_y, \quad a_y \neq 0, b_y \in \mathbb{R}, \end{aligned}$$

dann gilt $\varrho_{f(x)g(y)} = \text{sgn } (a_x a_y) \cdot \varrho_{xy}$.

2. Falls Funktionen $f : \mathbb{R} \rightarrow \mathbb{R}$ und $g : \mathbb{R} \rightarrow \mathbb{R}$ beide monoton wachsend oder beide monoton fallend sind, dann gilt

$$\varrho_{sp}(f(x), g(y)) = \varrho_{sp}(x, y).$$

Falls f monoton wachsend und g monoton fallend (oder umgekehrt) sind, dann gilt $\varrho_{sp}(f(x), g(y)) = -\varrho_{sp}(x, y)$.

Beweis Beweisen wir nur 1), weil 2) offensichtlich ist.

- 1.

$$\begin{aligned} \varrho_{f(x)g(y)} &= \frac{\sum_{i=1}^n ((a_x x_i + b_x) - (a_x \bar{x}_n + b_x))((a_y y_i + b_y) - (a_y \bar{y}_n + b_y))}{\sqrt{a_x^2 \sum_{i=1}^n (x_i - \bar{x}_n)^2} \sqrt{a_y^2 \sum_{i=1}^n (y_i - \bar{y}_n)^2}} \\ &= \frac{a_x a_y}{|a_x| |a_y|} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y}_n)^2}} = \text{sgn } (a_x a_y) \cdot \varrho_{xy}. \end{aligned}$$

□

Bemerkung 7.23

1. Da lineare Transformationen monoton sind, gilt Aussage 1) auch für Spearmans Korrelationskoeffizienten ϱ_{sp} .
2. Der Koeffizient ϱ_{xy} erfasst lineare Zusammenhänge, während ϱ_{sp} monotone Zusammenhänge aufspürt.

7.9.2 Einfache lineare Regression

Wenn man den Zusammenhang von Merkmalen X und Y mit Hilfe von Streudiagrammen visualisiert, wird oft ein linearer Trend erkennbar, obwohl der Bravais-Pearson-Korrelationskoeffizient einen Wert kleiner als 1 liefert, z.B. $\rho_{xy} \approx 0,6$ (vgl. Abb. 7.9). Dies ist der Fall, weil die Datenpunkte (x_i, y_i) ,

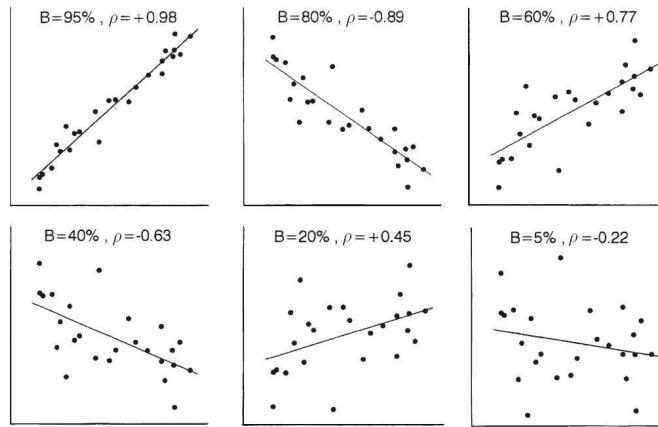


Abbildung 7.9: Vergleich verschiedenwertiger Bestimmtheitsmaße. Es sind Regressionsgerade, Bestimmtheitsmaß B und Korrelationskoeffizient ρ verschiedener (fiktiver) Punktwolken vom Umfang $n = 25$ dargestellt. Die Beschriftung der Achsen ist weggelassen, weil sie hier ohne Bedeutung ist.

$i = 1, \dots, n$ oft um eine Gerade streuen und nicht exakt auf einer Geraden liegen. Um solche Situationen stochastisch modellieren zu können, nimmt man den Zusammenhang der Form

$$Y = f(X) + \varepsilon$$

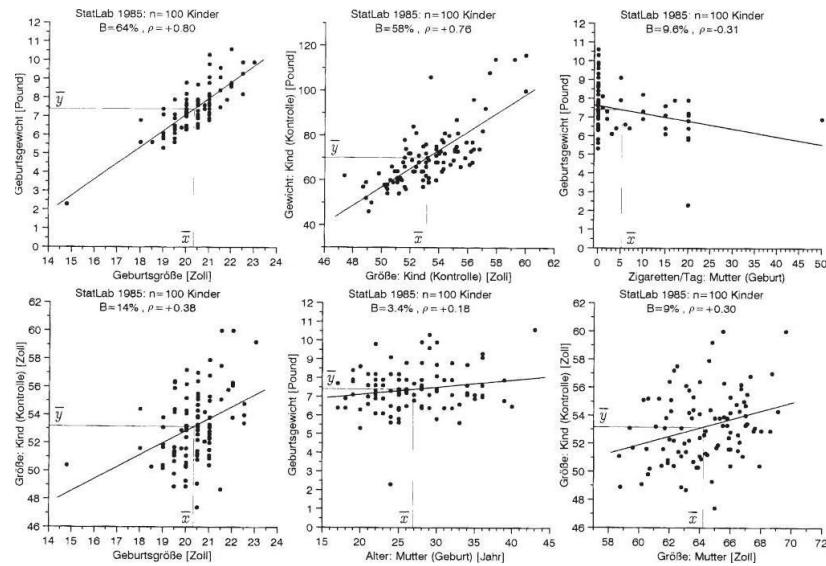
an, wobei ε die sogenannte Störgröße ist, die auf mehrere Ursachen wie z.B. Beobachtungsfehler (Messfehler, Berechnungsfehler, usw.) zurückzuführen sein kann. Dabei nennt man die Zufallsvariable Y Zielgröße oder *Regressand*, die Zufallsvariable X Einflussfaktor, *Regressor* oder *Ausgangsvariable*. Der Zusammenhang $Y = f(X) + \varepsilon$ wird *Regression* genannt, wobei man oft über ε voraussetzt, dass $E\varepsilon = 0$ (kein systematischer Beobachtungsfehler). Wenn $f(x) = \alpha + \beta x$ eine lineare Funktion ist, so spricht man von der *einfachen linearen Regression*. Es sind aber durchaus andere Arten der Zusammenhänge denkbar, wie z.B.

$$f(x) = \sum_{i=0}^n \alpha_i x^i$$

X	Y
Geschwindigkeit	Länge des Bremswegs
Körpergröße des Vaters	Körpergröße des Sohnes
Produktionsfaktor	Qualität des Produktes
Spraydosen-Verbrauch	Ozongehalt der Atmosphäre
Noten im Bachelor-Studium	Noten im Master-Studium

Tabelle 7.1: Beispiele möglicher Ausgangs- und Zielgrößen

(*polynomiale Regression*), usw. Beispiele für mögliche Ausgangs- bzw. Zielgrößen sind in Tabelle 7.1 zusammengefasst, einige Beispiele in Abbildung 7.10.

Abbildung 7.10: Punktwolken verschiedener Merkmale der StatLab-Auswahl 1985 mit Regressionsgerade, Bestimmtheitsmaß B und Korrelationskoeffizient ρ .

Auf Modellebene ist damit folgende Fragestellung gegeben: Es gebe Zufallsstichproben von Ziel- bzw. Ausgangsvariablen (Y_1, \dots, Y_n) und (X_1, \dots, X_n) , zwischen denen ein verrauschter linearer Zusammenhang $Y_i = \alpha + \beta X_i + \varepsilon_i$ besteht, wobei ε_i Störgrößen sind, die nicht direkt beobachtbar und uns somit unbekannt sind. Meistens nimmt man an, dass $E \varepsilon_i = 0 \quad \forall i = 1, \dots, n$ und $Cov(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij}$, d.h. $\varepsilon_1 \dots \varepsilon_n$ sind unkorreliert mit $\text{Var } \varepsilon_i = \sigma^2$. Wenn wir über die Eigenschaften der Schätzer für α , β und σ^2 reden, ge-

hen wir davon aus, dass die X -Werte nicht zufällig sind, also $X_i = x_i \quad \forall i = 1, \dots, n$. Wenn man von einer konkreten Stichprobe (y_1, \dots, y_n) für (Y_1, \dots, Y_n) ausgeht, so sollen anhand von den Stichproben (x_1, \dots, x_n) und (y_1, \dots, y_n) Regressionsparameter α (*Regressionskonstante*) und β (*Regressionskoeffizient*) sowie *Regressionsvarianz* σ^2 geschätzt werden. Dabei verwendet man die sogenannte *Methode der kleinsten Quadrate*, die den mittleren quadratischen Fehler von den Datenpunkten $(x_i, y_i)_{i=1, \dots, n}$ des Streudiagramms zur *Regressionsgeraden* $y = \alpha + \beta x$ minimiert:

$$(\alpha, \beta) = \operatorname{argmin}_{\alpha, \beta \in \mathbb{R}} e(\alpha, \beta) \quad \text{mit} \quad e(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2.$$

Da die Darstellung $y_i = \alpha + \beta x_i + \varepsilon_i$ gilt, kann man $e(\alpha, \beta) = 1/n \sum_{i=1}^n \varepsilon_i^2$

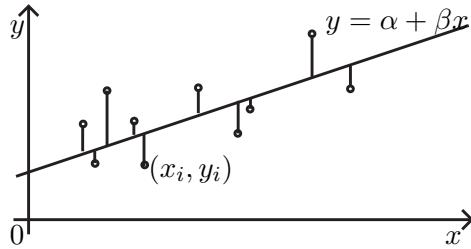


Abbildung 7.11: Methode kleinster Quadrate

schreiben. Es ist der vertikale mittlere quadratische Abstand von den Datenpunkten (x_i, y_i) zur Geraden $y = \alpha + \beta x$ (vgl. Abb. 7.11). Das Minimierungsproblem $e(\alpha, \beta) \mapsto \min$ löst man durch das zweifache Differenzieren von $e(\alpha, \beta)$. Somit erhält man $\hat{\alpha} = \bar{y}_n - \hat{\beta} \bar{x}_n$, wobei

$$\begin{aligned} \hat{\beta} &= \frac{S_{xy}^2}{S_{xx}^2}, \quad \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i, \\ S_{xy}^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n), \quad S_{xx}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2. \end{aligned}$$

Übungsaufgabe 7.24 Leiten Sie die Schätzer $\hat{\alpha}$ und $\hat{\beta}$ selbstständig her.

Die Varianz σ^2 schätzt man durch $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2$, wobei $\hat{\varepsilon}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$, $i = 1, \dots, n$ die sogenannten *Residuen* sind. Die Gründe, warum $\hat{\sigma}^2$ diese Gestalt hat, können an dieser Stelle noch nicht angegeben werden, weil wir noch nicht die Maximum-Likelihood-Methode kennen. Zu gegebener Zeit (in der Vorlesung Stochastik III) wird jedoch klar, dass diese Art der Schätzung sehr natürlich ist.

Bemerkung 7.25 Die angegebenen Schätzer für α und β sind nicht symmetrisch bzgl. Variablen x_i und y_i . Wenn man also die *horizontalen* Abstände (statt vertikaler) zur Bildung des mittleren quadratischen Fehlers nimmt

Kind i	1	2	3	4	5	6	7	8	9
Fernsehzeit x_i	0,3	2,2	0,5	0,7	1,0	1,8	3,0	0,2	2,3
Tiefschlafdauer y_i	5,8	4,4	6,5	5,8	5,6	5,0	4,8	6,0	6,1

Tabelle 7.2: Daten von Fernsehzeit und korrespondierender Tiefschlafdauer

(was dem Rollentausch $x \leftrightarrow y$ entspricht), so bekommt man andere Schätzer für α und β , die mit $\hat{\alpha}$ und $\hat{\beta}$ nicht übereinstimmen müssen:

$$d_i = y_i - \alpha - \beta x_i \mapsto d'_i = x_i - \frac{(y_i - \alpha)}{\beta}.$$

Ein Ausweg aus dieser asymmetrischen Situation wäre es, die orthogonalen

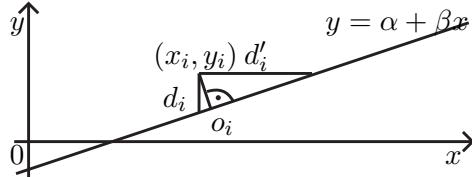


Abbildung 7.12: Orthogonale Abstände

Abstände o_i von (x_i, y_i) zur Geraden $y = \alpha + \beta x$ zu betrachten (vgl. Abb. 7.12). Diese Art der Regression, die „errors-in-variables regression“ genannt wird, hat aber eine Reihe von Eigenschaften, die sie zur Prognose von Zielvariablen y_i durch die Ausgangsvariablen x_i unbrauchbar machen. Sie sollte zum Beispiel nur dann verwendet werden, wenn die Standardabweichungen für X und Y etwa gleich groß sind.

Beispiel 7.26 Ein Kinderpsychologe vermutet, dass sich häufiges Fernsehen negativ auf das Schlafverhalten von Kindern auswirkt. Um diese Hypothese zu überprüfen, wurden 9 Kinder im gleichen Alter befragt, wie lange sie pro Tag fernsehen dürfen, und zusätzlich die Dauer ihrer Tiefschlafphase gemessen. So ergibt sich der Datensatz in Tabelle 7.2 und die Regressionsgerade aus Abbildung 7.13.

Es ergibt sich für die oben genannten Stichproben (x_1, \dots, x_9) und (y_1, \dots, y_9)

$$\bar{x}_9 = 1,33, \quad \bar{y}_9 = 5,56, \quad \hat{\beta} = -0,45, \quad \hat{\alpha} = 6,16.$$

Somit ist

$$y = 6,16 - 0,45x$$

die Regressionsgerade, die eine negative Steigung hat, was die Vermutung des Kinderpsychologen bestätigt. Außerdem ist es mit Hilfe dieser Geraden

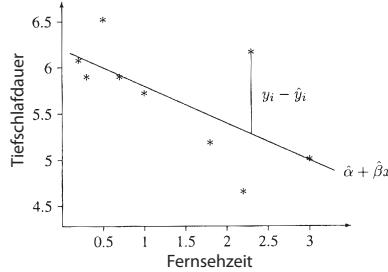


Abbildung 7.13: Streudiagramm und Ausgleichsgerade zur Regression der Dauer des Tiefschlafs auf die Fernsehzeit

möglich, Prognosen für die Dauer des Tiefschlafs für vorgegebene Fernsehzeiten anzugeben. So wäre z.B. für die Fernsehzeit von 1 Stunde der Tiefschlaf von $6,16 - 0,45 \cdot 1 = 5,71$ Stunden plausibel.

Bemerkung 7.27 (Eigenschaften der Regressionsgerade)

1. Es gilt $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\rho_{xy})$, was aus $\hat{\beta} = s_{xy}^2 / s_{xx}^2$ folgt. Dies bedeutet (falls $s_{yy}^2 > 0$):
 - (a) Die Regressionsgerade $y = \hat{\alpha} + \hat{\beta}x$ steigt an, falls die Stichproben (x_1, \dots, x_n) und (y_1, \dots, y_n) positiv korreliert sind.
 - (b) Die Regressionsgerade fällt ab, falls sie negativ korreliert sind.
 - (c) Die Regressionsgerade ist konstant, falls die Stichproben unkorreliert sind.

Falls $s_{yy}^2 = 0$, dann ist die Regressionsgerade konstant ($y = \bar{y}_n$).

2. Die Regressionsgerade $y = \hat{\alpha} + \hat{\beta}x$ verläuft immer durch den Punkt (\bar{x}_n, \bar{y}_n) : $\hat{\alpha} + \hat{\beta}\bar{x}_n = \bar{y}_n$.
3. Seien $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$, $i = 1, \dots, n$. Dann gilt

$$\overline{\hat{y}_n} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \bar{y}_n \quad \text{und somit} \quad \sum_{i=1}^n (\underbrace{y_i - \hat{y}_i}_{\hat{\varepsilon}_i}) = 0.$$

Dabei sind $\hat{\varepsilon}_i$ die schon vorher eingeführten Residuen. Mit ihrer Hilfe ist es möglich, die Güte der Regressionsprognose zu beurteilen.

Residualanalyse und Bestimmtheitsmaß

Definition 7.28 Der relative Anteil der Streuungsreduktion an der Gesamtstreuung S_{yy}^2 heißt das *Bestimmtheitsmaß* der Regressionsgeraden:

$$R^2 = \frac{S_{yy}^2 - \frac{1}{n-1} \sum_{i=1}^n \hat{\varepsilon}_i^2}{S_{yy}^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_n)^2}.$$

Es ist nur im Fall $S_{xx}^2 > 0$, $S_{yy}^2 > 0$ definiert, d.h., wenn nicht alle Werte x_i bzw. y_i übereinstimmen.

Warum R^2 in dieser Form eingeführt wird, zeigt folgende Überlegung, die *Streuungszerlegung* genannt wird:

Lemma 7.29 Die Gesamtstreuung („sum of squares total“) $\text{SQT} = (n - 1)S_{yy}^2 = \sum_{i=1}^n (y_i - \bar{y}_n)^2$ lässt sich in die Summe der sogenannten erklärten Streuung „sum of squares explained“ $\text{SQE} = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2$ und der Residualstreuung „sum of squared residuals“ $\text{SQR} = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ zerlegen:

$$\text{SQT} = \text{SQE} + \text{SQR}$$

bzw.

$$\sum_{i=1}^n (y_i - \bar{y}_n)^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Beweis

$$\begin{aligned} \text{SQT} &= \sum_{i=1}^n (y_i - \bar{y}_n)^2 = \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y}_n)^2 \\ &= \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{=\text{SQR}} + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}_n) + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2}_{=\text{SQE}} \\ &= \text{SQE} + \text{SQR} + 2 \sum_{i=1}^n \hat{y}_i(y_i - \hat{y}_i) - 2\bar{y}_n \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)}_{=0, \text{ vgl. Eig. 3 S. 121}} = \text{SQE} + \text{SQR} + E, \end{aligned}$$

wobei noch zu zeigen ist, dass $E = 2 \sum_{i=1}^n \hat{y}_i(y_i - \hat{y}_i) = 0$, also

$$\begin{aligned} E &= 2 \sum_{i=1}^n (\hat{\alpha} + \hat{\beta}x_i)(y_i - \hat{\alpha} - \hat{\beta}x_i) = 2\hat{\alpha} \underbrace{\sum_{i=1}^n \hat{\varepsilon}_i}_{=0} + 2\hat{\beta} \sum_{i=1}^n x_i(y_i - \hat{\alpha} - \hat{\beta}x_i) \\ &= 2\hat{\beta} \left(\sum_{i=1}^n x_i y_i - \hat{\alpha} \sum_{i=1}^n x_i - \hat{\beta} \sum_{i=1}^n x_i^2 \right) \underset{\hat{\alpha} = \bar{y}_n - \bar{x}_n \hat{\beta}}{=} 2\hat{\beta} \left(\underbrace{\sum_{i=1}^n x_i y_i - n\bar{x}_n \bar{y}_n}_{=(n-1)S_{xy}^2} + \hat{\beta} n\bar{x}_n^2 - \hat{\beta} \sum_{i=1}^n x_i^2 \right) \\ &= 2\hat{\beta} \left((n-1)S_{xy}^2 - \hat{\beta}(n-1)S_{xx}^2 \right) \underset{\hat{\beta} = \frac{S_{xy}^2}{S_{xx}^2}}{=} 2\hat{\beta}(n-1) \left(S_{xy}^2 - \frac{S_{xy}^2}{S_{xx}^2} \cdot S_{xx}^2 \right) = 0. \end{aligned}$$

□

Die erklärte Streuung gibt die Streuung der Regressionsgeradenwerte um \bar{y}_n an. Sie stellt damit die auf den linearen Zusammenhang zwischen X und Y zurückführende Variation der y -Werte dar. Das oben eingeführte Bestimmtheitsmaß ist somit der Anteil dieser Streuung an der Gesamtstreuung:

$$R^2 = \frac{\text{SQE}}{\text{SQT}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^n (y_i - \bar{y}_n)^2} = \frac{\text{SQT} - \text{SQR}}{\text{SQT}} = 1 - \frac{\text{SQR}}{\text{SQT}}.$$

Es folgt aus dieser Darstellung, dass $R^2 \in [0, 1]$ ist.

1. $R^2 = 0$ bedeutet $\text{SQE} = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2 = 0$ und somit $\hat{y}_i = \bar{y}_n \forall i$. Dies weist darauf hin, dass das lineare Modell in diesem Fall schlecht ist, denn aus $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i = \bar{y}_n$ folgt $\hat{\beta} = \frac{S_{xy}^2}{S_{xx}^2} = 0$ und somit $S_{xy}^2 = 0$. Also sind die Merkmale X und Y unkorreliert.
2. $R^2 = 1$ bedingt $\text{SQR} = \sum_{i=1}^n \hat{\varepsilon}_i^2 = 0$. Somit liegen alle (x_i, y_i) perfekt auf der Regressionsgeraden. Dies bedeutet, dass die Daten x_i und y_i , $i = 1, \dots, n$ perfekt linear abhängig sind.

Faustregel zur Beurteilung der Güte der Anpassung eines linearen Modells an Hand von Bestimmtheitsmaß R^2 :

R^2 ist deutlich von Null verschieden (d.h. es besteht noch ein linearer Zusammenhang), falls $R^2 > \frac{4}{n+2}$, wobei n der Stichprobenumfang ist.

Allgemein gilt folgender Zusammenhang zwischen dem Bestimmtheitsmaß R^2 und dem Bravais-Pearson-Korrelationskoeffizienten ϱ_{xy} :

Lemma 7.30

$$R^2 = \varrho_{xy}^2$$

Beweis Aus der Eigenschaft 3 S. 121 folgt $\bar{y}_n = \overline{\hat{y}_n}$. Somit gilt

$$\text{SQE} = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2 = \sum_{i=1}^n (\hat{y}_i - \overline{\hat{y}_n})^2 = \sum_{i=1}^n (\hat{\alpha} + \hat{\beta}x_i - \hat{\alpha} - \hat{\beta}\bar{x}_n)^2 = \hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

und damit

$$R^2 = \frac{\text{SQE}}{\text{SQT}} = \frac{\hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x}_n)^2}{\sum_{i=1}^n (y_i - \bar{y}_n)^2} = \frac{(S_{xy}^2)^2}{(S_{xx}^2)^2} \cdot \frac{(n-1)S_{xx}^2}{(n-1)S_{yy}^2} = \left(\frac{S_{xy}^2}{S_{yy}S_{xx}} \right)^2 = \varrho_{xy}^2$$

□

Folgerung 7.31

- Der Wert von R^2 ändert sich bei einer Lineartransformation der Daten (x_1, \dots, x_n) und (y_1, \dots, y_n) nicht. Grafisch kann man die Güte der Modellanpassung bei der linearen Regression folgendermaßen überprüfen:

Man zeichnet Punktpaare $(\hat{y}_i, \hat{\varepsilon}_i)_{i=1, \dots, n}$ als Streudiagramm (der sogenannte *Residualplot*). Falls diese Punktewolke gleichmäßig um Null streut, so ist das lineare Modell gut gewählt worden. Falls das Streudiagramm einen erkennbaren Trend aufweist, bedeutet das, dass die Annahme des linearen Modells für diese Daten ungeeignet sei (vgl. Abb. 7.14)

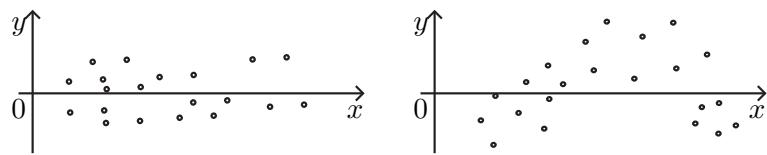


Abbildung 7.14: Links: Gute, Rechts: Schlechte Übereinstimmung mit dem linearen Modell

- Da $R^2 = \varrho_{xy}^2$, ist der Wert von R^2 symmetrisch bzgl. der Stichproben (x_1, \dots, x_n) und (y_1, \dots, y_n) :

$$\varrho_{xy}^2 = R^2 = \varrho_{yx}^2 \quad \text{bzw.} \quad R_{xy}^2 = R_{yx}^2,$$

wobei R_{xy}^2 das Bestimmtheitsmaß bezeichnet, das sich aus der normalen Regression ergibt und R_{yx}^2 das mit vertauschten Achsen.

Kapitel 8

Punktschätzer

8.1 Parametrisches Modell

Sei (x_1, \dots, x_n) eine konkrete Stichprobe. Es wird angenommen, dass (x_1, \dots, x_n) eine Realisierung einer Zufallsstichprobe (X_1, \dots, X_n) ist, wobei X_1, \dots, X_n unabhängige identisch verteilte Zufallsvariablen mit der unbekannten Verteilungsfunktion F sind und F zu einer bekannten parametrischen Familie $\{F_\theta : \theta \in \Theta\}$ gehört. Hier ist $\theta = (\theta_1, \dots, \theta_m) \in \Theta$ der *m-dimensionale Parametervektor* der Verteilung F_θ und $\Theta \subset \mathbb{R}^m$ der sogenannte *Parameterraum* (eine Borel-Teilmenge von \mathbb{R}^m , die die Menge aller zugelassenen Parameterwerte darstellt). Es wird vorausgesetzt, dass die Parametrisierung $\theta \rightarrow F_\theta$ *identifizierbar* ist, indem $F_{\theta_1} \neq F_{\theta_2}$ für $\theta_1 \neq \theta_2$ gilt.

Eine wichtige Aufgabe der Statistik, die wir in diesem Kapitel betrachten werden, besteht in der Schätzung des Parametervektors θ (oder eines Teils von θ) an Hand von der konkreten Stichprobe (x_1, \dots, x_n) . In diesem Fall spricht man von einem *Punktschätzer* $\hat{\theta} : \mathbb{R}^n \rightarrow \mathbb{R}^m$, der eine gültige Stichprobenfunktion ist. Meistens wird angenommen, dass

$$P(\hat{\theta}(X_1, \dots, X_n) \in \Theta) = 1,$$

wobei es zu dieser Regel auch Ausnahmen gibt.

Beispiel 8.1

1. Sei X die Dauer des fehlerfreien Arbeitszyklus eines technischen Systems. Oft wird $X \sim Exp(\lambda)$ angenommen. Dann stellt $\{F_\theta : \theta \in \Theta\}$ mit $m = 1$, $\theta = \lambda$, $\Theta = \mathbb{R}_+$ und

$$F_\theta(x) = (1 - e^{-\theta x}) \cdot I(x \geq 0)$$

ein parametrisches Modell dar, wobei der Parameterraum eindimensional ist. Später wird für λ der (Punkt-) Schätzer $\hat{\lambda}(x_1, \dots, x_n) = 1/\bar{x}_n$ vorgeschlagen.

2. In den Fragestellungen der statistischen Qualitätskontrolle werden n Erzeugnisse auf Mängel untersucht. Falls $p \in (0, 1)$ die unbekannte Wahrscheinlichkeit des Mangels ist, so wird mit $X \sim Bin(n, p)$ die Gesamtanzahl der mangelhaften Produkte beschrieben. Dabei wird folgendes parametrische Modell unterstellt:

$$\Theta = \{(n, p) : n \in \mathbb{N}, p \in (0, 1)\}, \quad \theta = (n, p), \quad m = 2,$$

$$F_\theta(x) = P_\theta(X \leq x) = \begin{cases} 1, & x > n \\ \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k}, & x \in [0, n] \\ 0, & x < 0. \end{cases}$$

Falls n bekannt ist, kann die Wahrscheinlichkeit p des Ausschusses durch den Punktschätzer $\hat{p}(x_1, \dots, x_n) = \bar{x}_n$, $x_i \in \{0, 1\}$ näherungsweise berechnet werden.

8.2 Parametrische Familien von statistischen Prüfverteilungen

In der Vorlesung Wahrscheinlichkeitsrechnung wurden bereits einige parametrische Familien von Verteilungen eingeführt. Hier geben wir weitere Verteilungsfamilien an, die in der Statistik eine besondere Stellung einnehmen, weil sie als Referenzverteilungen in der Schätztheorie, statistischen Tests und Vertrauensintervallen ihre Anwendung finden.

8.2.1 Gamma-Verteilung

Als erstes führen wir zwei spezielle Funktionen aus der Analysis ein:

1. Die *Gamma-Funktion*:

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx \quad \text{für } p > 0.$$

Es gelten folgende Eigenschaften:

$$\begin{aligned} \Gamma(1) &= 1, & \Gamma(1/2) &= \sqrt{\pi} \\ \Gamma(p+1) &= p\Gamma(p) & \forall p > 0, & \Gamma(n+1) = n!, & \forall n \in \mathbb{N}. \end{aligned}$$

2. Die *Beta-Funktion*:

$$B(p, q) = \int_0^1 t^{p-1} (1-t)^{q-1} dt, \quad p, q > 0.$$

Es gelten folgende Eigenschaften:

$$B(p, q) = B(q, p), \quad B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}, \quad p, q > 0.$$

Definition 8.2 Die *Gamma-Verteilung* mit Parametern $\lambda > 0$ und $p > 0$ ist eine absolut stetige Verteilung mit der Dichte

$$f_X(x) = \begin{cases} \frac{\lambda^p x^{p-1}}{\Gamma(p)} e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases} \quad (8.1)$$

Dabei verwenden wir die Bezeichnung $X \sim \Gamma(\lambda, p)$ für eine Zufallsvariable X , die Gamma-verteilt mit Parametern λ und p ist. Es gilt offensichtlich $X \geq 0$ fast sicher für $X \sim \Gamma(\lambda, p)$.

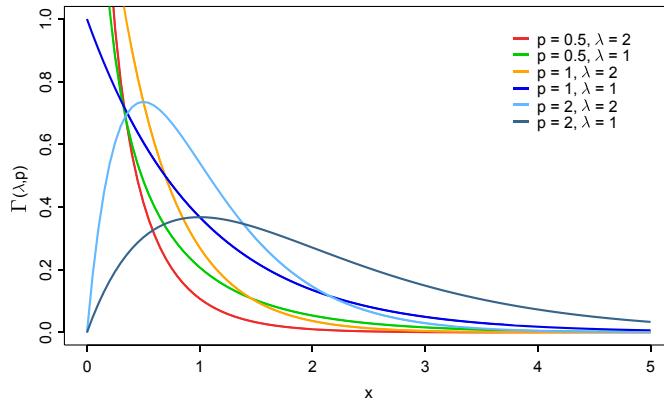


Abbildung 8.1: Dichte der Gammaverteilung

Übungsaufgabe 8.3 Zeigen Sie, dass (8.1) eine Dichte ist.

Beispiel 8.4

1. In der Kraftfahrzeugversicherung wird die Gamma-Verteilung oft zur Modellierung des Gesamtschadens verwendet.
2. Falls $p = 1$, dann ist $\Gamma(\lambda, 1) = \text{Exp}(\lambda)$.

Satz 8.5 (Momenterzeugende und charakteristische Funktion der Gamma-verteilung) Falls $X \sim \Gamma(\lambda, p)$, dann gilt Folgendes:

1. Die momenterzeugende Funktion der Gammaverteilung $\Psi_X(s)$ ist gegeben durch

$$\Psi_X(s) = \mathbb{E}e^{sX} = \frac{1}{(1 - s/\lambda)^p}, \quad s < \lambda.$$

2. k -te Momente:

$$\mathbb{E}X^k = \frac{p(p+1) \cdot \dots \cdot (p+k-1)}{\lambda^k}, \quad k \in \mathbb{N}.$$

Beweis 1. Betrachte

$$\begin{aligned}\Psi_X(s) &= \int_0^\infty e^{sx} f_X(x) dx = \frac{\lambda^p}{\Gamma(p)} \int_0^\infty x^{p-1} e^{\overbrace{(s-\lambda)x}^{<0}} dx \\ &\stackrel{-(s-\lambda)x=y}{=} \frac{\lambda^p}{\Gamma(p)} \int_0^\infty \frac{y^{p-1}}{-(s-\lambda)^p} e^{-y} dy = \frac{\lambda^p \Gamma(p)}{\Gamma(p)(\lambda-s)^p} \\ &= \left(\frac{\lambda}{\lambda-s}\right)^p = \frac{1}{(1-s/\lambda)^p}, \quad \lambda > s.\end{aligned}$$

Falls $s \in \mathbb{C}$, $\operatorname{Re}(s) < \lambda$, dann ist $\Psi_X(s)$ holomorph auf $D = \{z = x + iy \in \mathbb{C} : x < \lambda\}$.

2.

$$\mathbb{E}X^k = \Psi^{(k)}(0) \implies \mathbb{E}X^k = \frac{p \cdot (p+1) \cdot \dots \cdot (p+k-1)}{\lambda^k}, \quad k \in \mathbb{N}.$$

□

Folgerung 8.6 (Faltungsstabilität der Γ -Verteilung) Falls $X \sim \Gamma(\lambda, p_1)$ und $Y \sim \Gamma(\lambda, p_2)$, X, Y unabhängig, dann ist $X + Y \sim \Gamma(\lambda, p_1 + p_2)$.

Beweis Es gilt

$$\begin{aligned}\varphi_{X+Y}(s) &= \varphi_X(s) \cdot \varphi_Y(s) = \frac{1}{(1-is/\lambda)^{p_1}} \cdot \frac{1}{(1-is/\lambda)^{p_2}} = \left(\frac{1}{1-is/\lambda}\right)^{p_1+p_2} \\ &= \varphi_{\Gamma(\lambda, p_1+p_2)}(s).\end{aligned}$$

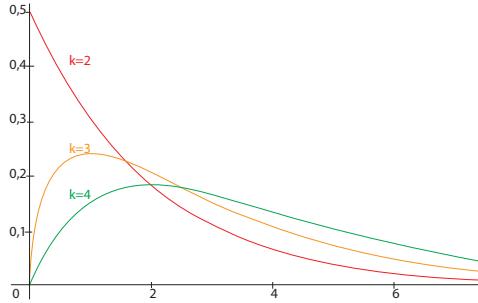
Da die charakteristischen Funktionen die Verteilungen eindeutig bestimmen, folgt damit $X + Y \sim \Gamma(\lambda, p_1 + p_2)$. □

Beispiel 8.7 Seien $X_1, \dots, X_n \sim \operatorname{Exp}(\lambda)$ unabhängig. Nach der Folgerung 8.6 gilt $X = X_1 + \dots + X_n \sim \Gamma(\lambda, \underbrace{1 + \dots + 1}_n) = \Gamma(\lambda, n)$, denn $\operatorname{Exp}(\lambda) = \Gamma(\lambda, 1)$. Dabei heißt X *Erlang-verteilt* mit Parametern λ und n . Man schreibt $X \sim \operatorname{Erl}(\lambda, n)$.

Zusammengefasst: $\operatorname{Erl}(\lambda, n) = \Gamma(\lambda, n)$

Interpretation: In der Risikotheorie z.B. sind X_i Zwischenankunftszeiten der Einzelschäden. Dann ist $X = \sum_{i=1}^n X_i$ die Ankunftszeit des n -ten Schadens, $X \sim \operatorname{Erl}(\lambda, n)$.

Definition 8.8 (χ^2 -Verteilung) X ist eine χ^2 -verteilte Zufallsvariable mit k Freiheitsgraden ($X \sim \chi_k^2$), falls $X \stackrel{d}{=} X_1^2 + \dots + X_k^2$, wobei $X_1, \dots, X_k \sim N(0, 1)$ unabhängige identisch verteilte Zufallsvariablen sind.

Abbildung 8.2: Dichte der χ^2 -Verteilung für $k = 2, 3, 4$

Satz 8.9 (χ^2 -Verteilung: Spezialfall der Γ -Verteilung mit $\lambda = 1/2, p = k/2$)
Falls $X \sim \chi_k^2$, dann gilt:

1. $X \sim \Gamma(1/2, k/2)$, d.h.

$$f_X(x) = \begin{cases} \frac{x^{k/2-1} e^{-x/2}}{2^{k/2} \Gamma(k/2)}, & x \geq 0 \\ 0, & x < 0 \end{cases}. \quad (8.2)$$

2. Insbesondere ist $\mathbb{E}X = k$, $\text{Var } X = 2k$.

Beweis 1. Sei $X = X_1^2 + \dots + X_k^2$ mit $X_i \sim N(0, 1)$ unabhängigen identisch verteilten Zufallsvariablen. Errechnen wir zunächst die *Verteilung der X_i^2* :

$$\begin{aligned} P(X_1^2 \leq x) &= P(X_1 \in [-\sqrt{x}, \sqrt{x}]) = \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\ &= \int_0^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy + \int_{-\sqrt{x}}^0 \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\ &\stackrel{y^2=t}{=} \int_0^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t}{2}} \frac{1}{2\sqrt{t}} dt + \int_x^0 \frac{1}{\sqrt{2\pi}} e^{-\frac{t}{2}} \frac{-1}{2\sqrt{t}} dt \\ &= \int_0^x \frac{(1/2)^{-1/2} t^{1/2-1}}{\Gamma(1/2)} e^{-t/2} dt, \quad x \geq 0. \end{aligned}$$

Somit folgt $X_1^2 \sim \Gamma(1/2, 1/2) \implies X \sim \Gamma(1/2, \underbrace{1/2 + \dots + 1/2}_k) = \Gamma(1/2, k/2)$
und daher gilt der Ausdruck (8.2) für die Dichte.

2. Wegen der Additivität des Erwartungswertes und der Unabhängigkeit von X_i gilt

$$\mathbb{E}X = k \cdot \mathbb{E}X_1^2, \quad \text{Var } X = k \text{Var } X_1^2, \quad \mathbb{E}(X_1^2) = \mathbb{E}(\Gamma(1/2, 1/2)).$$

Bitte zeigen Sie selbstständig, dass $\mathbb{E}X_1^2 = 1$, $\text{Var } X_1^2 = 2$. □

8.2.2 Student-Verteilung (t-Verteilung)

Definition 8.10 Seien X, Y unabhängige Zufallsvariablen, wobei $X \sim N(0, 1)$ und $Y \sim \chi_r^2$. Dann heißt die Zufallsvariable

$$U \stackrel{d}{=} \frac{X}{\sqrt{Y/r}}$$

Student- oder t -verteilt mit r Freiheitsgraden. Wir schreiben $U \sim t_r$.

Satz 8.11 (Dichte der t -Verteilung) Falls $X \sim t_r$, dann gilt:

1.

$$f_X(x) = \frac{1}{\sqrt{r}B\left(\frac{r}{2}, \frac{1}{2}\right)} \cdot \frac{1}{\left(1 + \frac{x^2}{r}\right)^{\frac{r+1}{2}}}, \quad x \in \mathbb{R}.$$

2. $\mathbb{E}X = 0$, $\text{Var } X = \frac{r}{r-2}$, $r \geq 3$.

Bemerkung 8.12

1. **Grafik von f_r :** Die t_r -Verteilung ist symmetrisch. Insbesondere gilt:

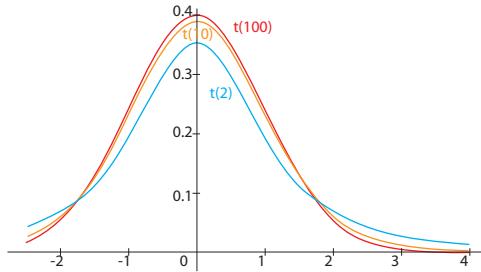


Abbildung 8.3: Dichte \hat{f} der t -Verteilung für $r = 2, 10, 100$

$$t_{r,\alpha} = -t_{r,1-\alpha}, \quad \alpha \in (0, 1),$$

wobei $t_{r,\alpha}$ das α -Quantil der Student-Verteilung mit r Freiheitsgraden ist.

2. Falls $r \rightarrow \infty$, dann $f_r(x) \rightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $x \in \mathbb{R}$. (Übungsaufgabe)
3. Für $r = 1$ gilt: $t_1 = \text{Cauchy}(0, 1)$ mit Dichte $f(x) = \frac{1}{\pi(1+x^2)}$. Der Erwartungswert von t_1 existiert nicht.

8.2.3 Fisher-Snedecor-Verteilung (F-Verteilung)

Definition 8.13 Falls $X \stackrel{d}{=} \frac{U_r/r}{U_s/s}$, wobei $U_r \sim \chi_r^2$, $U_s \sim \chi_s^2$, $r, s \in \mathbb{N}$, U_r, U_s unabhängig, dann hat X eine F-Verteilung mit Freiheitsgraden r, s . Bezeichnung: $X \sim F_{r,s}$.

Bemerkung 8.14 Sei $X \sim F_{r,s}$, $r, s \in \mathbb{N}$ mit Dichte f_X .

1. Einige Graphen der F-Verteilung sind in Abbildung ?? dargestellt.
2. Einige Eigenschaften der F-Verteilung:

Lemma 8.15 Es gilt:

$$(a) \quad \mathbb{E}X = \frac{s}{s-2}, \quad s \geq 3.$$

$$(b) \quad \text{Var } X = \frac{2s^2(r+s-2)}{r(s-4)(s-2)^2}, \quad s \geq 5.$$

(c) Falls $F_{r,s,\alpha}$ das α -Quantil der $F_{r,s}$ -Verteilung ist, dann gilt

$$F_{r,s,\alpha} = \frac{1}{F_{s,r,1-\alpha}}, \quad \alpha \in (0, 1).$$

8.3 Punktschätzer und ihre Grundeigenschaften

Sei (X_1, \dots, X_n) eine Zufallsstichprobe, definiert auf dem kanonischen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P_\theta)$. Seien X_i , $i = 1, \dots, n$ unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion $F \in \{F_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^m$. Finde einen Schätzer $\hat{\theta}(X_1, \dots, X_n)$ für den Parameter θ mit vorgegebenen Eigenschaften.

Unser Ziel im nächsten Abschnitt ist es, zunächst grundlegende Eigenschaften der Schätzer kennenzulernen.

8.3.1 Eigenschaften von Punktschätzern

Definition 8.16 (Erwartungstreue) Ein Schätzer $\hat{\theta}(X_1, \dots, X_n)$ für θ heißt *erwartungstreu* oder *unverzerrt*, falls

$$\mathbb{E}_\theta \hat{\theta}(X_1, \dots, X_n) = \theta, \quad \theta \in \Theta.$$

Dabei wird vorausgesetzt, dass

$$\mathbb{E}_\theta |\hat{\theta}(X_1, \dots, X_n)| < \infty, \quad \theta \in \Theta.$$

Der *Bias* (*Verzerrung*) eines Schätzers $\hat{\theta}(X_1, \dots, X_n)$ ist gegeben durch

$$\text{Bias}(\hat{\theta}) = \mathbb{E}_\theta \hat{\theta}(X_1, \dots, X_n) - \theta.$$

Falls $\hat{\theta}(X_1, \dots, X_n)$ erwartungstreu ist, dann gilt $\text{Bias}(\hat{\theta}) = 0$ (kein systematischer Schätzfehler).

Definition 8.17 (Asymptotische Erwartungstreue) Der Schätzer $\hat{\theta}(X_1, \dots, X_n)$ für θ heißt *asymptotisch erwartungstreu* (oder *asymptotisch unverzerrt*), falls (für große Datenmengen)

$$\mathbb{E}_\theta \hat{\theta}(X_1, \dots, X_n) \xrightarrow[n \rightarrow \infty]{} \theta, \quad \theta \in \Theta.$$

Definition 8.18 (Konsistenz) Falls

$$\hat{\theta}(X_1, \dots, X_n) \xrightarrow[n \rightarrow \infty]{} \theta, \quad \theta \in \Theta$$

in L^2 , stochastisch bzw. fast sicher, dann heißt der Schätzer $\hat{\theta}(X_1, \dots, X_n)$ ein *konsistenter Schätzer* für θ im *mittleren quadratischen, schwachen bzw. starken Sinne*.

- $\hat{\theta}$ *L^2 -konsistent*: für $\mathbb{E}_\theta \hat{\theta}^2(X_1, \dots, X_n) < \infty$ gilt

$$\hat{\theta} \xrightarrow[n \rightarrow \infty]{L^2} \theta \iff \mathbb{E}_\theta |\hat{\theta}(X_1, \dots, X_n) - \theta|^2 \xrightarrow[n \rightarrow \infty]{} 0, \quad \theta \in \Theta.$$

- $\hat{\theta}$ *schwach konsistent*:

$$\hat{\theta} \xrightarrow[n \rightarrow \infty]{P} \theta \iff P_\theta(|\hat{\theta}(X_1, \dots, X_n) - \theta| > \varepsilon) \xrightarrow[n \rightarrow \infty]{} 0, \quad \varepsilon > 0, \quad \theta \in \Theta.$$

- $\hat{\theta}$ *stark konsistent*:

$$\hat{\theta} \xrightarrow[n \rightarrow \infty]{\text{f.s.}} \theta \iff P_\theta \left(\lim_{n \rightarrow \infty} \hat{\theta}(X_1, \dots, X_n) = \theta \right) = 1, \quad \theta \in \Theta.$$

Daraus ergibt sich folgendes Diagramm (vgl. Wahrscheinlichkeitsrechungsskript, Kapitel 6).



Definition 8.19 (Mittlerer quadratischer Fehler (mean squared error)) Der mittlere quadratische Fehler eines Schätzers $\hat{\theta}(X_1, \dots, X_n)$ für θ ist definiert als

$$MSE(\hat{\theta}) = \mathbb{E}_\theta |\hat{\theta}(X_1, \dots, X_n) - \theta|^2.$$

Lemma 8.20 Falls $m = 1$ und $\mathbb{E}_\theta \hat{\theta}^2(X_1, \dots, X_n) < \infty$, $\theta \in \Theta$, dann gilt

$$MSE(\hat{\theta}) = \text{Var}_\theta \hat{\theta} + (\text{Bias}(\hat{\theta}))^2.$$

Beweis

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}_\theta (\hat{\theta} - \theta)^2 = \mathbb{E}_\theta (\hat{\theta} - \mathbb{E}_\theta \hat{\theta} + \mathbb{E}_\theta \hat{\theta} - \theta)^2 \\ &= \underbrace{\mathbb{E}_\theta (\hat{\theta} - \mathbb{E}_\theta \hat{\theta})^2}_{\text{Var}_\theta \hat{\theta}} + 2 \underbrace{\mathbb{E}_\theta (\hat{\theta} - \mathbb{E}_\theta \hat{\theta})(\mathbb{E}_\theta \hat{\theta} - \theta)}_{=0} + \underbrace{(\mathbb{E}_\theta \hat{\theta} - \theta)^2}_{=\text{Bias}(\hat{\theta})^2} \\ &= \text{Var}_\theta \hat{\theta} + (\text{Bias}(\hat{\theta}))^2. \end{aligned}$$

□

Bemerkung 8.21 Falls $\hat{\theta}$ erwartungstreu für θ ist, dann gilt $MSE(\hat{\theta}) = \text{Var}_{\theta} \hat{\theta}$.

Definition 8.22 (Vergleich von Schätzern) Seien $\hat{\theta}_1(X_1, \dots, X_n)$ und $\hat{\theta}_2(X_1, \dots, X_n)$ zwei Schätzer für θ . Man sagt, dass $\hat{\theta}_1$ besser ist als $\hat{\theta}_2$, falls

$$MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2), \quad \theta \in \Theta.$$

Falls $m = 1$ und die Schätzer $\hat{\theta}_1, \hat{\theta}_2$ erwartungstreu sind, so ist $\hat{\theta}_1$ besser als $\hat{\theta}_2$, falls $\hat{\theta}_1$ die kleinere Varianz besitzt. Dabei wird stets vorausgesetzt, dass $\mathbb{E}_{\theta} \hat{\theta}_i^2 < \infty, \quad \theta \in \Theta$.

Definition 8.23 (Asymptotische Normalverteiltheit) Sei $\hat{\theta}(X_1, \dots, X_n)$ ein Schätzer für θ ($m = 1$). Falls $0 < \text{Var}_{\theta} \hat{\theta}(X_1, \dots, X_n) < \infty, \quad \theta \in \Theta$ und

$$\frac{\hat{\theta}(X_1, \dots, X_n) - \mathbb{E}_{\theta} \hat{\theta}(X_1, \dots, X_n)}{\sqrt{\text{Var}_{\theta} \hat{\theta}(X_1, \dots, X_n)}} \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1),$$

dann ist $\hat{\theta}(X_1, \dots, X_n)$ asymptotisch normalverteilt.

Definition 8.24 (Bester erwartungstreuer Schätzer) Der Schätzer $\hat{\theta}(X_1, \dots, X_n)$ für θ ist der beste erwartungstreue Schätzer, falls

$$\mathbb{E}_{\theta} \hat{\theta}^2(X_1, \dots, X_n) < \infty, \quad \theta \in \Theta, \quad \mathbb{E}_{\theta} \hat{\theta}(X_1, \dots, X_n) = \theta, \quad \theta \in \Theta,$$

und $\hat{\theta}$ die minimale Varianz in der Klasse aller erwartungstreuen Schätzer für θ besitzt. Das heißt, dass für einen beliebigen erwartungstreuen Schätzer $\tilde{\theta}(X_1, \dots, X_n)$ mit

$$\mathbb{E}_{\theta} \tilde{\theta}^2(X_1, \dots, X_n) < \infty \quad \text{gilt} \quad \text{Var}_{\theta} \hat{\theta} \leq \text{Var}_{\theta} \tilde{\theta}, \quad \theta \in \Theta.$$

8.3.2 Schätzer des Erwartungswertes und empirische Momente

Sei $X \stackrel{d}{=} X_i, \quad i = 1, \dots, n$ ein statistisches Merkmal. Sei weiter $\mathbb{E}|X_i|^k < \infty$ für ein $k \in \mathbb{N}$, $m = 1$ und der zu schätzende Parameter $\theta = \mu_k = \mathbb{E}X_i^k$. Insbesondere gilt im Fall $k = 1$, dass $\theta = \mu_1 = \mu$ der Erwartungswert ist.

Definition 8.25 Das k -te empirische Moment von X wird als

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

definiert. Unter dieser Definition gilt, dass $\hat{\mu}_1 = \bar{X}_n$, also das erste empirische Moment gleich dem Stichprobenmittel ist.

Satz 8.26 (Eigenschaften der empirischen Momente) Unter obigen Voraussetzungen gelten folgende Eigenschaften:

1. $\hat{\mu}_k$ ist erwartungstreu für μ_k (insbesondere \bar{X}_n).
2. $\hat{\mu}_k$ ist stark konsistent.
3. Falls $\mathbb{E}_\theta |X|^{2k} < \infty$, $\forall \theta \in \Theta$, dann ist $\hat{\mu}_k$ asymptotisch normalverteilt.
4. Es gilt $\text{Var } \bar{X}_n = \frac{\sigma^2}{n}$, wobei $\sigma^2 = \text{Var}_\theta X$. Falls $X_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$ (eine normalverteilte Stichprobe), dann gilt:

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Beweis

1.
$$\mathbb{E}_\theta \hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\theta X_i^k = \frac{1}{n} \sum_{i=1}^n \mu_k = \frac{n\mu_k}{n} = \mu_k.$$

2. Aus dem starken Gesetz der großen Zahlen folgt

$$\frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow[n \rightarrow \infty]{\text{f.s.}} \mathbb{E}_\theta X_i^k = \mu_k.$$

3. Mit dem zentralen Grenzwertsatz gilt

$$\frac{\sum_{i=1}^n X_i^k - n \cdot \mathbb{E} X^k}{\sqrt{n \cdot \text{Var } X^k}} = \frac{\frac{1}{n} \sum_{i=1}^n X_i^k - \mu_k}{\frac{1}{\sqrt{n}} \sqrt{\text{Var } X^k}} = \sqrt{n} \frac{\hat{\mu}_k - \mu_k}{\sqrt{\text{Var } X^k}} \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1).$$

Insbesondere gilt für den Spezialfall $k = 1$

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1).$$

4.

$$\text{Var } \bar{X}_n = \text{Var} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) \underset{X_i \text{ u.i.v.}}{=} \frac{1}{n^2} \sum_{i=1}^n \text{Var } X_i = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Falls $X_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$, dann gilt wegen der Faltungsstabilität der Normalverteilung $\bar{X}_n \sim N(\cdot, \cdot)$, weil

$$\frac{1}{n} X_i \sim N\left(\frac{\mu}{n}, \frac{\sigma^2}{n^2}\right), \quad X_i \text{ u.i.v.}$$

Somit folgt aus 1) und 4) $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Damit ist der Satz bewiesen. □

8.3.3 Schätzer der Varianz

Seien X_i , $i = 1, \dots, n$ unabhängig identisch verteilt, $X_i \stackrel{d}{=} X$, $\mathbb{E}_\theta X^2 < \infty$ $\forall \theta \in \Theta$, $\theta = (\theta_1, \dots, \theta_m)^T$, $\theta_i = \sigma^2 = \text{Var}_\theta X$ für ein $i \in \{1, \dots, m\}$. Die Stichprobenvarianz

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

ist dann ein Schätzer für σ^2 . Falls der Erwartungswert $\mu = \mathbb{E}_\theta X$ der Stichprobenvariablen explizit benannt ist, so kann ein Schätzer für σ^2 auch als

$$\tilde{S}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

definiert werden.

Wir werden nun die Eigenschaften von S_n^2 und \tilde{S}_n^2 untersuchen und sie miteinander vergleichen.

Satz 8.27

1. Die Stichprobenvarianz S_n^2 ist erwartungstreue für σ^2 :

$$\mathbb{E}_\theta S_n^2 = \sigma^2, \quad \theta \in \Theta.$$

2. Wenn $\mathbb{E}_\theta X^4 < \infty$, dann gilt

$$\text{Var}_\theta S_n^2 = \frac{1}{n} \left(\mu'_4 - \frac{n-3}{n-1} \sigma^4 \right),$$

wobei $\mu'_4 = \mathbb{E}_\theta (X - \mu)^4$.

Satz 8.28 1. Der Schätzer \tilde{S}_n^2 für σ^2 ist erwartungstreue.

2. Es gilt $\text{Var}_\theta \tilde{S}_n^2 = 1/n(\mu'_4 - \sigma^4)$.

Beweis
$$\mathbb{E}_\theta \tilde{S}_n^2 = \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{E}_\theta (X_i - \mu)^2}_{=\text{Var}_\theta X_i} = \frac{1}{n} \sum_{i=1}^n \sigma^2 = \sigma^2.$$

□

Folgerung 8.29 Der Schätzer \tilde{S}_n^2 für σ^2 ist besser als S_n^2 , weil beide erwartungstreue sind und

$$\text{Var}_\theta \tilde{S}_n^2 = \frac{\mu'_4 - \sigma^4}{n} < \frac{\mu'_4 - \frac{n-3}{n-1} \sigma^4}{n} = \text{Var}_\theta S_n^2.$$

Diese Eigenschaft von \tilde{S}_n^2 im Vergleich zu S_n^2 ist intuitiv klar, da man in \tilde{S}_n^2 mehr Informationen über die Verteilung der Stichprobenvariablen X_i (nämlich den bekannten Erwartungswert μ) reingesteckt hat.

Satz 8.30 Die Schätzer S_n^2 bzw. \tilde{S}_n^2 sind stark konsistent und asymptotisch normalverteilt:

$$\begin{aligned} S_n^2 &\xrightarrow[n \rightarrow \infty]{\text{f.s.}} \sigma^2, & \sqrt{n} \frac{S_n^2 - \sigma^2}{\sqrt{\mu'_4 - \sigma^4}} &\xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1), \\ \tilde{S}_n^2 &\xrightarrow[n \rightarrow \infty]{\text{f.s.}} \sigma^2, & \sqrt{n} \frac{\tilde{S}_n^2 - \sigma^2}{\sqrt{\mu'_4 - \sigma^4}} &\xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1), \end{aligned}$$

falls $\mu'_4 < \infty$.

Folgerung 8.31 Es gilt

$$1. \quad \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1)$$

und somit

$$2. \quad P\left(\mu \in \left[\bar{X}_n - \frac{z_{1-\alpha/2} S_n}{\sqrt{n}}, \bar{X}_n + \frac{z_{1-\alpha/2} S_n}{\sqrt{n}}\right]\right) \xrightarrow[n \rightarrow \infty]{} 1 - \alpha \quad (8.3)$$

für ein $\alpha \in (0, 1)$, wobei z_α das α -Quantil der $N(0, 1)$ -Verteilung ist.

Bemerkung 8.32 Das Intervall in (8.3) nennt man *asymptotisches Konfidenz- oder Vertrauensintervall* für den Parameter μ . Falls α klein ist (z.B. $\alpha = 0,05$), so liegt μ mit einer asymptotisch großen Wahrscheinlichkeit $1 - \alpha$ im vorgegebenen Intervall. Diese Art der Schätzung von μ stellt eine Alternative zu den Punktschätzern dar und wird ausführlich in Kapitel 4 behandelt.

Beweis der Folgerung 8.31

1. Aus Satz 8.30 folgt

$$S_n^2 \xrightarrow[n \rightarrow \infty]{\text{f.s.}} \sigma^2 \implies \frac{\sigma}{S_n} \xrightarrow[n \rightarrow \infty]{\text{f.s.}} 1 \implies \frac{\sigma}{S_n} \xrightarrow[n \rightarrow \infty]{d} 1$$

und somit nach der Verwendung des Satzes von Slutsky

$$\sqrt{n} \frac{\bar{X}_n - \mu}{S_n} = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \cdot \frac{\sigma}{S_n} \xrightarrow[n \rightarrow \infty]{d} Y \cdot 1 = Y \sim N(0, 1),$$

wobei wir die asymptotische Normalverteiltheit von \bar{X}_n benutzt haben.

2. Aus 1) folgt

$$P_\theta \left(\sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \in [z_{\alpha/2}, z_{1-\alpha/2}] \right) \xrightarrow[n \rightarrow \infty]{} \Phi(z_{1-\alpha/2}) - \Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.$$

Daraus folgt das Intervall (8.3) nach der Auflösung der Ungleichung

$$z_{\alpha/2} \leq \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \leq z_{1-\alpha/2}$$

bzgl. μ .

□

Betrachten wir weiterhin den wichtigen Spezialfall der normalverteilten Stichprobenvariablen X_i , $i = 1, \dots, n$, also $X \sim N(\mu, \sigma^2)$.

Satz 8.33 Falls X_1, \dots, X_n normalverteilt sind mit Parametern μ und σ^2 , dann gilt

1. $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$,
2. $\frac{n\tilde{S}_n^2}{\sigma^2} \sim \chi_n^2$.

Lemma 8.34 Falls $X \sim N(\mu, \sigma^2)$, X_1, \dots, X_n unabhängige identisch verteilte Zufallsvariablen, $X_i \stackrel{d}{=} X$, dann sind \bar{X}_n und S_n^2 unabhängig.

Dieses Lemma wird unter Anderem gebraucht, um folgendes Ergebnis zu beweisen:

Satz 8.35 Unter den Voraussetzungen von Lemma 8.34 gilt

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \sim t_{n-1}.$$

Beweis des Satzes 8.35 Aus den Sätzen 8.26, 4) und 8.33 folgt

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{und} \quad \frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2,$$

also

$$Y_1 = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \sim N(0, 1) \quad \text{und} \quad Y_2 = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Nach dem Lemma 8.34 und Satz ?? sind Y_1 und Y_2 unabhängig. Dann gilt

$$T = \frac{Y_1}{\sqrt{\frac{Y_2}{n-1}}} \sim t_{n-1}$$

nach der Definition einer t -Verteilung, wobei

$$T = \frac{\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}}{\sqrt{\frac{(n-1)S_n^2}{\sigma^2(n-1)}}} = \sqrt{n} \frac{\bar{X}_n - \mu}{S_n}.$$

Somit gilt

$$\sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \sim t_{n-1}.$$

□

Bemerkung 8.36 Mit Hilfe des Satzes 8.35 kann folgendes Konfidenzintervall für den Erwartungswert μ einer normalverteilten Stichprobe (X_1, \dots, X_n) bei unbekannter Varianz σ^2 ($X_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$) konstruiert werden:

$$P\left(\mu \in \left[\bar{X}_n - \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} S_n, \bar{X}_n + \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} S_n\right]\right) = 1 - \alpha$$

für $\alpha \in (0, 1)$, denn

$$\begin{aligned} P\left(\sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \in \left[\underbrace{t_{n-1,\alpha/2}}_{= -t_{n-1,1-\alpha/2} \text{ wg. Sym. } t\text{-Vert.}}, t_{n-1,1-\alpha/2}\right]\right) &= F_{t_{n-1}}(t_{n-1,1-\alpha/2}) - F_{t_{n-1}}(t_{n-1,\alpha/2}) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha, \end{aligned} \tag{8.4}$$

wobei $t_{n-1,\alpha}$ das α -Quantil der t_{n-1} -Verteilung darstellt. Der Rest folgt aus (8.4) durch das Auflösen bzgl. μ .

8.3.4 Eigenschaften der Ordnungsstatistiken

In Abschnitt 7.6.2 haben wir bereits die Ordnungsstatistiken $x_{(1)}, \dots, x_{(n)}$ einer konkreten Stichprobe (x_1, \dots, x_n) betrachtet. Wenn wir nun auf der Modellebene arbeiten, also eine Zufallsstichprobe (X_1, \dots, X_n) von unabhängigen identisch verteilten Zufallsvariablen X_i mit Verteilungsfunktion $F(x)$ haben, welche Eigenschaften haben dann ihre Ordnungsstatistiken

$$X_{(1)}, \dots, X_{(n)}?$$

Satz 8.37

1. Die Verteilungsfunktion der Ordnungsstatistik $X_{(i)}$, $i = 1, \dots, n$ ist gegeben durch

$$F_{X_{(i)}}(x) = \sum_{k=i}^n \binom{n}{k} F^k(x)(1 - F(x))^{n-k}, \quad x \in \mathbb{R}. \tag{8.5}$$

2. Falls X_i absolut stetig verteilt sind mit Dichte f , die stückweise stetig ist, dann ist auch $X_{(i)}$, $i = 1, \dots, n$ absolut stetig verteilt mit der Dichte

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} f(x) F^{i-1}(x) (1-F(x))^{n-i}, \quad x \in \mathbb{R}.$$

Beweis

Führen wir die Zufallsvariable

$$Y = \#\{i : X_i \leq x\} = \sum_{i=1}^n I(X_i \leq x), \quad x \in \mathbb{R}$$

ein. Da X_1, \dots, X_n unabhängig identisch verteilt mit Verteilungsfunktion F sind, gilt $Y \sim \text{Bin}(n, F(x))$. Weiterhin gilt

$$F_{X_{(i)}}(x) = P(X_{(i)} \leq x) = P(Y \geq i) = \sum_{k=i}^n \binom{n}{k} F^k(x) (1-F(x))^{n-k}, \quad x \in \mathbb{R}.$$

□

Bemerkung 8.38 Für $i = 1$ und $i = n$ sieht die Formel (8.5) besonders einfach aus:

$$\begin{aligned} F_{X_{(1)}}(x) &= 1 - (1 - F(x))^n, & x \in \mathbb{R} \\ F_{X_{(n)}}(x) &= F^n(x), & x \in \mathbb{R}. \end{aligned}$$

Übungsaufgabe 8.39 Zeigen Sie für X_1, \dots, X_n unabhängig identisch verteilt, $X_i \sim U[0, \theta]$, $\theta > 0$, $i = 1, \dots, n$, dass

1. die Dichte von $X_{(i)}$ gleich

$$f_{X_{(i)}}(x) = \begin{cases} \frac{n!}{(i-1)!(n-i)!} \theta^{-n} x^{i-1} (\theta - x)^{n-i}, & x \in (0, \theta) \\ 0, & \text{sonst} \end{cases}$$

und

- 2.

$$\mathbb{E} X_{(i)}^k = \frac{\theta^k n! (i+k-1)!}{(n+k)!(i-1)!}, \quad k \in \mathbb{N}, \quad i = 1, \dots, n$$

sind. Insbesondere gilt $\mathbb{E} X_{(i)} = \frac{i}{n+1} \theta$ und $\text{Var } X_{(i)} = \frac{i(n-i+1)\theta^2}{(n+1)^2(n+2)}$.

8.3.5 Empirische Verteilungsfunktion

Im Folgenden betrachten wir die statistischen Eigenschaften der in Abschnitt 7.5.2 eingeführten empirischen Verteilungsfunktion $\hat{F}_n(x)$ einer Zufallsstichprobe (X_1, \dots, X_n) , wobei $X_i \stackrel{d}{=} X$ unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion $F(\cdot)$ sind.

Satz 8.40 Es gilt

1. $n\hat{F}_n(x) \sim \text{Bin}(n, F(x))$, $x \in \mathbb{R}$.
2. $\hat{F}_n(x)$ ist ein erwartungstreuer Schätzer für $F(x)$, $x \in \mathbb{R}$ mit

$$\text{Var } \hat{F}_n(x) = \frac{F(x)(1 - F(x))}{n}.$$

3. $\hat{F}_n(x)$ ist stark konsistent.
4. $\hat{F}_n(x)$ ist asymptotisch normalverteilt:

$$\sqrt{n} \frac{\hat{F}_n(x) - F(x)}{\sqrt{F(x)(1 - F(x))}} \xrightarrow{d} Y \sim N(0, 1), \quad \forall x : F(x) \in (0, 1).$$

Beweis 1. folgt aus der Darstellung

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x), \quad x \in \mathbb{R},$$

weil $I(X_i \leq x) \sim \text{Bernoulli}(F(x))$, $\forall i = 1, \dots, n$. Somit ist

$$\sum_{i=1}^n I(X_i \leq x) \sim \text{Bin}(n, F(x)).$$

2. Es folgt aus 1)

$$\begin{cases} \mathbb{E}(n\hat{F}_n(x)) = nF(x), & x \in \mathbb{R}, \\ \text{Var } (n\hat{F}_n(x)) = nF(x) \cdot (1 - F(x)), & x \in \mathbb{R}, \end{cases}$$

woraus $\mathbb{E}\hat{F}_n(x) = F(x)$ und $\text{Var } \hat{F}_n(x) = F(x)(1 - F(x))/n$ folgen.

3. Da $Y_i = I(X_i \leq x)$, $i = 1, \dots, n$, $x \in \mathbb{R}$ unabhängige identisch verteilte Zufallsvariablen sind, gilt nach dem starken Gesetz der großen Zahlen

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{n \rightarrow \infty} \mathbb{E}Y_i = F(x).$$

4. folgt aus der Anwendung des zentralen Grenzwertsatzes auf die oben genannte Folge $\{Y_i\}_{i \in \mathbb{N}}$.

□

In Satz 8.40, 3) wird behauptet, dass

$$\hat{F}_n(x) \xrightarrow[n \rightarrow \infty]{\text{f.s.}} F(x), \quad \forall x \in \mathbb{R}.$$

Der nachfolgende Satz von Gliwenko-Cantelli behauptet, dass diese Konvergenz gleichmäßig in $x \in \mathbb{R}$ stattfindet. Um diesen Satz formulieren zu können, betrachten wir den *gleichmäßigen Abstand* zwischen \hat{F}_n und F

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|.$$

Dieser Abstand ist eine Zufallsvariable, die auch *Kolmogorow-Abstand* genannt wird. Er gibt den maximalen Fehler an, den man bei der Schätzung von $F(x)$ durch $\hat{F}_n(x)$ macht.

Übungsaufgabe 8.41 Zeigen Sie, dass

$$D_n = \max_{i \in \{1, \dots, n\}} \max \left\{ F(X_{(i)} - 0) - \frac{i-1}{n}, \frac{i}{n} - F(X_{(i)}) \right\}. \quad (8.6)$$

Beachten Sie dabei die Tatsache, dass $\hat{F}_n(x)$ eine Treppenfunktion mit Sprungstellen $X_{(i)}$, $i = 1, \dots, n$ ist.

Satz 8.42 (Gliwenko-Cantelli) Es gilt $D_n \xrightarrow[n \rightarrow \infty]{\text{f.s.}} 0$.

Satz 8.43 Für jede stetige Verteilungsfunktion F gilt

$$D_n \stackrel{d}{=} \sup_{y \in [0,1]} |\hat{G}_n(y) - y|, \text{ wobei } \hat{G}_n(y) = \frac{1}{n} \sum_{i=1}^n I(Y_i \leq y), \quad y \in \mathbb{R}$$

die empirische Verteilungsfunktion der Zufallsstichprobe (Y_1, \dots, Y_n) mit unabhängigen identisch verteilten Zufallsvariablen $Y_i \sim U[0,1]$, $i = 1, \dots, n$ ist.

Folgerung 8.44 Falls F eine stetige Verteilungsfunktion ist, dann gilt

$$D_n \stackrel{d}{=} \max_{i=1, \dots, n} \max \left\{ Y_{(i)} - \frac{i-1}{n}, \frac{i}{n} - Y_{(i)} \right\},$$

wobei $Y_{(1)}, \dots, Y_{(n)}$ die Ordnungsstatistiken der auf $[0,1]$ gleichverteilten Stichprobenvariablen Y_1, \dots, Y_n sind.

Beweis Benutze dazu die Darstellung (8.6), den Satz 8.43 sowie die Tatsache, dass

$$F(x) = x, \quad x \in [0,1]$$

für die Verteilungsfunktion der $U[0,1]$ -Verteilung ist. \square

Folgende Ergebnisse werden ohne Beweis angegeben:

Bemerkung 8.45 Für die Zwecke des statistischen Testens (vgl. den Anpassungstest von Kolmogorow-Smirnow, Bemerkung 8.47, 2)) ist es notwendig, die Quantile der Verteilung von D_n zu nennen. Auf Grund der Komplexität der Verteilung von D_n ist es jedoch unmöglich, sie explizit anzugeben. Mit Hilfe des Satzes 8.43 ist es möglich, diese Quantile durch Monte-Carlo-Simulationen numerisch zu berechnen. Dazu simuliert man mehrere Stichproben (Y_1, \dots, Y_n) von $U[0, 1]$ -verteilten Pseudozufallszahlen, bildet $\hat{G}_n(x)$ und berechnet D_n nach Folgerung 8.44.

Satz 8.46 (Kolmogorow) Falls die Verteilungsfunktion F der unabhängigen und identisch verteilten StichprobenvARIABLEN X_i , $i = 1, \dots, n$ stetig ist, dann gilt

$$\sqrt{n}D_n \xrightarrow[n \rightarrow \infty]{d} Y,$$

wobei Y eine Zufallsvariable mit der Verteilungsfunktion

$$K(x) = \begin{cases} \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 x^2} = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 x^2}, & x > 0, \\ 0, & \text{sonst} \end{cases}$$

(Kolmogorow-Verteilung) ist.

Bemerkung 8.47

1. Aus Satz 8.46 folgt

$$P(\sqrt{n}D_n \leq x) \underset{n \rightarrow \infty}{\approx} K(x), \quad x \in \mathbb{R}.$$

Die daraus resultierende Näherungsformel

$$P(D_n \leq x) \approx K(x\sqrt{n})$$

ist ab $n > 40$ praktisch brauchbar.

2. *Kolmogorow-Smirnow-Anpassungstest*: Mit Hilfe der Aussage des Satzes 8.46 ist es möglich, folgenden *asymptotischen Anpassungstest von Komogorow-Smirnow* zu entwickeln. Es wird die Haupthypothese $H_0 : F = F_0$ (die unbekannte Verteilungsfunktion der StichprobenvARIABLEN X_1, \dots, X_n ist gleich F_0) gegen die Alternative $H_1 : F \neq F_0$ getestet. Dabei wird H_0 verworfen, falls

$$\sqrt{n}D_n \notin [k_{\alpha/2}, k_{1-\alpha/2}]$$

ist, wobei

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$

und k_α das α -Quantil der Kolmogorow-Verteilung ist. Somit ist die Wahrscheinlichkeit, die richtige Hypothese H_0 zu verwerfen (Wahrscheinlichkeit des *Fehlers 1. Art*) asymptotisch gleich

$$P\left(\sqrt{n}D_n \notin [k_{\alpha/2}, k_{1-\alpha/2}] \mid H_0\right) \xrightarrow{n \rightarrow \infty} 1 - K(k_{1-\alpha/2}) + K(k_{\alpha/2}) = 1 - (1 - \alpha/2) + \alpha/2 = \alpha.$$

In der Praxis wird α klein gewählt, z.B. $\alpha \approx 0,05$. Somit ist im Fall, dass H_0 stimmt, die Wahrscheinlichkeit einer Fehlentscheidung in Folge des Testens klein.

Dieser Test ist nur ein Beispiel dessen, wie der Satz von Kolmogorow in der statistischen Testtheorie verwendet wird. Die allgemeine Philosophie des Testens wird in Stochastik III erläutert.

Mit Hilfe von \hat{F}_n lassen sich sehr viele Schätzer durch die sogenannte *Plug-in-Methode* konstruieren. Dies werden wir jetzt näher erläutern: Sei $M = \{\text{Menge aller Verteilungsfunktionen}\}$.

8.4 Methoden zur Gewinnung von Punktschätzern

Sei (X_1, \dots, X_n) eine Stichprobe von unabhängigen identisch verteilten Zufallsvariablen X_i mit Verteilungsfunktion $F \in \{F_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^m$ (Parametrisches Modell). Sei die Parametrisierung $\theta \mapsto F_\theta$ unterscheidbar, d.h. $F_\theta \neq F_{\theta'} \iff \theta \neq \theta'$.

Zielstellung: Konstruiere einen Schätzer $\hat{\theta}(X_1, \dots, X_n)$ für $\theta = (\theta_1, \dots, \theta_m)$.

8.4.1 Momentenschätzer

Aus der Wahrscheinlichkeitsrechnung (Satz 4.8) folgt, dass unter gewissen Voraussetzungen (z.B. Gleichverteilung auf einem kompakten Intervall) an die Verteilung F diese Verteilung aus der Kenntnis von Momenten $\mathbb{E}X^k$, $k \in \mathbb{N}$ wiedergewonnen werden kann. Auf dieser Idee der Schätzung von F aus den Momenten basiert die von Karl Pearson am Ende des XIX. Jh. vorgeschlagene *Momentenmethode*.

Annahme: Es existiert ein $r \geq m$, so dass $\mathbb{E}_\theta|X_i|^r < \infty$. Seien die Momente $\mathbb{E}_\theta X_i^k = g_k(\theta)$, $k = 1, \dots, r$ als Funktionen des Parametervektors $\theta = (\theta_1, \dots, \theta_m) \in \Theta$ gegeben.

Momenten-Gleichungssystem: $\hat{\mu}_k = g_k(\theta)$, $k = 1, \dots, r$, wobei $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$ die k -ten empirischen Momente sind.

Definition 8.48 Falls das obige Gleichungssystem eindeutig lösbar bzgl. θ ist, so heißt die Lösung $\hat{\theta}(X_1, \dots, X_n)$ *Momentenschätzer (M-Schätzer)* von θ .

Lemma 8.49 Falls die Funktion $g = (g_1, \dots, g_r) : \Theta \rightarrow C \subset \mathbb{R}^r$ eineindeutig und ihre Inverse $g^{-1} : C \rightarrow \Theta$ stetig ist, dann ist der M-Schätzer $\hat{\theta}(X_1, \dots, X_n)$ von θ stark konsistent.

Beweis Es gilt $\hat{\theta}(X_1, \dots, X_n) = g^{-1}(\hat{\mu}_1, \dots, \hat{\mu}_r) \xrightarrow[n \rightarrow \infty]{\text{f.s.}} \theta$, weil $\hat{\mu}_k \xrightarrow[n \rightarrow \infty]{\text{f.s.}} g_k(\theta)$, $k = 1, \dots, r$ (starke Konsistenz der empirischen Momente) und g^{-1} stetig. \square

Bemerkung 8.50

1. Andere Eigenschaften gelten für M-Schätzer im Allgemeinen nicht. Zum Beispiel sind nicht alle M-Schätzer erwartungstreue (vgl. Beispiel 8.51, 1)).
2. Manchmal sind $r > m$ Gleichungen im Momentensystem notwendig, um einen M-Schätzer zu bekommen. Dies ist zum Beispiel dann der Fall, wenn manche Funktionen $g_i = \text{const}$ sind, d.h. sie enthalten keine Information über θ (vgl. Beispiel 8.51, 2)).

Beispiel 8.51

1. *Normalverteilung:* $X_i \stackrel{d}{=} X$, $i = 1, \dots, n$, $X \sim N(\mu, \sigma^2)$; Gesucht ist ein M-Schätzer für μ und σ^2 , also $\theta = (\mu, \sigma^2)$. Es gilt

$$\begin{aligned} g_1(\mu, \sigma^2) &= \mathbb{E}_\theta X = \mu, \\ g_2(\mu, \sigma^2) &= \mathbb{E}_\theta X^2 = \text{Var}_\theta X + (\mathbb{E}_\theta X)^2 = \sigma^2 + \mu^2. \end{aligned}$$

Somit ergibt sich das Gleichungssystem

$$\begin{cases} \frac{1}{n} \sum_{i=1}^n X_i = \mu, \\ \frac{1}{n} \sum_{i=1}^n X_i^2 = \mu^2 + \sigma^2. \end{cases}$$

Damit folgt

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n, \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \hat{\mu}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i^2 - \bar{X}_n^2) \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{n-1}{n} S_n^2. \end{aligned}$$

Das heißt, dass die M-Schätzer $\hat{\mu} = \bar{X}_n$, $\hat{\sigma}^2 = \frac{n-1}{n} S_n^2$ sind. Dabei ist $\hat{\sigma}^2$ nicht erwartungstreue:

$$\mathbb{E}_\theta \hat{\sigma}^2 = \frac{n-1}{n} \cdot \mathbb{E}_\theta S_n^2 = \frac{n-1}{n} \sigma^2.$$

2. *Gleichverteilung:* $X_i \stackrel{d}{=} X$, $i = 1, \dots, n$, $X \sim U[-\theta, \theta]$, $\theta > 0$. Gesucht ist ein Momentenschätzer für θ . Es gilt

$$\begin{aligned} g_1(\theta) &= \mathbb{E}_\theta X = 0, \\ g_2(\theta) &= \mathbb{E}_\theta X^2 = \text{Var}_\theta X = \frac{(\theta - (-\theta))^2}{12} = \frac{(2\theta)^2}{12} = \frac{\theta^2}{3}. \end{aligned}$$

Damit ergibt sich das Gleichungssystem

$$\begin{cases} \frac{1}{n} \sum_{i=1}^n X_i = 0 & \text{unbrauchbar,} \\ \frac{1}{n} \sum_{i=1}^n X_i^2 = \frac{\theta^2}{3}. \end{cases}$$

Es folgt, dass $\hat{\theta} = \sqrt{\frac{3}{n} \sum_{i=1}^n X_i^2}$ der Momentenschätzer für θ ist. Wir haben somit 2 Gleichungen für die Schätzung eines einzigen Parameters θ benötigt, d.h. $r = 2 > m = 1$.

8.4.2 Maximum-Likelihood-Schätzer

Diese wurden von Carl Friedrich Gauss (Anfang des XIX. Jh.) und Sir Ronald Fisher (1922) entdeckt. Seien entweder alle Verteilungen aus der parametrischen Familie $\{F_\theta : \theta \in \Theta\}$ diskret oder alle absolut stetig.

Definition 8.52

1. Falls die Stichprobenvariablen $X_i, i = 1, \dots, n$ absolut stetig verteilt mit Dichte $f_\theta(x)$ sind, dann heißt

$$L(x_1, \dots, x_n, \theta) = \prod_{i=1}^n f_\theta(x_i), \quad (x_1, \dots, x_n) \in \mathbb{R}^n, \quad \theta \in \Theta$$

die *Likelihood-Funktion* der Stichprobe (x_1, \dots, x_n) .

2. Falls die Stichprobenvariablen $X_i, i = 1, \dots, n$ diskret verteilt mit Zähldichte $p_\theta(x) = P_\theta(X_i = x), x \in C$ sind (C ist der Wertebereich von X), dann heißt

$$L(x_1, \dots, x_n, \theta) = \prod_{i=1}^n p_\theta(x_i), \quad (x_1, \dots, x_n) \in C^n, \quad \theta \in \Theta$$

die *Likelihood-Funktion* der Stichprobe (x_1, \dots, x_n) .

Nach dieser Definition gilt im

- *diskreten Fall* $L(x_1, \dots, x_n, \theta) = P_\theta(X_1 = x_1, \dots, X_n = x_n)$
- *absolut stetigen Fall*

$$\begin{aligned} L(x_1, \dots, x_n, \theta) \Delta x_1 \cdot \dots \cdot \Delta x_n &= f_{(X_1, \dots, X_n), \theta}(x_1, \dots, x_n) \Delta x_1 \cdot \dots \cdot \Delta x_n \\ &\approx P_\theta(X_1 \in [x_1, x_1 + \Delta x_1], \dots, X_n \in [x_n, x_n + \Delta x_n]), \quad \Delta x_i \rightarrow 0, \quad i = 1, \dots, n. \end{aligned}$$

Nun wird ein Schätzer für θ so gewählt, dass die Wahrscheinlichkeit

$$P_\theta(X_1 = x_1, \dots, X_n = x_n) \quad \text{bzw.} \quad P_\theta(X_i \in [x_i, x_i + \Delta x_i], \quad i = 1, \dots, n)$$

maximal wird. \implies Maximum-Likelihoodmethode:

Definition 8.53 Sei das Maximierungsproblem $L(x_1, \dots, x_n, \theta) \mapsto \max_{\theta \in \Theta}$ eindeutig lösbar. Dann heißt

$$\hat{\theta}(x_1, \dots, x_n) = \operatorname{argmax}_{\theta \in \Theta} L(x_1, \dots, x_n, \theta)$$

der *Maximum-Likelihood-Schätzer* von θ (*ML-Schätzer*).

Bemerkung 8.54

1. In relativ wenigen Fällen ist ein ML-Schätzer $\hat{\theta}$ für θ explizit auffindbar. In diesen Fällen wird meistens der konstante Faktor von $L(x_1, \dots, x_n, \theta)$ weggeworfen und vom Rest der Logarithmus gebildet:

$$\log L(x_1, \dots, x_n, \theta) \quad (\text{die sog. Loglikelihood-Funktion}).$$

Dadurch wird

$$\prod_{i=1}^n f_\theta(x_i) \quad \text{bzw.} \quad \prod_{i=1}^n p_\theta(x_i)$$

zu einer Summe

$$\sum_{i=1}^n \log f_\theta(x_i) \quad \text{bzw.} \quad \sum_{i=1}^n \log p_\theta(x_i),$$

die leichter bzgl. θ zu differenzieren ist. Danach betrachtet man

$$\frac{\partial \log L(x_1, \dots, x_n, \theta)}{\partial \theta_j} = 0, \quad j = 1, \dots, m.$$

Dies ist die notwendige Bedingung eines Extremums von $\log L$ (und somit von L , weil $\log \nearrow$). Falls dieses System eindeutig lösbar ist, und die Lösung eine Maximum-Stelle ist, dann wird sie zum ML-Schätzer $\hat{\theta}(X_1, \dots, X_n)$ erklärt.

2. In den meisten praxisrelevanten Fällen sind ML-Schätzer jedoch nur numerisch auffindbar.

Beispiel 8.55

1. *Bernoulli-Verteilung*: $X_i \stackrel{d}{=} X$, $i = 1, \dots, n$, $X \sim \text{Bernoulli}(p)$, für ein $p \in [0, 1]$. Da

$$X = \begin{cases} 1, & \text{mit Wkt. } p \\ 0, & \text{sonst} \end{cases}$$

mit Zähldichte

$$p_\theta(x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\},$$

ist die *Likelihood-Funktion der Stichprobe* (X_1, \dots, X_n) gegeben durch

$$L(x_1, \dots, x_n, \theta) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \stackrel{\text{def.}}{=} h(p).$$

- (a) Falls $\sum_{i=1}^n x_i = 0$ ($\iff x_1 = x_2 = \dots = x_n = 0$), es folgt $h(p) = (1-p)^n \rightarrow \max_{p \in [0,1]}$ bei $p = 0$. Dann ist der ML-Schätzer $\hat{p}(0, \dots, 0) = 0$.
- (b) Falls $\sum_{i=1}^n x_i = n$ ($\iff x_1 = x_2 = \dots = x_n = 1$), es folgt $h(p) = p^n \rightarrow \max_{p \in [0,1]}$ bei $p = 1$. Dann ist der ML-Schätzer $\hat{p}(1, 1, \dots, 1) = 1$.
- (c) Falls $0 < \sum_{i=1}^n x_i < n$, dann gilt

$$\log L(x_1, \dots, x_n, p) = n\bar{x}_n \log p + n(1 - \bar{x}_n) \log(1 - p) = n \cdot g(p).$$

Da $g(p) \xrightarrow[p \rightarrow 0,1]{} -\infty$ und

$$\begin{aligned} \frac{\partial g(p)}{\partial p} &= \frac{\bar{x}_n}{p} + \frac{1 - \bar{x}_n}{1 - p} \cdot (-1) = \frac{\bar{x}_n}{p} + \frac{\bar{x}_n - 1}{1 - p} = 0 \\ \iff (1 - p)\bar{x}_n + (\bar{x}_n - 1)p &= 0 \implies p = \bar{x}_n, \end{aligned}$$

folgt aufgrund der Stetigkeit von g , dass g genau ein Extremum $\operatorname{argmax}_p g(p) = \bar{x}_n$ besitzt.

Der ML-Schätzer ist also gegeben durch $\hat{p}(X_1, \dots, X_n) = \bar{X}_n$.

2. *Gleichverteilung:* $X \sim U[0, \theta]$, $\theta > 0$, (X_1, \dots, X_n) unabhängig identisch verteilt, gesucht ist ein ML-Schätzer für θ . Es gilt

$$f_{X_i}(x) = 1/\theta \cdot I(x \in [0, \theta]), \quad i = 1, \dots, n.$$

Somit ist die Likelihood-Funktion durch

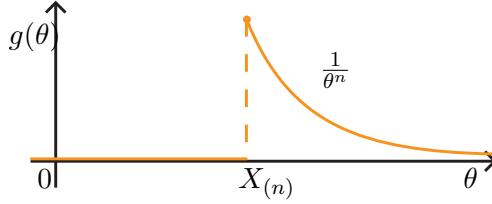
$$\begin{aligned} L(x_1, \dots, x_n, \theta) &= \begin{cases} (1/\theta)^n, & 0 \leq x_1, \dots, x_n \leq \theta \\ 0, & \text{sonst} \end{cases} \\ &= \begin{cases} (1/\theta)^n, & \text{falls } \min\{x_1, \dots, x_n\} \geq 0, \quad \max\{x_1, \dots, x_n\} \leq \theta \\ 0, & \text{sonst} \end{cases} \\ &= g(\theta), \quad \theta > 0 \end{aligned}$$

gegeben. Damit folgt $\hat{\theta} = \operatorname{argmax}_{\theta > 0} g(\theta) = \max\{x_1, \dots, x_n\} = x_{(n)}$, wodurch der ML-Schätzer durch $\hat{\theta}(X_1, \dots, X_n) = X_{(n)}$ gegeben ist.

Satz 8.56 (Schwache Konsistenz von ML-Schätzern) Sei $m = 1$ und Θ ein offenes Intervall aus \mathbb{R} . Sei $L(x_1, \dots, x_n, \theta)$ *unimodal*, d.h. für $\hat{\theta}$ ML-Schätzer für θ gilt

$$\begin{cases} \forall \theta < \hat{\theta}(x_1, \dots, x_n) \implies L(x_1, \dots, x_n, \theta) \text{ ist steigend} \\ \forall \theta > \hat{\theta}(x_1, \dots, x_n) \implies L(x_1, \dots, x_n, \theta) \text{ ist fallend} \end{cases}$$

(d.h. es existiert genau ein $\max_{\theta \in \Theta} L(x_1, \dots, x_n, \theta)$). Dann gilt $\hat{\theta}(X_1, \dots, X_n) \xrightarrow[n \rightarrow \infty]{P} \theta$.

Abbildung 8.4: Illustration der Funktion g .

Definition 8.57 Sei $X = (X_1, \dots, X_n)$ eine Zufallsstichprobe von unabhängigen identisch verteilten Zufallsvariablen $X_i \sim F_\theta$, $\theta \in \Theta$. Sei $L(x, \theta)$ die Likelihood-Funktion von X_i . Dann heißt der Ausdruck

$$I(\theta) = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \log L(X_1, \theta) \right)^2, \quad \theta \in \Theta \quad (8.7)$$

die *Fisher-Information* der Stichprobe (X_1, \dots, X_n) .

Es wird in Zukunft vorausgesetzt, dass $0 < I(\theta) < \infty$. Wir stellen nun einige Bedingungen auf, die für die asymptotische Normalverteiltheit von ML-Schätzern notwendig sind.

1. $\Theta \subset \mathbb{R}$ ist ein offenes Intervall ($m = 1$).
2. Es gelte $P_\theta \neq P_{\theta'}$ genau dann, wenn $\theta \neq \theta'$.
3. Die Familie $\{P_\theta, \theta \in \Theta\}$, $\theta \in \Theta$ bestehe nur aus diskreten oder nur aus absolut stetigen Verteilungen, also nicht aus Mischungen von diskreten und absolut stetigen Verteilungen.
4. $B = \text{supp } L(x, \theta) = \{x \in \mathbb{R} : L(x, \theta) > 0\}$ hängt nicht von $\theta \in \Theta$ ab. Dabei heißt supp (von englisch „support“) der „Träger“ einer Funktion f und ist definiert als

$$\text{supp } f = \{x \in \mathbb{R} : f(x) \neq 0\}$$

und die Likelihood-Funktion $L(x, \theta)$ ist durch

$$L(x, \theta) = \begin{cases} p(x, \theta), & \text{im diskreten Fall,} \\ f(x, \theta), & \text{im absolut stetigen Fall} \end{cases} \quad (8.8)$$

gegeben, wobei $p(x, \theta)$ bzw. $f(x, \theta)$ die Wahrscheinlichkeitsfunktion bzw. Dichte von P_θ ist.

5. Die Abbildung $L(x, \theta)$ ist dreimal stetig differenzierbar und es gilt

$$0 = \frac{d^k}{d\theta^k} \int_B L(x, \theta) dx = \int_B \frac{\partial^k}{\partial \theta^k} L(x, \theta) dx, \quad k = 1, 2, \theta \in \Theta.$$

Da das Integral über die Dichte $L(x, \theta)$ gleich 1 ist, ist die Ableitung gleich 0. Dabei sind im diskreten Fall die Integrale durch Summen zu ersetzen.

6. Für alle $\theta_0 \in \Theta$ existiert eine Konstante $\delta_{\theta_0} > 0$ und eine messbare Funktion $g_{\theta_0} : B \rightarrow [0, \infty)$, so dass

$$\left| \frac{\partial^3 \log L(x, \theta)}{\partial \theta^3} \right| \leq g_{\theta_0}(x), \quad \forall x \in B, \quad |\theta - \theta_0| < \delta_{\theta_0},$$

wobei $\mathbb{E}_{\theta_0} g_{\theta_0}(X_1) < \infty$.

Bemerkung 8.58 Es gilt folgende Relation:

$$n \cdot I(\theta) = \text{Var}_{\theta} \left(\frac{\partial}{\partial \theta} \log L(X_1, \dots, X_n, \theta) \right),$$

wobei

$$L(X_1, \dots, X_n, \theta) = \prod_{i=1}^n L(X_i, \theta) \tag{8.9}$$

die Likelihood-Funktion der Stichprobe (X_1, \dots, X_n) ist mit $L(X_i, \theta)$ nach (8.8).

Beweis Es gilt

$$\frac{\partial}{\partial \theta} \log L(X_1, \dots, X_n, \theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^n \log L(X_i, \theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \log L(X_i, \theta) = \sum_{i=1}^n \frac{L'(X_i, \theta)}{L(X_i, \theta)}.$$

Ferner

$$\mathbb{E}_{\theta} \left(\frac{\partial}{\partial \theta} \log L(X_1, \dots, X_n, \theta) \right) = \sum_{i=1}^n \mathbb{E}_{\theta} \frac{L'(X_i, \theta)}{L(X_i, \theta)} = \sum_{i=1}^n \int_B \frac{L'(X, \theta)}{L(X, \theta)} \cdot L(X, \theta) dx \stackrel{5)}{=} 0.$$

Insgesamt gilt also

$$\begin{aligned} \text{Var}_{\theta} \left(\frac{\partial}{\partial \theta} \log L(X_1, \dots, X_n, \theta) \right) &= \text{Var}_{\theta} \left(\sum_{i=1}^n \frac{\partial}{\partial \theta} \log L(X_i, \theta) \right) \\ &\stackrel{X_i \text{ unabhg.}}{=} \sum_{i=1}^n \text{Var}_{\theta} \left(\frac{\partial}{\partial \theta} \log L(X_i, \theta) \right) \stackrel{X_i \text{ ident.}}{\underset{\text{vert.}}{=}} n \cdot \text{Var}_{\theta} \left(\frac{\partial}{\partial \theta} \log L(X_1, \theta) \right) \\ &= n \cdot \mathbb{E}_{\theta} \left(\frac{\partial}{\partial \theta} \log L(X_1, \theta) \right)^2 = n \cdot I(\theta). \end{aligned}$$

□

Satz 8.59 Sei (X_1, \dots, X_n) eine Stichprobe von Zufallsvariablen, für die die Bedingungen 1) bis 6) erfüllt sind und $0 < I(\theta) < \infty$, $\theta \in \Theta$. Falls $\hat{\theta}(X_1, \dots, X_n)$ ein schwach konsistenter ML-Schätzer für θ ist, dann ist $\hat{\theta}(X_1, \dots, X_n)$ asymptotisch normalverteilt:

$$\sqrt{n \cdot I(\theta)} (\hat{\theta}(X_1, \dots, X_n) - \theta) \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1).$$

8.4.3 Bayes-Schätzer

Sei (X_1, \dots, X_n) eine Zufallsstichprobe, wobei X_i unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion $F_\theta, \theta \in \Theta$ sind. Sei F_θ entweder eine diskrete oder eine absolut stetige Verteilung. Sei aber auch θ eine Zufallsvariable $\tilde{\theta}$ mit Verteilung $Q(\cdot)$ auf dem Messraum $(\Theta, \mathcal{B}_\Theta)$, die entweder diskret mit Zähldichte $q(\cdot)$ oder absolut stetig mit Dichte $q(\cdot)$ ist. Nach wie vor werden beide Fälle gemeinsam betrachtet, dabei entsprechen sich die Summation und Integration im diskreten bzw. absolut stetigen Fall.

Definition 8.60 Die Verteilung $Q(\cdot)$ heißt *a-priori-Verteilung* des Parameters θ (von $\tilde{\theta}$) (a-priori bedeutet hier „vor dem Experiment (X_1, \dots, X_n) “).

Definition 8.61 Die *a-posteriori-Verteilung* des Parameters θ (von $\tilde{\theta}$) ist gegeben durch die (Zähl-)Dichte

$$q_{X_1, \dots, X_n}(\theta, X_1, \dots, X_n) = \begin{cases} P(\tilde{\theta} = \theta \mid X_1 = x_1, \dots, X_n = x_n), & \text{falls } Q \text{ diskret ist,} \\ f_{\tilde{\theta}|X_1, \dots, X_n}(\theta, x_1, \dots, x_n), & \text{falls } Q \text{ absolut stetig ist.} \end{cases}$$

Dabei ist

$$\begin{aligned} P(\tilde{\theta} = \theta \mid X_1 = x_1, \dots, X_n = x_n) &= \frac{P(\tilde{\theta} = \theta, X_1 = x_1, \dots, X_n = x_n)}{P(X_1 = x_1, \dots, X_n = x_n)} \\ &= \frac{P_\theta(X_i = x_i, i = 1, \dots, n) \cdot q(\theta)}{\sum_{\theta_1 \in \Theta} P_{\theta_1}(X_i = x_i, i = 1, \dots, n) \cdot q(\theta_1)} \end{aligned}$$

die *Bayesche Formel*, bzw.

$$f_{\tilde{\theta}|X_1, \dots, X_n}(\theta, x_1, \dots, x_n) = \frac{f_{(\tilde{\theta}, X_1, \dots, X_n)}(\theta, x_1, \dots, x_n)}{f_{X_1, \dots, X_n}(x_1, \dots, x_n)} = \frac{L(x_1, \dots, x_n, \theta) \cdot q(\theta)}{\int_{\Theta} L(x_1, \dots, x_n, \theta_1) \cdot q(\theta_1) d\theta_1},$$

mit $L(x_1, \dots, x_n, \theta)$ nach (8.9).

Definition 8.62 Eine *Verlustfunktion* $V : \Theta^2 \rightarrow \mathbb{R}_+$ ist eine Θ^2 -messbare Funktion.

Verlustfunktionen spielen in unseren Betrachtungen folgende Rolle: $\mathbb{E}_* V(\tilde{\theta}, a)$ stellt den *erwarteten Verlust* (mittleres Risiko) dar, der bei der Schätzung des Parameters θ durch a entsteht. Dabei stellt \mathbb{E}_* den Erwartungswert bezüglich der *a-posteriori-Verteilung* von $\tilde{\theta}$ dar. Es sind offensichtlich die konkreten Stichprobenwerte x_1, \dots, x_n in die a-posteriori-Verteilung eingegangen, deshalb ist $\mathbb{E}_* V(\tilde{\theta}, a)$ eine Funktion von a und x_1, \dots, x_n :

$$\mathbb{E}_* V(\tilde{\theta}, a) = \varphi(x_1, \dots, x_n, a).$$

Definition 8.63 Ein Schätzer $\hat{\theta}$ heißt *Bayes-Schätzer* des Parameters θ , falls

$$\hat{\theta}(x_1, \dots, x_n) = \operatorname{argmin}_a \mathbb{E}_* V(\tilde{\theta}, a) \quad (8.10)$$

existiert und eindeutig ist.

Bemerkung 8.64

1. Manchmal gilt $\hat{\theta} \notin \Theta$, was mit der Existenz des Minimums von $\varphi(x_1, \dots, x_n, a)$ auf Θ zu tun hat.
2. Der Name „Bayesscher Ansatz“ stammt von dem englischen Mathematiker Thomas Bayes (1702–1761), der die Bayessche Formel

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{\sum_j P(A|B_j) \cdot P(B_j)} \quad (8.11)$$

nur ideenhaft eingeführt hat. Der eigentliche Entdecker der Formel (8.11) ist Pierre-Simon Laplace (1749–1827) (Ende des XVIII. Jahrhunderts). Diese Formel wurde bei der Herleitung der *a-posteriori-Verteilung* von $\tilde{\theta}$ implizit benutzt.

3. Die Vorgehensweise in Definition 8.63 ist in konkreten praxisrelevanten Fällen meistens nur numerisch möglich. Es gibt sehr wenige Beispiele für analytische Lösungen des in (8.10) gestellten Minimierungsproblems.

Beispiel 8.65 (Quadratische Verlustfunktion) Ist $V(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$, so ist

$$\operatorname{argmin}_a (\varphi(x_1, \dots, x_n, a)) = \operatorname{argmin}_a (\mathbb{E}_*(\tilde{\theta} - a)^2) = \operatorname{argmin}_a (\mathbb{E}_*\tilde{\theta}^2 - 2a\mathbb{E}_*\tilde{\theta} + a^2) = \mathbb{E}_*\tilde{\theta}$$

und daher der *Bayes-Schätzer* $\hat{\theta}(x_1, \dots, x_n)$ für θ durch $\mathbb{E}_*\tilde{\theta}$ gegeben.

Beispiel 8.66 (Bernoulli-Verteilung) Sei (X_1, \dots, X_n) eine unabhängig identisch verteilte Stichprobe von $X_i \sim \text{Bernoulli}(p)$, $p \in (0, 1)$. Weiter sei die a-priori-Verteilung

$$\tilde{p} \sim \text{Beta}(\alpha, \beta), \quad \alpha, \beta > 0, \text{ mit Zähldichte} \quad q(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} \cdot I(p \in [0, 1]),$$

die a-posteriori-Verteilung von \tilde{p} ist dann gleich

$$q^*(p) = f_{\tilde{p}|X_1=x_1, \dots, X_n=x_n}(p) = \frac{P_p(X_1 = x_1, \dots, X_n = x_n) \cdot q(p)}{\int_0^1 P_{p_1}(X_1 = x_1, \dots, X_n = x_n) \cdot q(p_1) dp_1}.$$

Es ist immer möglich die a-posteriori-Verteilung nicht bezüglich des Vektors (X_1, \dots, X_n) , sondern bezüglich einer Funktion $g(X_1, \dots, X_n)$, zu berechnen (*Komplexitätsreduktion*).

Hier ist $Y = g(X_1, \dots, X_n) = \sum_{i=1}^n X_i$ die Gesamtanzahl aller Erfolge in n Experimenten, wobei

$$X_i = \begin{cases} 1, & \text{mit Wahrscheinlichkeit } p, \\ 0, & \text{sonst.} \end{cases}$$

Daher gilt für die a-posteriori-Verteilung bzgl. Y :

$$\begin{aligned} q^*(p) &= f_{\tilde{p}|Y=k}(p) = \frac{P_p(Y=k) \cdot q(p)}{\int_0^1 P_{p_1}(Y=k) q(p_1) dp_1} \\ &\stackrel{Y \sim \text{Bin}(n,p), \text{ falls } \tilde{p}=p}{=} \frac{\binom{n}{k} p^k (1-p)^{n-k} \cdot (B(\alpha, \beta))^{-1} \cdot p^{\alpha-1} (1-p)^{\beta-1}}{B(\alpha, \beta) \cdot \int_0^1 p_1^{k+\alpha-1} (1-p_1)^{n-k+\beta-1} dp_1} \\ &= \frac{p^{k+\alpha-1} (1-p)^{n-k+\beta-1}}{B(k+\alpha, n-k+\beta)}, \quad p \in [0, 1]. \end{aligned}$$

Daher ist die a-posteriori-Verteilung von \tilde{p} unter der Bedingung $Y = k$ durch

$$\text{Beta}(k + \alpha, n - k + \beta)$$

gegeben.

Für den *Bayes-Schätzer* gilt:

$$\begin{aligned} \hat{p}(x_1, \dots, x_n) &= \mathbb{E}_{*}\tilde{p} = \int_0^1 p \cdot q^*(p) dp = \frac{\int_0^1 p^{k+\alpha} (1-p)^{n-k+\beta-1} dp}{B(k+\alpha, n-k+\beta)} \\ &= \frac{B(k+\alpha+1, n-k+\beta)}{B(k+\alpha, n-k+\beta)} = \dots = \frac{k+\alpha}{\alpha+\beta+n} = \frac{\sum_{i=1}^n x_i + \alpha}{\alpha+\beta+n} = \frac{\alpha+n\bar{x}_n}{\alpha+\beta+n}. \end{aligned}$$

Interpretation:

$$\hat{p}(X_1, \dots, X_n) = \underbrace{\frac{n}{\alpha+\beta+n}}_{=:c_1} \bar{X}_n + \underbrace{\frac{\alpha+\beta}{\alpha+\beta+n}}_{=:c_2} \cdot \frac{\alpha}{\alpha+\beta} = c_1 \cdot \bar{X}_n + c_2 \cdot \mathbb{E}_{apr} \tilde{\theta},$$

wobei $c_1 + c_2 = 1$ ist. Dies heißt, dass die Bayessche Methode einen Mittelweg zwischen dem Schätzer $\mathbb{E}_{apr} \tilde{\theta}$ (in Abwesenheit der Information über die Stichprobe (X_1, \dots, X_n)) und dem M-Schätzer \bar{X}_n (in Abwesenheit der a-priori-Information über die Verteilung von \tilde{p}) für p einschlägt.

8.4.4 Resampling-Methoden zur Gewinnung von Punktschätzern

Sei (X_1, \dots, X_n) eine Stichprobe im parametrischen Modell. Gesucht ist ein Schätzer $\hat{\theta}$ für den Parameter θ . Um diesen Schätzer zu konstruieren, werden bei Resampling-Methoden neue Stichproben (X_1^*, \dots, X_n^*) durch das unabhängige Ziehen mit Zurücklegen aus der alten Stichprobe (X_1, \dots, X_n) generiert und auf ihrer Basis Mittelwerte, Stichprobenvarianzen und andere Schätzer gebildet. Dabei ist die Dimension m des Parameterraums Θ beliebig.

Wir werden im Folgenden die *Resampling*-Methoden

1. *Jackknife* (dt. „Taschenmesser“, weist auf Mittel, die jedem immer zur Hand sein sollten)

2. *Bootstrap* (engl. „self-sufficient“, dt. „mit eigenen Ressourcen“) betrachten.

1. *Jackknife-Methoden zur Schätzung der Varianz bzw. der Verzerrung von Schätzern:*

Als einführendes Beispiel betrachten wir $\theta = \mathbb{E}X = \mu$ bzw. $\theta = \text{Var } X = \sigma^2$ und ihre (erwartungstreue) Schätzer $\hat{\mu} = \bar{X}_n$ bzw. $\hat{\sigma}^2 = S_n^2$.

Wie wir bereits wissen, gilt

$$\text{Var } \hat{\mu} = \frac{\sigma^2}{n}, \quad \text{Var } \hat{\sigma}^2 = \frac{1}{n} \left(\mu'_4 - \frac{n-3}{n-1} \sigma^4 \right).$$

Nun ist ein Schätzer für die Varianz von $\hat{\mu}$ bzw. $\hat{\sigma}^2$ gesucht. Dazu verwenden wir die Plug-in Methode

$$\widehat{\text{Var}} \hat{\mu} = \frac{S_n^2}{n}, \quad \widehat{\text{Var}} \hat{\sigma}^2 = \frac{1}{n} \left(\hat{\mu}'_4 - \frac{n-3}{n-1} S_n^4 \right),$$

wobei $\hat{\mu}'_4$ das vierte zentrierte empirische Moment ist.

Im Allgemeinen sind jedoch *keine* Formeln von $\text{Var } \hat{\theta}$ bekannt. Hier kommt nun die *Jackknife*-Methode zum Einsatz:

- Sei $X_{[i]}$ die Stichprobe $(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$, $i = 1, \dots, n$. Falls

$$\hat{\theta}(X_1, \dots, X_n) = \varphi_n(X_1, \dots, X_n),$$

so bilden wir

$$\hat{\theta}_{[i]} = \varphi_{n-1}(X_{[i]}), \quad \bar{\theta}_{[:] \cdot} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{[i]}, \quad \widehat{\text{Var}}_{jn}(\hat{\theta}) \stackrel{\text{def.}}{=} \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_{[i]} - \bar{\theta}_{[:] \cdot} \right)^2.$$

Definition 8.67 Der Schätzer $\bar{\theta}_{[:] \cdot}$ bzw. $\widehat{\text{Var}}_{jn}(\hat{\theta})$ heißt *Jackknife-Schätzer* für den Erwartungswert bzw. die Varianz des Schätzers $\hat{\theta}$ von θ .

Beispiel 8.68 Sei $\theta = \mu$, $\hat{\theta} = \hat{\mu} = \bar{X}_n$, so gilt

$$\varphi_n(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i,$$

womit folgt, dass

$$\begin{aligned} \hat{\theta}_{[i]} &= \frac{1}{n-1} \sum_{j \neq i} X_j = \frac{1}{n-1} \left(-X_i + \sum_{j=1}^n X_j \right) = \frac{n}{n-1} \bar{X}_n - \frac{1}{n-1} X_i, \quad \forall i = 1, \dots, n, \\ \bar{\theta}_{[:] \cdot} &= \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{[i]} = \frac{n}{n-1} \bar{X}_n - \frac{1}{n(n-1)} \sum_{i=1}^n X_i = \frac{n \cdot \bar{X}_n}{n-1} - \frac{\bar{X}_n}{n-1} = \frac{n-1}{n-1} \bar{X}_n = \bar{X}_n. \end{aligned}$$

Daher ist ein *Jackknife-Schätzer für μ* gleich \bar{X}_n .

Konstruieren wir nun einen *Jackknife-Schätzer der Varianz*:

$$\begin{aligned}\widehat{\text{Var}}_{jn}(\hat{\theta}) &= \frac{n-1}{n} \sum_{i=1}^n \left(\frac{n}{n-1} \bar{X}_n - \frac{1}{n-1} X_i - \bar{X}_n \right)^2 = \frac{n-1}{n} \sum_{i=1}^n \left(\frac{1}{n-1} (\bar{X}_n - X_i) \right)^2 \\ &= \frac{n-1}{n(n-1)^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} S_n^2,\end{aligned}$$

wobei dies genau der Plug-in Schätzer der Varianz von $\hat{\mu}$ ist.

- *Jackknife-Schätzer für die Verzerrung eines Schätzers*
Sei $\hat{\theta}(X_1, \dots, X_n)$ ein Schätzer für θ . Der Bias von $\hat{\theta}$ ist $\mathbb{E}_{\theta} \hat{\theta} - \theta = \text{Bias}(\hat{\theta})$.

Definition 8.69 Ein *Jackknife-Schätzer der Verzerrung (Bias)* von $\hat{\theta}$ ist durch

$$\widehat{\text{Bias}}_{jn}(\hat{\theta}) = (n-1)(\bar{\theta}_{[.]} - \hat{\theta})$$

gegeben.

An folgenden Beispielen wird klar, dass der oben beschriebene Vorgang zur Verringerung der Verzerrung beiträgt:

Der *Schätzer*

$$\tilde{\theta} = \hat{\theta} - \widehat{\text{Bias}}_{jn}(\hat{\theta}) = n\hat{\theta} - (n-1)\bar{\theta}_{[.]} \quad (8.12)$$

hat in der Regel einen *kleineren Bias* als $\hat{\theta}$. Dabei ist wiederum

$$\hat{\theta}_{[i]} = \varphi_{n-1}(X_{[i]}) \quad \text{und} \quad \bar{\theta}_{[.]} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{[i]} \quad \text{mit} \quad \hat{\theta}(X_1, \dots, X_n) = \varphi_n(X_1, \dots, X_n).$$

Beispiel 8.70

- (a) Ist $\theta = \mathbb{E}X_i = \mu$, so ist $\hat{\theta} = \bar{X}_n$ ein unverzerrter Schätzer für μ . Was ist der Bias-korrigierte Schätzer $\tilde{\mu}$? (Dieser sollte schließlich nicht schlechter werden!)

Es gilt $\bar{\theta}_{[.]} = \bar{X}_n$, daher ist der Bias-Schätzer von Jackknife $\widehat{\text{Bias}}_{jn}(\hat{\theta}) = (n-1)(\bar{X}_n - \bar{X}_n) = 0$ und somit $\tilde{\theta} = \hat{\theta} - 0 = \bar{X}_n$. Wir haben also gesehen, dass die Jackknife-Methode die unverzerrten Schätzer (zumindest in diesem Beispiel) richtig behandelt, indem sie keinen zusätzlichen Bias einbaut.

- (b) $\theta = \sigma^2 = \text{Var } X_i$, $\hat{\theta} = \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ ein verzerrter M-Schätzer der Varianz. Was ist $\tilde{\theta}$ in diesem Fall?

Übungsaufgabe 8.71 Zeigen Sie, dass $\tilde{\theta} = S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{n}{n-1} \hat{\sigma}^2$ ein erwartungstreuer Schätzer der Varianz

ist. Somit wird der Bias von $\hat{\sigma}^2$ durch die Anwendung der Jackknife-Methode vollständig beseitigt.

Beweisidee: Zeigen Sie hierzu zunächst, dass

$$\widehat{\text{Bias}}_{jn}(\hat{\theta}) = -\frac{1}{n(n-1)} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Bemerkung 8.72 Die Beispiele 8.70 a), b), in denen sich der Jackknife-Schätzer analytisch bestimmen ließ, sind eher eine Ausnahme als die Regel. In den meisten Fällen erfolgt die Bias-Reduktion mit Hilfe der Monte-Carlo-Methoden auf Basis der Formel (8.12).

2. Bootstrap-Schätzer:

Die Bootstrap-Methode besteht in dem Erzeugen einer neuen Stichprobe (X_1^*, \dots, X_n^*) , die aus einer approximativen Verteilung \hat{F} der StichprobenvARIABLEN $X_i, i = 1, \dots, n$ gewonnen wird. Seien \mathbb{E}_* und Var_* die wahrscheinlichkeitstheoretischen Größen, die auf dem Verteilungsgesetz P_* der neuen Stichprobe (X_1^*, \dots, X_n^*) beruhen. Dabei gibt es folgende Möglichkeiten, \hat{F} zu konstruieren:

- i) $\hat{F}(x) = \hat{F}_n(x)$ die empirische Verteilungsfunktion von X_i , falls X_i unabängig identisch verteilt sind.
- ii) \hat{F} ist ein parametrischer Schätzer von F , der parametrischen Verteilungsfunktion von X_i . Das heißt, falls $X_i \sim F_\theta, i = 1, \dots, n$ für ein $\theta \in \Theta$ und $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ ein Schätzer für θ ist, so setzen wir $\hat{F} = F_{\hat{\theta}}$ (Plug-in Methode).

Definition 8.73 Ein *Bootstrap-Schätzer* für den *Erwartungswert* (bzw. *Bias* oder *Varianz*) von Schätzer $\hat{\theta}(X_1, \dots, X_n)$ ist gegeben durch

- (a) $\hat{\mathbb{E}}_{boot}(\hat{\theta}) = \mathbb{E}_* \hat{\theta}(X_1^*, \dots, X_n^*)$.
- (b) $\widehat{\text{Bias}}_{boot}(\hat{\theta}) = \hat{\mathbb{E}}_{boot} \hat{\theta} - \hat{\theta}$.
- (c) $\widehat{\text{Var}}_{boot}(\hat{\theta}) = \text{Var}_*(\hat{\theta}(X_1^*, \dots, X_n^*))$.

Beispiel 8.74 Sei $\theta = \mu = \mathbb{E} X_i$ und $\hat{F} = \hat{F}_n$ die empirische Verteilungsfunktion. Wie generiert man eine Stichprobe X_1^*, \dots, X_n^* , wobei $X_i^* \sim \hat{F}_n$?

\hat{F}_n gewichtet jede Beobachtung x_i der ursprünglichen Stichprobe mit dem Gewicht $1/n$, deshalb genügt es, einen der Einträge (x_1, \dots, x_n) auszuwählen (mit Wahrscheinlichkeit $1/n$, Urnenmodell „Ziehen mit Zurücklegen“), um $X_j^*, j = 1, \dots, n$ zu generieren.

Bootstrap-Schätzer für den Erwartungswert von $\hat{\mu} = \bar{X}_n$:

$$\hat{\mathbb{E}}_{boot} \hat{\mu} = \mathbb{E}_* \left(\frac{1}{n} \sum_{i=1}^n X_i^* \right) \stackrel{X_i^* \text{ u.i.v.}}{=} \frac{1}{n} \cdot n \mathbb{E}_*(X_1^*) = \int x d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n.$$

Somit folgt $\widehat{\text{Bias}}_{\text{boot}} \hat{\mu} = 0$.

$$\widehat{\text{Var}}_{\text{boot}}(\hat{\mu}) = \text{Var}_{*} \left(\frac{1}{n} \sum_{i=1}^n X_i^* \right) \stackrel{X_i^* \text{ u.i.v.}}{=} \frac{1}{n^2} \cdot n \cdot \text{Var}_{*}(X_1^*) = \frac{1}{n} \cdot \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{\hat{\sigma}^2}{n},$$

ein Plug-in Schätzer für $\text{Var } \bar{X}_n = \sigma^2/n$.

Monte-Carlo-Methoden zur numerischen Berechnung von Bootstrap-Schätzern:

Was kann man tun, wenn keine expliziten Formeln für z.B. $\widehat{\text{Var}}_{\text{Boot}}(\hat{\theta})$ vorliegen (der Regelfall in der Statistik)?

Generiere M unabhängige Stichproben $(X_{i1}^*, \dots, X_{in}^*)$, $i = 1, \dots, M$ nach der Regel i) oder ii) mit Hilfe der Monte-Carlo-Simulation. Dann berechne

$$\hat{\theta}_i = \hat{\theta}(X_{i1}^*, \dots, X_{in}^*), \quad i = 1, \dots, M \quad \text{und setze} \quad \hat{\mathbb{E}}_{\text{boot}} \hat{\theta} \approx \frac{1}{M} \sum_{i=1}^M \hat{\theta}_i.$$

Ähnlich gewinnt man approximative Bootstrap-Schätzer für Bias $\hat{\theta}$ und Var $\hat{\theta}$:

$$\widehat{\text{Bias}}_{\text{boot}} \hat{\theta} \approx \hat{\mathbb{E}}_{\text{boot}} \hat{\theta} - \hat{\theta}, \quad \widehat{\text{Var}}_{\text{boot}} \hat{\theta} \approx \frac{1}{M-1} \sum_{i=1}^M \left(\hat{\theta}_i - \hat{\mathbb{E}}_{\text{boot}} \hat{\theta} \right)^2.$$

Mehr sogar, man kann die Verteilungsfunktion von X_{ij}^* durch die empirische Verteilungsfunktion bestimmen:

$$\hat{F}_{\text{boot}}(x) = \frac{1}{M} \sum_{i=1}^M \frac{1}{n} \sum_{j=1}^n I(X_{ij}^* \leq x), \quad x \in \mathbb{R}.$$

Ferner lassen sich mit Hilfe von oben genannten Methoden *Bootstrap-Konfidenzintervalle* für $\hat{\theta}$ ableiten:

Dafür lassen sich Quantile $\hat{F}_{\hat{\theta}}^{-1}(\alpha_1)$ und $\hat{F}_{\hat{\theta}}^{-1}(\alpha_2)$ der Verteilung von $\hat{\theta}(X_1^*, \dots, X_n^*)$ aus der Stichprobe $(\hat{\theta}_1, \dots, \hat{\theta}_n)$ empirisch bestimmen. Damit gilt

$$P \left(\hat{F}_{\hat{\theta}}^{-1}(\alpha_1) \leq \hat{\theta}(X_1^*, \dots, X_n^*) \leq \hat{F}_{\hat{\theta}}^{-1}(\alpha_2) \right) \approx 1 - \alpha_1 - \alpha_2 = 1 - \alpha,$$

wobei $\alpha = \alpha_1 + \alpha_2$ klein ist. Beachte dabei, dass man hofft, dass X_i^* sehr ähnlich verteilt ist wie X_i und somit

$$P \left(\hat{F}_{\hat{\theta}}^{-1}(\alpha_1) \leq \hat{\theta}(X_1, \dots, X_n) \leq \hat{F}_{\hat{\theta}}^{-1}(\alpha_2) \right) \approx 1 - \alpha_1 - \alpha_2 = 1 - \alpha$$

gilt.

8.5 Weitere Güteeigenschaften von Punktschätzern

8.5.1 Ungleichung von Cramér-Rao

Sei (X_1, \dots, X_n) eine Stichprobe von unabhängigen identisch verteilten Zufallsvariablen X_i mit Verteilungsfunktion F_θ , $\theta \in \Theta$. Sei $\hat{\theta}(X_1, \dots, X_n)$ ein Schätzer für θ . Falls $\hat{\theta}$ erwartungstreu ist, dann misst man die Güte eines anderen erwartungstreuen Schätzers $\tilde{\theta}$ von θ am Wert seiner Varianz. Das bedeutet, falls $\text{Var}_{\theta} \tilde{\theta} < \text{Var}_{\theta} \hat{\theta}$, dann ist der Schätzer $\tilde{\theta}$ besser. Wir werden uns nun mit der Frage befassen, ob immer wieder neue, bessere Schätzer $\tilde{\theta}$ mit immer kleinerer Varianz konstruiert werden können. Die Antwort hierauf ist unter gewissen Voraussetzungen negativ. Die untere Schranke der Varianz $\text{Var}_{\theta} \hat{\theta}$ hierzu liefert der Satz von Cramér-Rao.

Sei $L(x, \theta)$ die Likelihood-Funktion von X_i , d.h.

$$L(x, \theta) = \begin{cases} P_\theta(x), & \text{im diskreten Fall,} \\ f_\theta(x), & \text{im stetigen Fall} \end{cases}$$

und $L(x_1, \dots, x_n, \theta) = \prod_{i=1}^n L(x_i, \theta)$ die Likelihood-Funktion von der gesamten Stichprobe (X_1, \dots, X_n) . Es gelten die Bedingungen 1) bis 5), die für die asymptotische Normalverteiltheit von ML-Schätzern auf Seite 148 gestellt wurden, wobei die Bedingung 5) für $k = 1$ gilt.

Satz 8.75 (Ungleichung von Cramér-Rao) Sei $\hat{\theta}(X_1, \dots, X_n)$ ein Schätzer für θ mit den folgenden Eigenschaften:

1. $\mathbb{E}_\theta \hat{\theta}^2(X_1, \dots, X_n) < \infty \quad \forall \theta \in \Theta$.

2. Für alle $\theta \in \Theta$ existiert

$$\frac{d}{d\theta} \mathbb{E}_\theta \hat{\theta}(X_1, \dots, X_n) = \begin{cases} \int_{\mathbb{R}} \dots \int_{\mathbb{R}} \hat{\theta}(x_1, \dots, x_n) \frac{\partial}{\partial \theta} L(x_1, \dots, x_n, \theta) dx_1 \dots dx_n, & \text{im stetigen Fall,} \\ \sum_{x_1, \dots, x_n} \hat{\theta}(x_1, \dots, x_n) \frac{\partial}{\partial \theta} L(x_1, \dots, x_n, \theta), & \text{im diskr. Fall.} \end{cases}$$

Dann gilt

$$\text{Var}_{\theta} \hat{\theta}(X_1, \dots, X_n) \geq \frac{\left(\frac{d}{d\theta} \mathbb{E}_\theta \hat{\theta}(X_1, \dots, X_n) \right)^2}{n \cdot I(\theta)}, \quad \theta \in \Theta,$$

wobei $I(\theta)$ die Fisher-Information aus (8.7) ist.

Beweis Führen wir die Funktion

$$\varphi_\theta(x_1, \dots, x_n) = \frac{\partial}{\partial \theta} \log L(x_1, \dots, x_n, \theta)$$

ein. In Bemerkung 8.58 haben wir bewiesen, dass

$$\mathbb{E}_\theta \varphi_\theta(X_1, \dots, X_n) = 0, \quad \text{Var}_{\theta} \varphi_\theta(X_1, \dots, X_n) = n \cdot I(\theta).$$

Wenden wir die Ungleichung von Cauchy-Schwarz auf $\text{Cov}_{\theta}(\varphi_{\theta}(X_1, \dots, X_n), \hat{\theta}(X_1, \dots, X_n))$ an:

$$\begin{aligned}\text{Cov}_{\theta}(\varphi_{\theta}(X_1, \dots, X_n), \hat{\theta}(X_1, \dots, X_n)) &= \mathbb{E}_{\theta}(\varphi_{\theta}(X_1, \dots, X_n) \cdot \hat{\theta}(X_1, \dots, X_n)) - 0 \\ &\leq \sqrt{\text{Var}_{\theta} \varphi_{\theta}(X_1, \dots, X_n)} \sqrt{\text{Var}_{\theta} \hat{\theta}(X_1, \dots, X_n)}\end{aligned}$$

Somit folgt

$$\text{Var}_{\theta} \hat{\theta}(X_1, \dots, X_n) \geq \frac{\overbrace{(\mathbb{E}_{\theta}(\varphi_{\theta}(X_1, \dots, X_n) \cdot \hat{\theta}(X_1, \dots, X_n)))}^{=:A}^2}{\text{Var}_{\theta} \varphi_{\theta}(X_1, \dots, X_n)} = \frac{A^2}{n \cdot I(\theta)}.$$

Es bleibt zu zeigen, dass

$$A = \frac{d}{d\theta} \mathbb{E}_{\theta} \hat{\theta}(X_1, \dots, X_n).$$

Wir zeigen die Aussage für den absolut stetigen Fall (im diskreten Fall sind die Integrale durch Summen zu ersetzen):

$$\begin{aligned}A &= \int \frac{\partial}{\partial \theta} \log L(x_1, \dots, x_n, \theta) \cdot \hat{\theta}(x_1, \dots, x_n) \cdot L(x_1, \dots, x_n, \theta) dx_1 \dots dx_n \\ &= \int \frac{\partial}{\partial \theta} L(x_1, \dots, x_n, \theta) \cdot \hat{\theta}(x_1, \dots, x_n) dx_n \stackrel{\text{Vor. 2)}}{=} \frac{d}{d\theta} \mathbb{E}_{\theta} \hat{\theta}(X_1, \dots, X_n).\end{aligned}$$

□

Folgerung 8.76 Falls $\hat{\theta}$ ein erwartungstreuer Schätzer für θ ist und die Voraussetzungen des Satzes 8.75 erfüllt sind, so gilt

$$\text{Var}_{\theta} \hat{\theta}(X_1, \dots, X_n) \geq \frac{1}{n \cdot I(\theta)}.$$

Beweis Wende die Ungleichung von Cramér-Rao an $\hat{\theta}$ mit

$$\frac{d}{d\theta} (\mathbb{E}_{\theta} \hat{\theta}(X_1, \dots, X_n)) = \frac{d}{d\theta} \theta = 1$$

an.

□

An folgenden Beispielen werden wir sehen, dass der Schätzer \bar{X}_n des Erwartungswertes μ in der Klasse aller Schätzer für μ , die die Voraussetzungen des Satzes 8.75 erfüllen, die kleinste Varianz besitzt. Somit ist \bar{X}_n der beste erwartungstreue Schätzer in dieser Klasse für mindestens zwei parametrische Familien von Verteilungen:

- Normalverteilung und

- Poisson-Verteilung.

Beispiel 8.77

1. $X_i \sim N(\mu, \sigma^2)$, $\hat{\mu} = \bar{X}_n$ als Schätzer für μ . Dabei ist $\hat{\mu}$ erwartungstreu mit $\text{Var } \hat{\mu} = \sigma^2/n$. Zeigen wir, dass die Cramér-Rao-Schranke für die Varianz eines erwartungstreuen Schätzers $\hat{\theta}$ für μ ebenso gleich σ^2/n ist. Prüfen wir zunächst die Voraussetzungen des Satzes 8.75:

Zeigen wir, dass

$$0 = \frac{d}{d\mu} \int_{\mathbb{R}} L(x, \mu) dx = \int_{\mathbb{R}} \frac{\partial}{\partial \mu} L(x, \mu) dx \quad \text{mit} \quad L(x, \mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} :$$

$$\begin{aligned} \frac{\partial}{\partial \mu} L(x, \mu) &= \frac{2(x-\mu)}{2\sigma^2} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} = \frac{x-\mu}{\sigma^2} \cdot L(x, \mu), \\ \int_{\mathbb{R}} \frac{\partial}{\partial \mu} L(x, \mu) dx &= \mathbb{E} \left(\frac{X-\mu}{\sigma^2} \right) = 0. \end{aligned}$$

Zeigen wir weiterhin die Gültigkeit der Bedingung 2) des Satzes 8.75:

$$\frac{d}{d\mu} \mathbb{E} \bar{X}_n = \frac{d}{d\mu} (\mu) = 1 \stackrel{?}{=} \frac{1}{n} \int_{\mathbb{R}^n} (x_1 + \dots + x_n) \frac{\partial}{\partial \mu} \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2} \right) dx_1 \dots dx_n .$$

Induktion bzgl. n :

- Induktionsanfang $n = 1$:

$$\int_{\mathbb{R}} x \frac{\partial}{\partial \mu} L(x, \mu) dx = \int_{\mathbb{R}} \frac{x(x-\mu)}{\sigma^2} L(x, \mu) dx = \frac{1}{\sigma^2} (\mathbb{E}_{\mu} X^2 - \mu^2) = \frac{\text{Var}_{\mu} X}{\sigma^2} = 1 .$$

- Induktionshypothese: Für n gilt

$$\int_{\mathbb{R}^n} (x_1 + \dots + x_n) \cdot \frac{\partial}{\partial \mu} L(x_1, \dots, x_n, \mu) dx_1 \dots dx_n = n .$$

- Induktionsschritt $n \rightarrow n+1$:

$$A = \int_{\mathbb{R}^{n+1}} (x_1 + \dots + x_{n+1}) \frac{\partial}{\partial \mu} \underbrace{L(x_1, \dots, x_{n+1}, \mu)}_{=L(x_1, \dots, x_n, \mu) \cdot L(x_{n+1}, \mu)} dx_1 \dots dx_{n+1} \stackrel{?}{=} n+1 .$$

Dabei gilt für A :

$$\begin{aligned}
A &= \int_{\mathbb{R}^{n+1}} (x_1 + \dots + x_n) \cdot \left(\frac{\partial}{\partial \mu} L(x_1, \dots, x_n, \mu) \cdot L(x_{n+1}, \mu) + L(x_1, \dots, x_n, \mu) \cdot \right. \\
&\quad \left. \cdot \frac{\partial}{\partial \mu} L(x_{n+1}, \mu) \right) dx_1 \dots dx_n dx_{n+1} + \int_{\mathbb{R}^{n+1}} x_{n+1} \left(\frac{\partial}{\partial \mu} L(x_1, \dots, x_n, \mu) \cdot \right. \\
&\quad \left. \cdot L(x_{n+1}, \mu) + L(x_1, \dots, x_n, \mu) \cdot \frac{\partial}{\partial \mu} L(x_{n+1}, \mu) \right) dx_1 \dots dx_n dx_{n+1} \\
&= n \cdot \underbrace{\int_{\mathbb{R}} L(x_{n+1}, \mu) dx_{n+1}}_{=1} + \int_{\mathbb{R}^n} (x_1 + \dots + x_n) \cdot L(x_1, \dots, x_n, \mu) dx_1 \dots dx_n \cdot \\
&\quad \cdot \underbrace{\int_{\mathbb{R}} \frac{\partial}{\partial \mu} L(x_{n+1}, \mu) dx_{n+1}}_{=0} + \int_{\mathbb{R}} x_{n+1} L(x_{n+1}, \mu) dx_{n+1} \cdot \\
&\quad \cdot \underbrace{\int_{\mathbb{R}^n} \frac{\partial}{\partial \mu} L(x_1, \dots, x_n, \mu) dx_1 \dots dx_n}_{=0} + \underbrace{\int_{\mathbb{R}} x_{n+1} \frac{\partial}{\partial \mu} L(x_{n+1}, \mu) dx_{n+1}}_{=\frac{d}{d\mu} \mathbb{E}_\mu X = \frac{d}{d\mu} \mu = 1} \cdot \\
&\quad \cdot \underbrace{\int_{\mathbb{R}^n} L(x_1, \dots, x_n, \mu) dx_1 \dots dx_n}_{=1} = n + 1.
\end{aligned}$$

Nachdem alle Voraussetzungen erfüllt sind, berechnen wir die Schranke

$$\frac{1}{n \cdot I(\mu)} \quad \text{mit} \quad I(\mu) = \mathbb{E}_\mu \left(\frac{\partial}{\partial \mu} \log L(X, \mu) \right)^2.$$

Es gilt

$$\frac{\partial}{\partial \mu} \log L(x, \mu) = \frac{\partial}{\partial \mu} \left(-\log \sqrt{2\pi\sigma^2} - \frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right) = -\frac{2(x-\mu)}{2\sigma^2} \cdot (-1) = \frac{x-\mu}{\sigma^2},$$

woraus folgt, dass

$$I(\mu) = \frac{1}{\sigma^4} \mathbb{E}_\mu (X - \mu)^2 = \frac{1}{\sigma^4} \cdot \text{Var}_\mu X = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2} \quad \Rightarrow \quad n \cdot I(\mu) = \frac{n}{\sigma^2}.$$

Insgesamt gilt also

$$\text{Var}_\mu \hat{\theta} \geq \frac{1}{\frac{n}{\sigma^2}} = \frac{\sigma^2}{n} = \text{Var}_\mu \bar{X}_n$$

für einen beliebigen erwartungstreuen Schätzer $\hat{\theta}$ für μ , der die Voraussetzungen des Satzes 8.75 erfüllt.

2. Das zweite Beispiel sei folgende Übungsaufgabe:

Übungsaufgabe 8.78 Seien $X_i \sim \text{Poisson}(\lambda)$, $i = 1, \dots, n$. Zeigen Sie, dass die Schranke von Cramér-Rao

$$\frac{1}{n \cdot I(\lambda)} = \frac{\lambda}{n} = \text{Var}_{\lambda} \bar{X}_n$$

ist. Dies bedeutet, dass auch hier \bar{X}_n der beste erwartungstreue Schätzer ist, der die Voraussetzungen des Satzes 8.75 erfüllt.

An Hand des nächsten Beispiels wollen wir zeigen, dass die Konstruktion von Schätzern mit einer Varianz, die kleiner als die Cramér-Rao-Schranke ist, möglich ist, falls die Voraussetzungen von Satz 8.75 nicht erfüllt sind.

Beispiel 8.79 Seien $X_i \sim U[0, \theta]$, $\theta > 0$. Dann ist die Bedingung „ $\text{supp } f_{\theta}(x) = [0, \theta]$ “ unabhängig von θ verletzt und auch eine weitere Bedingung:

$$0 \neq \int_{\mathbb{R}} \frac{\partial}{\partial \theta} L(x, \theta) dx = \int_0^{\theta} \left(\frac{1}{\theta} \right)' dx = -\frac{1}{\theta^2} \cdot \theta = -\frac{1}{\theta}.$$

Sei $\hat{\theta}$ ein erwartungstreuer Schätzer für θ , so würde nach der Ungleichung von Cramér-Rao folgen, dass $\text{Var}_{\theta} \hat{\theta} \geq (n \cdot I(\theta))^{-1}$, wobei

$$I(\theta) = \mathbb{E} \left(\frac{\partial}{\partial \theta} \log L(X, \theta) \right)^2 = \int_0^{\theta} \frac{1}{\theta} \left(\frac{\partial}{\partial \theta} \log \left(\frac{1}{\theta} \right) \right)^2 dx = \frac{1}{\theta} \int_0^{\theta} dx \cdot \left(-\frac{1}{\theta} \right)^2 = \frac{1}{\theta^2}.$$

Damit hätten wir

$$\text{Var}_{\theta} \hat{\theta} \geq \frac{\theta^2}{n}.$$

Betrachten wir

$$\hat{\theta}(X_1, \dots, X_n) = \frac{n+1}{n} \max\{X_1, \dots, X_n\} = \frac{n+1}{n} X_{(n)}.$$

Zeigen wir, dass

$$\mathbb{E}_{\theta} \hat{\theta}(X_1, \dots, X_n) = \theta \quad \text{und} \quad \text{Var}_{\theta} \hat{\theta}(X_1, \dots, X_n) < \frac{\theta^2}{n}.$$

Berechnen wir dazu $\mathbb{E}_{\theta} X_{(n)}^k$, $k \in \mathbb{N}$. Es gilt

$$F_{X_{(n)}}(x) = F_{X_i}^n(x) = \begin{cases} \frac{x^n}{\theta^n}, & x \in [0, \theta], \\ 1, & x \geq \theta, \\ 0, & x < 0, \end{cases}$$

$$f_{X_{(n)}}(x) = F'_{X_{(n)}}(x) = \frac{nx^{n-1}}{\theta^n} \cdot I(x \in [0, \theta]),$$

$$\mathbb{E}_{\theta} X_{(n)}^k = \int_0^{\theta} x^k \frac{nx^{n-1}}{\theta^n} dx = \frac{n}{\theta^n} \int_0^{\theta} x^{n+k-1} dx = \frac{n \cdot \theta^{n+k}}{\theta^n \cdot (n+k)} = \frac{n\theta^k}{n+k}.$$

Damit folgt

$$\mathbb{E}_\theta \hat{\theta} = \frac{n+1}{n} \cdot \mathbb{E}_\theta X_{(n)} = \frac{n+1}{n} \cdot \frac{n\theta}{n+1} = \theta,$$

das heißt, $\hat{\theta}$ ist erwartungstreu. Weiterhin gilt

$$\begin{aligned} \text{Var}_\theta \hat{\theta} &= \left(\frac{n+1}{n}\right)^2 \cdot \text{Var}_\theta X_{(n)} = \left(\frac{n+1}{n}\right)^2 \cdot \left(\frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2}\right) \\ &= \frac{(n+1)^2}{n^2} \cdot \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2} \cdot \theta^2 \\ &= \frac{\theta^2}{n(n+2)}(n^2 + 2n + 1 - n^2 - 2n) = \frac{\theta^2}{n(n+2)} \end{aligned}$$

und somit

$$\text{Var}_\theta \hat{\theta} = \frac{\theta^2}{n(n+2)} < \frac{\theta^2}{n}.$$

Kapitel 9

Konfidenzintervalle

9.1 Einführung

Konfidenz- oder Vertrauensintervalle wurden bereits in Kapitel 3 exemplarisch behandelt (vgl. Folgerung 3.3.2 und Bemerkung 3.3.4). In diesem Kapitel werden wir eine formale Definition eines Konfidenzintervall es angeben, um Vertrauensintervalle in größerer Tiefe studieren zu können. Dabei werden sowohl *Ein-* als auch *Zweistichprobenprobleme* behandelt.

Rufen wir uns die Annahmen eines parametrischen Modells in Erinnerung: es sei eine Stichprobe (X_1, \dots, X_n) von unabhängigen, identisch verteilten Zufallsvariablen mit $X_i \sim F_\theta$ gegeben, wobei F_θ eine Verteilungsfunktion aus einer parametrischen Familie von Verteilungen $\{F_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}$ ist.

Die Punktschätzer von θ liefern jeweils einen Wert für den Parametervektor. Es wäre allerdings auch vorteilhaft, die Genauigkeit solcher Schätzansätze zu nennen, das heißt, einen Bereich anzugeben, in dem θ mit hoher Wahrscheinlichkeit $1-\alpha$ liegt. Dabei heißt α *Irrtumswahrscheinlichkeit*; übliche Werte für α sind $\alpha = 0,01; 0,05; 0,1$. Die Wahrscheinlichkeit $1 - \alpha$, daß θ für $m = 1$ im vorgegebenen *Konfidenzintervall* liegt, heißt dann *Überdeckungswahrscheinlichkeit* oder *Konfidenzniveau* und soll dann entsprechend hoch ausfallen, z.B. $0,99; 0,95; 0,9$.

Definition 9.1 Es sei $1 - \alpha$ ein Konfidenzniveau und $\underline{\theta} : \mathbb{R}^n \rightarrow \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, $\bar{\theta} : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ zwei Stichprobenfunktionen mit der Eigenschaft

$$\underline{\theta}(x_1, \dots, x_n) \leq \bar{\theta}(x_1, \dots, x_n) \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Falls

1. $P_\theta \left(\theta \in [\underline{\theta}(X_1, \dots, X_n), \bar{\theta}(X_1, \dots, X_n)] \right) \geq 1 - \alpha, \quad \theta \in \Theta$
2. $\inf_{\theta \in \Theta} P_\theta \left(\theta \in [\underline{\theta}(X_1, \dots, X_n), \bar{\theta}(X_1, \dots, X_n)] \right) = 1 - \alpha$

$$3. \lim_{n \rightarrow \infty} P_\theta \left(\theta \in [\underline{\theta}(X_1, \dots, X_n), \bar{\theta}(X_1, \dots, X_n)] \right) = 1 - \alpha, \quad \theta \in \Theta$$

dann heißt $I = [\underline{\theta}(X_1, \dots, X_n), \bar{\theta}(X_1, \dots, X_n)]$ ein

1. Konfidenzintervall
2. minimales Konfidenzintervall
3. asymptotisches Konfidenzintervall

zum Konfidenzniveau $1 - \alpha$.

Dabei heißt $l_\theta(X_1, \dots, X_n) = \bar{\theta}(X_1, \dots, X_n) - \underline{\theta}(X_1, \dots, X_n)$ die Länge des Konfidenzintervalls. Es ist erwünscht, möglichst kleine Konfidenzintervalle (mit minimaler Länge) bei großem Konfidenzniveau für θ zu konstruieren.

Wie bereits bei den Beispielen von Kapitel 8 ersichtlich ist, folgt die Konstruktion eines Konfidenzintervalls einem bestimmten Muster, das wir jetzt genauer studieren werden:

1. Finde eine Statistik $T(X_1, \dots, X_n, \theta)$, die
 - vom Parameter θ abhängt und
 - eine bekannte (Prüf-) Verteilung F besitzt (möglicherweise asymptotisch für $n \rightarrow \infty$).
2. Bestimme die Quantile $F^{-1}(\alpha_1)$ und $F^{-1}(1 - \alpha_2)$ von der Verteilung F für Niveaus α_1 und $1 - \alpha_2$, sodaß $\alpha_1 + \alpha_2 = \alpha$.
3. Löse (falls möglich) die Ungleichung
 $F^{-1}(\alpha_1) \leq T(X_1, \dots, X_n, \theta) \leq F^{-1}(1 - \alpha_2)$ bzgl. θ auf. Das entsprechende Ergebnis $I = [T^{-1}(F^{-1}(\alpha_1)), T^{-1}(F^{-1}(1 - \alpha_2))]$ (im Falle einer monoton in θ steigenden Statistik T) ist ein Konfidenzintervall für θ zum Niveau $1 - \alpha$, denn es gilt

$$\begin{aligned} P_\theta(\theta \in I) &= P_\theta(T_\theta^{-1}(F^{-1}(\alpha_1)) \leq \theta \leq T_\theta^{-1}(F^{-1}(1 - \alpha_2))) \\ &= P_\theta(F^{-1}(\alpha_1) \leq T_\theta(X_1, \dots, X_n, \theta) \leq F^{-1}(1 - \alpha_2)) \\ &= F(F^{-1}(1 - \alpha_2)) - F(F^{-1}(\alpha_1)) \\ &= 1 - \alpha_2 - \alpha_1 \\ &= 1 - \alpha \text{ für alle } \theta \in \Theta. \end{aligned}$$

Für asymptotische Konfidenzintervalle soll überall noch $\lim_{n \rightarrow \infty}$ geschrieben werden:

$\lim_{n \rightarrow \infty} P_\theta(\theta \in I) = \dots = 1 - \alpha$. Hierbei ist T_θ^{-1} die Inverse von $T(X_1, \dots, X_n, \theta)$ bezüglich θ . Grafisch kann dies auf Abb. 4.1 veranschaulicht werden.

Definition 9.2

1. Falls $\alpha_1 = \alpha_2 = \alpha/2$, dann heißt das Konfidenzintervall
 $I = [T^{-1}(F^{-1}(\frac{\alpha}{2})), T^{-1}(F^{-1}(1 - \frac{\alpha}{2}))]$ symmetrisch.
2. Falls $\alpha_1 = 0$ (bzw. $\underline{\theta}(X_1, \dots, X_n) = -\infty$), dann heißt das Konfidenzintervall $(-\infty, \bar{\theta}(X_1, \dots, X_n)]$ einseitig. Dasselbe gilt für $\alpha_2 = 0$ (bzw. $\bar{\theta}(X_1, \dots, X_n) = +\infty$) und $[\underline{\theta}(X_1, \dots, X_n), +\infty)$.

In der Zukunft werden wir oft, ohne Beschränkung der Allgemeinheit, symmetrische Konfidenzintervalle konstruieren, obwohl man auch ein allgemeineres, nicht-symmetrisches Intervall leicht angeben kann.

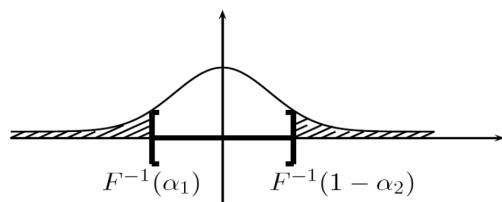


Abbildung 9.1: asymptotisches Konfidenzintervall

Bemerkung 9.3 Man sieht leicht, daß der Algorithmus zur Konstruktion eines Vertrauensbereiches sich sehr dem eines statistischen Tests ähnelt. Im letzten Fall heißt $T(X_1, \dots, X_n)$ *Teststatistik*. Im Allgemeinen kann man für jedes Konfidenzintervall einen entsprechenden statistischen Test angeben, aber nicht umgekehrt. In der Vorlesung Stochastik III werden wir einige Beispiele dieser Übertragung „Konfidenzintervall \mapsto Test“ sehen.

ref Kapitel

9.2 Ein-Stichproben-Probleme

In diesem Abschnitt werden wir einige Beispiele von Vertrauensbereichen für Parameter einiger bekannter Verteilungen nach dem oben genannten Schema konstruieren. Dabei werden wir immer mit einer Stichprobe (X_1, \dots, X_n) wie in Abschnitt 9.1 arbeiten.

9.2.1 Normalverteilung

Es seien X_1, \dots, X_n unabhängig, identisch verteilt, mit $X_i \sim N(\mu, \sigma^2)$.

Konfidenzintervalle für den Erwartungswert μ

- bei bekannter Varianz σ^2 Wenn wir annehmen, daß σ^2 bekannt ist, so ermöglicht uns der Satz 3.3.1, 4., ein exaktes Konfidenzintervall für

μ zum Niveau $1 - \alpha$ zu berechnen. Denn es gilt $\bar{X}_n \sim N(\mu, \sigma^2/n)$ und somit

$$T(X_1, \dots, X_n, \mu) = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \sim N(0, 1)$$

Es seien z_{α_1} und $z_{1-\alpha_2}$ Quantile der $N(0, 1)$ -Verteilung, $\alpha_1 + \alpha_2 = \alpha$ und $1 - \alpha$ das vorgegebene Konfidenzniveau.

Dann gilt

$$\begin{aligned} 1 - \alpha &= P(z_{\alpha_1} \leq T(X_1, \dots, X_n, \mu) \leq z_{1-\alpha_2}) \\ &= P\left(z_{\alpha_1} \leq \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \leq z_{1-\alpha_2}\right) \\ &\stackrel{(-z_{\alpha_1} = z_{1-\alpha_1})}{=} P\left(\bar{X}_n - \frac{z_{1-\alpha_2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{z_{1-\alpha_1}\sigma}{\sqrt{n}}\right). \end{aligned}$$

Somit ist $[\underline{\theta}(X_1, \dots, X_n), \bar{\theta}(X_1, \dots, X_n)]$ mit $\underline{\theta}(X_1, \dots, X_n) = \bar{X}_n - z_{1-\alpha_2} \frac{\sigma}{\sqrt{n}}$ und $\bar{\theta}(X_1, \dots, X_n) = \bar{X}_n + z_{1-\alpha_1} \frac{\sigma}{\sqrt{n}}$ ein exaktes Konfidenzintervall für μ zum Niveau $1 - \alpha$.

Es hat die Länge $l_\mu(X_1, \dots, X_n) = \frac{\sigma}{\sqrt{n}} (z_{1-\alpha_2} + z_{1-\alpha_1})$. Es gilt $l_\mu(X_1, \dots, X_n) \rightarrow 0$ für $n \rightarrow \infty$ was bedeutet, daß bei wachsendem Informationsumfang ($n \rightarrow \infty$) die Präzision der Schätzung immer besser wird.

Im Symmetriefall ($\alpha_1 = \alpha_2 = \alpha/2$) gilt $\underline{\theta}(X_1, \dots, X_n) = \bar{X}_n - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$, $\bar{\theta}(X_1, \dots, X_n) = \bar{X}_n + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$ und $l_\mu(X_1, \dots, X_n) = \frac{2\sigma}{\sqrt{n}} z_{1-\alpha/2}$.

Daraus folgt, daß man bei vorgegebener Länge $\varepsilon > 0$ die Anzahl der Beobachtungen n bestimmen kann, die dann notwendig sind, um die vorgegebene Präzision zu erreichen:

$$\frac{2\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \varepsilon \iff n \geq \left(\frac{2\sigma z_{1-\alpha/2}}{\varepsilon} \right)^2 \quad (9.1)$$

Für $\alpha_1 = 0$ bzw. $\alpha_2 = 0$ kann man einseitige Intervalle $(-\infty, \bar{X}_n + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}]$ und $[\bar{X}_n - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}, +\infty)$ genauso angeben.

- **bei unbekannter Varianz σ^2 :** siehe Bemerkung 3.3.4.

Dort wurde das Konfidenzintervall $[\bar{X}_n - \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} S_n, \bar{X}_n + \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} S_n]$ für μ zum Konfidenzniveau $1 - \alpha$ konstruiert, wobei $t_{n-1,1-\alpha/2}$ das $(1 - \frac{\alpha}{2})$ -Quantil der t_{n-1} -Verteilung ist.

Wie man sieht, ist sie Länge des Konfidenzintervalls zufällig: $l_\mu(X_1, \dots, X_n) = \frac{2S_n}{\sqrt{n}} t_{n-1, 1-\alpha/2}$, somit macht es Sinn, mit erwarteter Länge

$$\mathbb{E} l_\mu(X_1, \dots, X_n) = \frac{2}{\sqrt{n}} \mathbb{E} S_n t_{n-1, 1-\alpha/2}$$

zu arbeiten, um zum Beispiel die Frage nach der notwendigen Anzahl n von Beobachtungen bei vorgegebener Genauigkeit $\varepsilon > 0$ (vergleiche Gleichung (9.1)) zu beantworten.

Konfidenzintervalle für die Varianz σ^2

- bei bekanntem Erwartungswert μ :

Betrachten wir den Schätzer $\tilde{S}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ für σ^2 . Aus Satz 3.3.5, 2. folgt $\frac{n\tilde{S}_n^2}{\sigma^2} \sim \chi_n^2$. Wir setzen $T(X_1, \dots, X_n, \sigma^2) = \frac{n\tilde{S}_n^2}{\sigma^2}$ und bekommen

$$P\left(\chi_{n,\alpha_2}^2 \leq \frac{n\tilde{S}_n^2}{\sigma^2} \leq \chi_{n,1-\alpha_1}^2\right) = P\left(\frac{n\tilde{S}_n^2}{\chi_{n,1-\alpha_1}^2} \leq \sigma^2 \leq \frac{n\tilde{S}_n^2}{\chi_{n,\alpha_2}^2}\right) = 1 - \alpha.$$

Somit ist $\left[\frac{n\tilde{S}_n^2}{\chi_{n,1-\alpha_1}^2}, \frac{n\tilde{S}_n^2}{\chi_{n,\alpha_2}^2}\right]$ ein Konfidenzintervall für σ^2 zum Niveau $1 - \alpha$, $\alpha = \alpha_1 + \alpha_2$ mit der mittleren Länge $\mathbb{E} l_{\sigma^2} = n\sigma^2 \left(\frac{1}{\chi_{n,\alpha_2}^2} - \frac{1}{\chi_{n,1-\alpha_1}^2}\right)$.

- bei unbekanntem Erwartungswert μ :

Ähnlich wie oben beschrieben folgt das Konfidenzintervall $\left[\frac{(n-1)S_n^2}{\chi_{n-1,1-\alpha_1}^2}, \frac{(n-1)S_n^2}{\chi_{n-1,\alpha_2}^2}\right]$ zum Niveau $1 - \alpha$, $\alpha = \alpha_1 + \alpha_2$ aus Satz 3.3.5, 1., weil $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$ für die Stichprobenvarianz $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. Die erwartete Länge ist $\mathbb{E} l_{\sigma^2} = (n-1)\sigma^2 \left(\frac{1}{\chi_{n-1,\alpha_2}^2} - \frac{1}{\chi_{n-1,1-\alpha_1}^2}\right)$.

9.2.2 Konfidenzintervalle aus stochastischen Ungleichungen

Eine alternative Methode zur Gewinnung von Konfidenzintervallen besteht in der Anwendung stochastischer Ungleichungen. So kann man zum Beispiel bei einer Stichprobe (X_1, \dots, X_n) von unabhängigen und identisch verteilten Zufallsvariablen mit $\mathbb{E} X_i = \mu$, $\text{Var } X_i = \sigma^2 \in (0, \infty)$ die Ungleichung von Tschebyschew benutzen, um ein einfaches, aber grobes Konfidenzintervall

für μ zu konstruieren:

$$\begin{aligned}
 P(|\bar{X}_n - \mu| > \varepsilon) &\leq \frac{\text{Var } \bar{X}_n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} = \alpha \\
 \Rightarrow \text{ für } \varepsilon = \frac{\sigma}{\sqrt{n\alpha}} \text{ gilt: } 1 - \alpha &\leq P\left(|\bar{X}_n - \mu| \leq \varepsilon\right) \\
 &= P\left(-\frac{\sigma}{\sqrt{n\alpha}} \leq -\bar{X}_n + \mu \leq \frac{\sigma}{\sqrt{n\alpha}}\right) \\
 &= P\left(\bar{X}_n - \frac{\sigma}{\sqrt{n\alpha}} \leq \mu \leq \bar{X}_n + \frac{\sigma}{\sqrt{n\alpha}}\right).
 \end{aligned}$$

Das Konfidenzintervall $\left[\bar{X}_n - \frac{\sigma}{\sqrt{n\alpha}}, \bar{X}_n + \frac{\sigma}{\sqrt{n\alpha}}\right]$ für μ bei bekannter Varianz σ^2 ist verteilungsunabhängig, da keinerlei Annahmen über die Verteilung von X_i gemacht wurden.

Präzisere Konfidenzintervalle können bei der Verwendung folgender *Ungleichung von Hoeffding* konstruiert werden:

Satz 9.4 (Ungleichung von Hoeffding) Es seien Y_1, \dots, Y_n unabhängige Zufallsvariablen mit $E Y_i = 0, a_i \leq Y_i \leq b_i$ fast sicher, $i = 1, \dots, n$. Für alle $\varepsilon > 0$ gilt

$$P\left(\sum_{i=1}^n Y_i \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

(ohne Beweis).

Nehmen wir z.B. an, daß X_1, \dots, X_n unabhängige, identisch verteilte Zufallsvariablen sind, $X_i \sim \text{Bernoulli}(p)$, $p \in (0, 1)$. Wir wollen ein Konfidenzintervall für p bestimmen.

Folgerung 9.5 Es seien X_1, \dots, X_n unabhängige Bernoulli(p)-verteilte Zufallsvariablen. Dann gilt

$$P(|\bar{X}_n - p| > \varepsilon) \leq 2e^{-2n\varepsilon^2}, \quad \varepsilon > 0.$$

Beweis Es gilt

$$\bar{X}_n - p = \frac{1}{n} \sum_{i=1}^n \underbrace{(X_i - p)}_{Y_i}, \quad Y_i \in [-p, 1-p],$$

das heißt $a_i = -p$, $b_i = 1-p$, $b_i - a_i = 1$, $i = 1, \dots, n$, $E Y_i = p - p = 0$.

Dann gilt:

$$\begin{aligned}
 P_p(|\bar{X}_n - p| > \varepsilon) &\leq P_p\left(\left|\sum_{i=1}^n Y_i\right| \geq \varepsilon n\right) \\
 &= P_p\left(\sum_{i=1}^n Y_i \geq \varepsilon n\right) + P_p\left(\sum_{i=1}^n (-Y_i) \geq \varepsilon n\right) \\
 &\stackrel{\text{(Satz 9.4)}}{\leq} 2e^{-\frac{2\varepsilon^2 n^2}{n}} = 2e^{-2\varepsilon^2 n},
 \end{aligned}$$

wobei man den Satz 9.4 sowohl für die Folge $\{Y_i\}$ als auch $\{-Y_i\}$ anwendet. Damit ist die Behauptung bewiesen. \square

Bemerkung 9.6 Die Form der Ungleichung von Hoeffding ähnelt sehr der von Dvoretzky-Kiefer-Wolfowitz, Satz 3.3.10.

Nun fixieren wir $\alpha > 0$ und wählen $\varepsilon_n = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}$. Durch Anwendung von Folgerung 9.5 mit diesem ε_n erhalten wir $P_p(|\bar{X}_n - p| > \varepsilon_n) \leq \alpha$, somit $P_p(|\bar{X}_n - p| \leq \varepsilon_n) \geq 1 - \alpha$ und darum ist $[\bar{X}_n - \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}, \bar{X}_n + \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}]$ ein Konfidenzintervall für p zum Niveau $1 - \alpha$.

9.2.3 Asymptotische Konfidenzintervalle

Die Philosophie der Konstruktion von asymptotischen Konfidenzintervallen ist relativ einfach: Wir erläutern sie am Beispiel eines asymptotisch normalverteilten Schätzers $\hat{\theta}$ für einen Parameter θ .

Sei (X_1, \dots, X_n) eine Stichprobe von unabhängigen und identisch verteilten Zufallsvariablen, $X_i \sim F_\theta$, $\theta \in \Theta \subseteq \mathbb{R}$. Sei $\hat{\theta}_n = \hat{\theta}(X_1, \dots, X_n)$ ein Schätzer für θ , der asymptotisch normalverteilt ist. Dann gilt für erwartungstreue $\hat{\theta}_n$

$$\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \xrightarrow{d} Y \sim N(0, 1),$$

wobei $\hat{\sigma}_n$ ein konsistenter Schätzer der asymptotischen Varianz von $\hat{\theta}_n$ ist.

$$\begin{aligned}
 &\lim_{n \rightarrow \infty} P_\theta\left(z_{\alpha/2} \leq \frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \leq z_{1-\alpha/2}\right) \\
 &= \lim_{n \rightarrow \infty} P_\theta\left(\theta \in [\hat{\theta}_n - z_{1-\alpha/2}\hat{\sigma}_n, \hat{\theta}_n + z_{1-\alpha/2}\hat{\sigma}_n]\right) = 1 - \alpha.
 \end{aligned}$$

Somit ist $[\hat{\theta}_n - z_{1-\alpha/2}\hat{\sigma}_n, \hat{\theta}_n + z_{1-\alpha/2}\hat{\sigma}_n]$ ein asymptotisches Konfidenzintervall für θ zum Niveau $1 - \alpha$.

Diese Vorgehensweise werden wir jetzt anhand von zwei Beispielen klar machen:

- **Bernoulli-Verteilung:**

Seien $X_i \sim \text{Bernoulli}(p)$ -verteilt, $i = 1, \dots, n$. Dann gilt $\theta = p$, $\hat{\theta}_n = \hat{p}_n = \bar{X}_n$, $E_p \hat{p}_n = p$, $\text{Var}_p \hat{p}_n = \frac{p(1-p)}{n}$. Wir wählen $\hat{\sigma}^2 = \frac{1}{n} \hat{p}_n(1-\hat{p}_n) = \frac{\bar{X}_n}{n}(1-\bar{X}_n)$ als Plug-In-Schätzer für σ^2 . Dann gilt nach dem zentralen Grenzwertsatz (Satz 7.2.1, WR) und dem Satz von Slutsky (Satz 6.4.2, 3. WR):

$$\sqrt{n} \frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1),$$

das heißt $p \in \left[\bar{X}_n - z_{1-\alpha/2} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{1-\alpha/2} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right]$ stellt ein asymptotisches Konfidenzintervall für p zum Niveau $1 - \alpha$ dar. Da aber $p \in [0, 1]$ sein soll, betrachtet man

$$\underline{p}(X_1, \dots, X_n) = \max \left\{ 0, \bar{X}_n - z_{1-\alpha/2} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right\}$$

und

$$\bar{p}(X_1, \dots, X_n) = \min \left\{ 1, \bar{X}_n + z_{1-\alpha/2} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right\}.$$

Bemerkung 9.7 Ein anderes asymptotisches Konfidenzintervall für den Parameter p der Bernoulli-Verteilung bekommt man, wenn man die Aussage des zentralen Grenzwertsatzes

$\lim_{n \rightarrow \infty} P_p \left(-z_{1-\alpha/2} \leq \sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \leq z_{1-\alpha/2} \right) = 1 - \alpha$ nimmt und die quadratische Ungleichung dann bezüglich p auflöst.

Übungsaufgabe 9.8 Lösen Sie die Ungleichung auf!

- **Poissonverteilung:**

Es seien $X_i \sim \text{Poisson}(\lambda)$, $i = 1, \dots, n$, dann gilt $\theta = \lambda$, $\hat{\theta}_n = \hat{\lambda} = \bar{X}_n$. Da $E_\lambda X_i = \text{Var}_\lambda X_i = \lambda$, kann man den zentralen Grenzwertsatz (Satz 7.2.1, WR) anwenden

$$\sqrt{n} \frac{\bar{X}_n - \lambda}{\sqrt{\lambda}} \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1),$$

Da \bar{X}_n stark konsistent für λ ist, gilt nach dem Satz von Slutsky (Satz 6.4.2, 4, WR)

$$\sqrt{n} \frac{\bar{X}_n - \lambda}{\sqrt{\bar{X}_n}} \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1).$$

Daraus folgt ein asymptotisches Konfidenzintervall

$$\left[\bar{X}_n - z_{1-\alpha/2} \sqrt{\frac{\bar{X}_n}{n}}, \bar{X}_n + z_{1-\alpha/2} \sqrt{\frac{\bar{X}_n}{n}} \right]$$

für den Parameter λ zum Konfidenzniveau $1 - \alpha$.

Bemerkung 9.9 1. Ähnlich wie in Bemerkung 9.7 angegeben, kann man durch Auflösen der quadratischen Ungleichung in

$$\lim_{n \rightarrow \infty} P_\lambda \left(\sqrt{n} \frac{\bar{X}_n - \lambda}{\sqrt{\lambda}} \in [-z_{1-\alpha/2}, z_{1-\alpha/2}] \right) = 1 - \alpha$$

bezüglich λ ein alternatives asymptotisches Konfidenzintervall für λ angeben.

Übungsaufgabe 9.10 Bitte führen Sie diese Berechnungen durch.

2. Da $\lambda > 0$ ist, kann man die untere Schranke diesbezüglich korrigieren:

$$\underline{\lambda}(X_1, \dots, X_n) = \max \left\{ 0, \bar{X}_n - z_{1-\alpha/2} \sqrt{\frac{\bar{X}_n}{n}} \right\}$$

9.3 Zwei-Stichproben-Probleme

In diesem Abschnitt werden Charakteristiken bzw. Parameter von zwei unterschiedlichen Stichproben miteinander verglichen, indem man Konfidenzintervalle für einfache Funktionen dieser Parameter konstruiert.

Betrachten wir zwei Zufallsstichproben $Y_1 = (X_{11}, \dots, X_{1n_1})$ und $Y_2 = (X_{21}, \dots, X_{2n_2})$ von Zufallsvariablen X_{i1}, \dots, X_{in_i} , $i = 1, 2$, die innerhalb der Stichprobe Y_i jeweils unabhängig und identisch verteilt sind, $X_{ij} \stackrel{d}{=} X_i$, $j = 1, \dots, n_i$, $i = 1, 2$ und die Prototyp-Zufallsvariable $X_i \sim F_{\theta_i}$, $\theta_i \in \Theta \subset \mathbb{R}^m$. Es wird im Allgemeinen nicht gefordert, daß Y_1 und Y_2 unabhängig sind. Falls sie voneinander abhängen, spricht man von *verbundenen Stichproben* Y_1 und Y_2 . Betrachten wir eine Funktion $g : \mathbb{R}^{2m} \rightarrow \mathbb{R}$ von den Parametervektoren θ_1 und θ_2 . In diesem Skript werden dabei meistens die Fälle $m = 1, 2$, $g(\theta_1, \theta_2) = \theta_{1j} - \theta_{2j}$, $g(\theta_1, \theta_2) = \frac{\theta_{1j}}{\theta_{2j}}$ untersucht, wobei $\theta_i = (\theta_{i1}, \dots, \theta_{im})$, $i = 1, 2$.

Unsere Zielstellung wird sein, ein (möglicherweise asymptotisches) Konfidenzintervall für $g(\theta_1, \theta_2)$ mit Hilfe der Stichprobe (Y_1, Y_2) zu gewinnen.

Dabei wird dieselbe Philosophie wie in Abschnitt 4.1 beschrieben verfolgt. Es wird eine Statistik $T(Y_1, Y_2, g(\theta_1, \theta_2))$ gesucht, die eine (möglicherweise asymptotische) Prüfverteilung F besitzt und von $g(\theta_1, \theta_2)$ explizit abhängt.

Durch das Auflösen der Ungleichung $F_{\alpha_1}^{-1} \leq T(Y_1, Y_2, g(\theta_1, \theta_2)) \leq F_{1-\alpha_2}^{-1}$ bzgl. $g(\theta_1, \theta_2)$ bekommt man dann ein (möglicherweise asymptotisches) Konfidenzintervall zum Niveau $1 - \alpha$, $\alpha = \alpha_1 + \alpha_2$.

9.3.1 Normalverteilte Stichproben

Hier wird angenommen, daß $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2$.

Konfidenzintervall für die Differenz $\mu_1 - \mu_2$ bei bekannten Varianzen σ_1^2 und σ_2^2 und unabhängigen Stichproben

Seien Y_1 und Y_2 voneinander unabhängig und σ_1^2, σ_2^2 bekannt. Wir betrachten die Parameterfunktion $g(\mu_1, \mu_2) = \mu_1 - \mu_2$. Es seien $\bar{X}_{in_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$, $i = 1, 2$ die Stichprobenmittel der Stichproben Y_1 und Y_2 . Es gilt $\bar{X}_{in_i} \sim N(\mu_i, \frac{\sigma_i^2}{n_i})$, $i = 1, 2$. Nach Satz 3.3.3, 4) sind \bar{X}_{1n_1} und \bar{X}_{2n_2} unabhängig. Dann ist wegen der Faltungsstabilität der Normalverteilung $\bar{X}_{1n_1} - \bar{X}_{2n_2} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$. Nach dem Normieren erhält man die Statistik $T(Y_1, Y_2, \mu_1 - \mu_2) = \frac{\bar{X}_{1n_1} - \bar{X}_{2n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$. Daraus bekommt man das Konfidenzintervall

$$\left[\bar{X}_{1n_1} - \bar{X}_{2n_2} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{X}_{1n_1} - \bar{X}_{2n_2} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

für $\mu_1 - \mu_2$ zum Niveau $1 - \alpha$.

Konfidenzintervall für den Quotienten $\frac{\sigma_1^2}{\sigma_2^2}$ bei unbekannten Erwartungswerten μ_1 und μ_2 und unabhängigen Stichproben

Seien Y_1 und Y_2 stochastisch unabhängig voneinander. Sei $g(\sigma_1, \sigma_2) = \sigma_1^2 / \sigma_2^2$. Wir konstruieren die Statistik $T(Y_1, Y_2, \sigma_1^2 / \sigma_2^2)$ folgendermaßen:

Seien $S_{in_i}^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{in_i})^2$, $i = 1, 2$ die Stichprobenvarianzen der Stichproben Y_1 und Y_2 . Dann gilt $\frac{(n_i-1)S_{in_i}^2}{\sigma_i^2} \sim \chi_{n_i-1}^2$, $i = 1, 2$ nach Satz 3.3.5.

Da die $S_{in_i}^2$ voneinander unabhängig sind, gilt

$$T\left(Y_1, Y_2, \frac{\sigma_1^2}{\sigma_2^2}\right) = \frac{\frac{(n_2-1)S_{2n_2}^2}{(n_2-1)\sigma_2^2}}{\frac{(n_1-1)S_{1n_1}^2}{(n_1-1)\sigma_1^2}} = \frac{S_{2n_2}^2}{S_{1n_1}^2} \cdot \frac{\sigma_1^2}{\sigma_2^2} \sim F_{n_2-1, n_1-1}$$

nach der Definition der F -Verteilung. Daraus ergibt sich das Konfidenzintervall

$$\left[\frac{S_{1n_1}^2}{S_{2n_2}^2} F_{n_2-1, n_1-1, \alpha_1}, \frac{S_{1n_1}^2}{S_{2n_2}^2} F_{n_2-1, n_1-1, 1-\alpha_2} \right]$$

für $\frac{\sigma_1^2}{\sigma_2^2}$ zum Niveau $1 - \alpha$.

Konfidenzintervall für die Differenz $\mu_1 - \mu_2$ der Erwartungswerte bei verbundenen Stichproben

Dieses Mal seien Y_1 und Y_2 verbunden, $X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma^2)$ für ein unbekanntes $\sigma^2 > 0$, $n_1 = n_2 = n$. Da $X_{ij}, j = 1, \dots, n$ unabhängig und identisch verteilt sind, gilt $Z_j = X_{1j} - X_{2j} \sim N(\mu_1 - \mu_2, \sigma^2)$, $j = 1, \dots, n$.

Unser Ziel ist es, ein Konfidenzintervall für $\mu_1 - \mu_2$ zu bekommen. Wenn wir die Stichprobe (Z_1, \dots, Z_n) betrachten, und Ergebnisse des Abschnittes 4.2.1, 2. anwenden, so erhalten wir sofort folgendes Konfidenzintervall:

$$\left[\bar{Z}_n - t_{n-1, 1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}, \bar{Z}_n + t_{n-1, 1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \right]$$

für $\mu_1 - \mu_2$ zum Niveau $1 - \frac{\alpha}{2}$, wobei $\bar{Z}_n = \frac{1}{n} \sum_{j=1}^n Z_j = \frac{1}{n} \sum_{j=1}^n (X_{1j} - X_{2j}) = \bar{X}_{1n} - \bar{X}_{2n}$, $S_n^2 = \frac{1}{n-1} \sum_{j=1}^n (Z_j - \bar{Z}_n)^2 = \frac{1}{n-1} \sum_{j=1}^n (X_{1j} - X_{2j} - \bar{X}_{1n} + \bar{X}_{2n})^2$

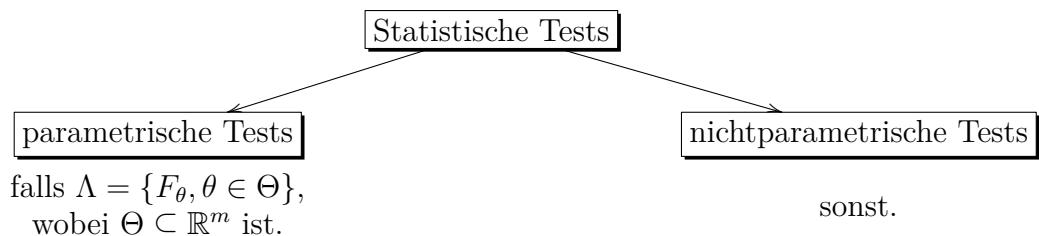
Kapitel 10

Tests statistischer Hypothesen

In der Vorlesung Stochastik I haben wir schon Beispiele von statistischen Tests kennengelernt, wie etwa den Kolmogorow-Smirnow-Test (vergleiche Bemerkung 3.3.38, 3), Skript Stochastik I). Jetzt sollen statistische Signifikanztests formal eingeführt und ihre Eigenschaften untersucht werden.

10.1 Allgemeine Philosophie des Testens

Es sei eine Zufallsstichprobe (X_1, \dots, X_n) von unabhängigen, identisch verteilten Zufallsvariablen X_i gegeben, mit Verteilungsfunktion $F \in \Lambda$, wobei Λ eine Klasse von Verteilungsfunktionen ist. Es sei (x_1, \dots, x_n) eine konkrete Stichprobe, die als Realisierung von (X_1, \dots, X_n) interpretiert wird. In der Theorie des statistischen Testens werden Hypothesen über die Beschaffenheit der (unbekannten) Verteilungsfunktion F gestellt und geprüft. Dabei unterscheidet man



Bei parametrischen Tests prüft man, ob der Parameter θ bestimmte Werte annimmt (zum Beispiel $\theta = 0$). Bekannte Beispiele von nichtparametrischen Tests sind Anpassungstests, bei denen man prüft, ob die Verteilungsfunktion F gleich einer vorgegebenen Funktion F_0 ist.

Formalisieren wir zunächst den Begriff *Hypothese*. Die Menge Λ von zulässigen Verteilungsfunktionen F wird in zwei disjunkte Teilmengen Λ_0 und Λ_1 zerlegt, $\Lambda_0 \cup \Lambda_1 = \Lambda$. Die Aussage

„Teste die *Haupthypothese* $H_0 : F \in \Lambda_0$ gegen die *Alternative* $H_1 : F \in \Lambda_1$ “ bedeutet, dass man anhand der konkreten Stichprobe (x_1, \dots, x_n) versucht, eine Entscheidung zu fällen, ob die Verteilungsfunktion der Zufallsvariable X_i zu Λ_0 oder zu Λ_1 gehört. Dies passiert auf Grund einer statistischen *Entscheidungsregel*

$$\varphi : \mathbb{R}^n \rightarrow [0, 1],$$

die eine Statistik mit folgender Interpretation ist:

Der Stichprobenraum \mathbb{R}^n wird in drei disjunkte Bereiche K_0, K_{01} und K_1 unterteilt, sodass $\mathbb{R}^n = K_0 \cup K_{01} \cup K_1$, wobei

$$\begin{aligned} K_0 &= \varphi^{-1}(\{0\}) &= \{x \in \mathbb{R}^n : \varphi(x) = 0\}, \\ K_1 &= \varphi^{-1}(\{1\}) &= \{x \in \mathbb{R}^n : \varphi(x) = 1\}, \\ K_{01} &= \varphi^{-1}((0, 1)) &= \{x \in \mathbb{R}^n : 0 < \varphi(x) < 1\}. \end{aligned}$$

Dementsprechend wird $H_0 : F \in \Lambda_0$

- verworfen, falls $\varphi(x) = 1$, also $x \in K_1$,
- nicht verworfen, falls $\varphi(x) = 0$, also $x \in K_0$;
- falls $\varphi(x) \in (0, 1)$, also $x \in K_{01}$, wird $\varphi(x)$ als Bernoulli-Wahrscheinlichkeit interpretiert, und es wird eine Zufallsvariable $Y \sim \text{Bernoulli}(\varphi(x))$ generiert, für die gilt:

$$Y = \begin{cases} 1 & \Rightarrow H_0 \text{ wird verworfen} \\ 0 & \Rightarrow H_0 \text{ wird nicht verworfen} \end{cases}$$

Falls $K_{01} \neq \emptyset$, wird eine solche Entscheidungsregel *randomisiert* genannt. Bei $K_{01} = \emptyset$, also $\mathbb{R}^n = K_0 \cup K_1$ spricht man dagegen von *nicht-randomisierten* Tests. Dabei heißt K_0 bzw. K_1 *Annahmebereich* bzw. *Ablehnungsbereich* (*kritischer Bereich*) von H_0 . K_{01} heißt *Randomisierungsbereich*.

Bemerkung 10.1 1. Man sagt absichtlich „ H_0 wird nicht verworfen“, statt „ H_0 wird akzeptiert“, weil die schließende Statistik generell keine positiven, sondern nur negative Entscheidungen treffen kann. Dies ist generell ein philosophisches Problem der Falsifizierbarkeit von Hypothesen oder wissenschaftlichen Theorien, von denen aber keiner behaupten kann, dass sie der Wahrheit entsprechen (vergleiche die *wissenschaftliche Erkenntnistheorie von Karl Popper (1902-1994)*).

2. Die randomisierten Tests sind hauptsächlich von theoretischem Interesse (vergleiche Abschnitt 2.3). In der Praxis werden meistens nicht-randomisierte Regeln verwendet, bei denen man aus der Stichprobe (x_1, \dots, x_n) allein die Entscheidung über H_0 treffen kann. Hier gilt $\varphi(x) = \mathbb{1}_{K_1}, x = (x_1, \dots, x_n) \in \mathbb{R}^n$.

In diesem und in folgendem Abschnitt betrachten wir ausschließlich nicht-randomisierte Tests, um in Abschnitt 2.3 zu der allgemeinen Situation zurückzukehren.

Definition 10.2 Man sagt, dass die nicht-randomisierte Testregel $\varphi : \mathbb{R}^n \rightarrow \{0, 1\}$ einen (*nichtrandomisierten*) *statistischen Test zum Signifikanzniveau α* angibt, falls für $F \in \Lambda_0$ gilt

$$P_F(\varphi(X_1, \dots, X_n) = 1) = P(H_0 \text{ verwerfen} \mid H_0 \text{ richtig}) \leq \alpha.$$

Definition 10.3 1. Wenn man H_0 verwirft, obwohl H_0 richtig ist, begeht man den sogenannten *Fehler 1. Art*. Die Wahrscheinlichkeit

$$\alpha_n(F) = P_F(\varphi(x_1, \dots, x_n) = 1), \quad F \in \Lambda_0$$

heißt die *Wahrscheinlichkeit des Fehlers 1. Art* und soll unter dem Niveau α bleiben.

2. Den *Fehler 2. Art* begeht man, wenn man die falsche Hypothese H_0 nicht verwirft. Dabei ist

$$\beta_n(F) = P_F(\varphi(x_1, \dots, x_n) = 0), \quad F \in \Lambda_1$$

die *Wahrscheinlichkeit des Fehlers 2. Art*.

Eine Zusammenfassung aller Möglichkeiten wird in folgender Tabelle festgehalten:

	H_0 richtig	H_0 falsch
H_0 verwerfen	Fehler 1. Art, Wahrscheinlichkeit $\alpha_n(F) \leq \alpha$	richtige Entscheidung
H_0 nicht verwerfen	richtige Entscheidung	Fehler 2. Art mit Wahrscheinlichkeit $\beta_n(F)$

Dabei sollen α_n und β_n möglichst klein sein, was gegenläufige Tendenzen darstellt, weil beim Kleinwerden von α die Wahrscheinlichkeit des Fehlers 2. Art notwendigerweise wächst.

Definition 10.4 1. Die Funktion

$$G_n(F) = P_F(\varphi(X_1, \dots, X_n) = 1), \quad F \in \Lambda$$

heißt *Gütfunktion* eines Tests φ .

2. Die Einschränkung von G_n auf Λ_1 heißt *Stärke*, *Schärfe* oder *Macht* (englisch *power*) des Tests φ .

Es gilt

$$\begin{cases} G_n(F) = \alpha_n(F) \leq \alpha, F \in \Lambda_0 \\ G_n(F) = 1 - \beta_n(F), F \in \Lambda_1 \end{cases}$$

Beispiel 10.5 Parametrische Tests. Wie sieht ein parametrischer Test aus? Der Parameterraum Θ wird als $\Theta_0 \cup \Theta_1$ dargestellt, wobei $\Theta_0 \cap \Theta_1 = \emptyset$. Es gilt $\Lambda_0 = \{F_\theta : \theta \in \Theta_0\}$, $\Lambda_1 = \{F_\theta : \theta \in \Theta_1\}$. P_F wird zu P_θ , α_n , G_n und β_n werden statt auf Λ auf Θ definiert.

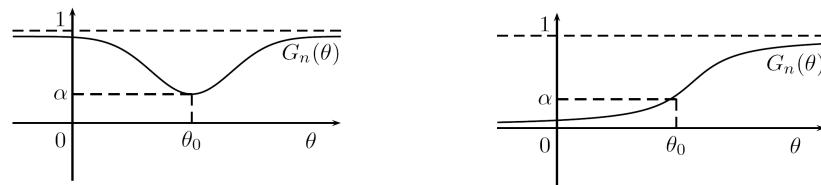
Welche Hypothesen H_0 und H_1 kommen oft bei parametrischen Tests vor? Zur Einfachheit betrachten wir den Spezialfall $\Theta = \mathbb{R}$.

1. $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$
2. $H_0 : \theta \geq \theta_0$ vs. $H_1 : \theta < \theta_0$
3. $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$
4. $H_0 : \theta \in [a, b]$ vs. $H_1 : \theta \notin [a, b]$

Im Fall (1) heißt der parametrische Test *zweiseitig*, in den Fällen (2) und (3) *einseitig (rechts- bzw. linksseitig)*. In Fall (4) spricht man von der *Intervallhypothese* H_0 .

Bei einem zweiseitigen bzw. einseitigen Test kann die Gütfunktion wie in Abbildung 10.1 (a) bzw. 10.1 (b) aussehen,

Abbildung 10.1: Gütfunktion

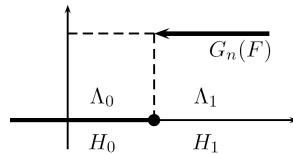


(a) eines zweiseitigen Tests

(b) eines einseitigen Tests

Bei einem allgemeinen (nicht notwendigerweise parametrischen) Modell kann man die ideale Gütfunktion wie in Abbildung 10.2 schematisch darstellen.

Abbildung 10.2: Schematische Darstellung der idealen Gütfunktion



- Man sieht aus Definition 10.3, dem Fehler 1. und 2. Art und der Ablehnungsregel, dass die Hypothesen H_0 und H_1 nicht symmetrisch behandelt werden, denn nur die Wahrscheinlichkeit des Fehlers 1. Art wird kontrolliert. Dies ist der Grund dafür, dass Statistiker die eigentlich interessierende Hypothese nicht als H_0 , sondern als H_1 formulieren, damit, wenn man sich für H_1 entscheidet, man mit Sicherheit sagen kann, dass die Wahrscheinlichkeit der Fehlentscheidung unter dem Niveau α liegt.
- Wie wird ein statistischer, nicht randomisierter Test praktisch konstruiert? Die Konstruktion der Ablehnungsregel φ ähnelt sich sehr der von Konfidenzintervallen:
 1. Finde eine Teststatistik $T : \mathbb{R}^n \rightarrow \mathbb{R}$, die unter H_0 eine (möglicherweise asymptotisch für $n \rightarrow \infty$) bestimmte Prüfverteilung hat.
 2. Definiere $B_0 = [t_{\alpha_1}, t_{1-\alpha_2}]$, wobei t_{α_1} und $t_{1-\alpha_2}$ Quantile der Prüfverteilung von T sind, $\alpha_1 + \alpha_2 = \alpha \in [0, 1]$.
 3. Falls $T(X_1, \dots, X_n) \in \mathbb{R} \setminus B_0 = B_1$, setze $\varphi(X_1, \dots, X_n) = 1$. H_0 wird verworfen. Ansonsten setze $\varphi(X_1, \dots, X_n) = 0$.
- Falls die Verteilung von T nur asymptotisch bestimmt werden kann, so heißt φ *asymptotischer Test*.
- Sehr oft aber ist auch die asymptotische Verteilung von T nicht bekannt. Dann verwendet man sogenannte *Monte-Carlo Tests*, in denen dann Quantile t_α näherungsweise aus sehr vielen Monte-Carlo-Simulationen von T (unter H_0) bestimmt werden: Falls t^i , $i = 1, \dots, m$ die Werte von T in m unabhängigen Simulationsvorgängen sind, das heißt $t^i = T(x_1^i, \dots, x_n^i)$, x_j^i sind unabhängige Realisierungen von

$X_j \sim F \in \Lambda_0$, $j = 1, \dots, n$, $i = 1, \dots, m$ dann bildet man ihre Ordnungsstatistiken $t^{(1)}, \dots, t^{(m)}$ und setzt $t_\alpha \approx t^{(\lfloor \alpha \cdot m \rfloor)}$, $\alpha \in [0, 1]$, wobei $t^{(0)} = -\infty$.

Bemerkung 10.6 Man sieht deutlich, dass aus einem beliebigen Konfidenzintervall

$$I_\theta = [I_1^\theta(X_1, \dots, X_n), I_2^\theta(X_1, \dots, X_n)]$$

zum Niveau $1 - \alpha$ für einen Parameter $\theta \in \mathbb{R}$ ein Test für θ konstruierbar ist. Die Hypothese $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ wird mit folgender Entscheidungsregel getestet:

$$\varphi(X_1, \dots, X_n) = 1, \text{ falls } \theta_0 \notin [I_1^{\theta_0}(X_1, \dots, X_n), I_2^{\theta_0}(X_1, \dots, X_n)].$$

Das Signifikanzniveau des Tests ist α .

Beispiel 10.7 *Normalverteilung, Test des Erwartungswertes bei bekannter Varianz.* Es seien

$$X_1, \dots, X_n \sim N(\mu, \sigma^2)$$

mit bekannter Varianz σ^2 . Ein Konfidenzintervall für μ ist

$$I^\mu = [I_1^\mu(X_1, \dots, X_n), I_2^\mu(X_1, \dots, X_n)] = \left[\bar{X}_n - \frac{z_{1-\alpha/2} \cdot \sigma}{\sqrt{n}}, \bar{X}_n + \frac{z_{1-\alpha/2} \cdot \sigma}{\sqrt{n}} \right]$$

(vergleiche Stochastik I, 4.2.1) H_0 wird verworfen, falls $|\mu_0 - \bar{X}_n| > \frac{z_{1-\alpha/2} \cdot \sigma}{\sqrt{n}}$. In der Testsprache bedeutet es, dass

$$\varphi(x_1, \dots, x_n) = \mathbb{1}((x_1, \dots, x_n) \in K_1),$$

wobei

$$K_1 = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : |\mu_0 - \bar{x}_n| > \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}} \right\}$$

der Ablehnungsbereich ist. Für die Teststatistik $T(X_1, \dots, X_n)$ gilt:

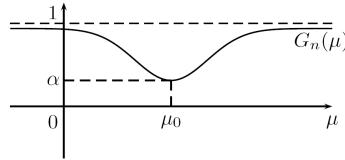
$$T(X_1, \dots, X_n) = \frac{\bar{X}_n - \mu_0}{\sigma} \sqrt{n} \sim N(0, 1) \mid \text{unter } H_0,$$

$$\alpha_n(\mu) = \alpha.$$

Berechnen wir nun die Gütfunktion (vergleiche Abbildung 10.3).

$$\begin{aligned}
 G_n(\mu) &= P_\mu \left(|\mu_0 - \bar{X}_n| > \frac{z_{1-\alpha/2}}{\sqrt{n}} \right) = 1 - P_\mu \left(\left| \bar{X}_n - \mu_0 \right| \leq \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}} \right) \\
 &= 1 - P_\mu \left(\left| \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} + \frac{\mu - \mu_0}{\sigma} \sqrt{n} \right| \leq z_{1-\alpha/2} \right) \\
 &= 1 - P_\mu \left(-z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma} \sqrt{n} \leq \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \leq z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma} \sqrt{n} \right) \\
 &= 1 - \Phi \left(z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma} \sqrt{n} \right) + \Phi \left(-z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma} \sqrt{n} \right) \\
 &= \Phi \left(-z_{1-\alpha/2} + \frac{\mu - \mu_0}{\sigma} \sqrt{n} \right) + \Phi \left(-z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma} \sqrt{n} \right).
 \end{aligned}$$

Abbildung 10.3: Gütfunktion für den zweiseitigen Test des Erwartungswertes einer Normalverteilung bei bekannter Varianz



Die „Ja-Nein“-Entscheidung des Testens wird oft als zu grob empfunden. Deswegen versucht man, ein feineres Maß der Verträglichkeit der Daten mit den Hypothesen H_0 und H_1 zu bestimmen. Dies ist der sogenannte p -Wert, der von den meisten Statistik-Softwarepaketen ausgegeben wird.

Definition 10.8 Es sei (x_1, \dots, x_n) die konkrete Stichprobe von Daten, die als Realisierung von (X_1, \dots, X_n) interpretiert wird und $T(X_1, \dots, X_n)$ die Teststatistik, mit deren Hilfe die Entscheidungsregel φ konstruiert wurde. Der p -Wert des statistischen Tests φ ist das kleinste Signifikanzniveau, zu dem der Wert $t = T(x_1, \dots, x_n)$ zur Verwerfung der Hypothese H_0 führt.

Im Beispiel eines einseitigen Tests mit dem Ablehnungsbereich $B_1 = (t, \infty)$ sagt man grob, dass

$$p = „P(T(X_1, \dots, X_n) \geq t \mid H_0)“,$$

wobei die Anführungszeichen bedeuten, dass dies keine klassische, sondern eine bedingte Wahrscheinlichkeit ist, die später präzise angegeben wird.

Bei der Verwendung des p -Wertes verändert sich die Ablehnungsregel: die Hypothese H_0 wird zum Signifikanzniveau α abgelehnt, falls $\alpha \geq p$.

Früher hat man die Signifikanz der Testentscheidung (Ablehnung von H_0) anhand folgender Tabelle festgesetzt:

p -Wert	Interpretation
$p \leq 0,001$	sehr stark signifikant
$0,001 < p \leq 0,01$	stark signifikant
$0,01 < p \leq 0,05$	schwach signifikant
$0,05 < p$	nicht signifikant

Da aber heute der p -Wert an sich verwendet werden kann, kann der Anwender der Tests bei vorgegebenem p -Wert selbst entscheiden, zu welchem Niveau er seine Tests durchführen will.

Bemerkung 10.9 1. Das Signifikanzniveau darf nicht in Abhängigkeit von p festgelegt werden. Dies würde die allgemeine Testphilosophie zerstören!

2. Der p -Wert ist keine Wahrscheinlichkeit, sondern eine Zufallsvariable, denn er hängt von (X_1, \dots, X_n) ab. Der Ausdruck $p = P(T(X_1, \dots, X_n) \geq t | H_0)$, der in Definition 10.8 für den p -Wert eines einseitigen Tests mit Teststatistik T gegeben wurde, soll demnach als *überschreitungswahrscheinlichkeit* interpretiert werden, dass bei Wiederholung des Zufallsexperiments unter H_0 der Wert $t = T(x_1, \dots, x_n)$ oder extremere Werte in Richtung der Hypothese H_1 betrachtet werden:

$$p = P(T(X'_1, \dots, X'_n) \geq T(x_1, \dots, x_n) | H_0),$$

wobei $(X'_1, \dots, X'_n) \stackrel{d}{=} (X_1, \dots, X_n)$. Falls wir von einer konkreten Realisierung (x_1, \dots, x_n) zur Zufallsstichprobe (X_1, \dots, X_n) übergehen, erhalten wir

$$p = p(X_1, \dots, X_n) = P(T(X'_1, \dots, X'_n) \geq T(X_1, \dots, X_n) | H_0)$$

3. Für andere Hypothesen H_0 wird der p -Wert auch eine andere Form haben. Zum Beispiel für

(a) einen symmetrischen zweiseitigen Test ist

$$B_0 = [-t_{1-\alpha/2}, t_{1-\alpha/2}]$$

der Akzeptanzbereich für H_0 .

$$\Rightarrow p = P(|T(X'_1, \dots, X'_n)| \geq t | H_0), t = |T(X_1, \dots, X_n)|$$

(b) einen rechtsseitigen Test mit $B_0 = [t_\alpha, \infty]$ gilt

$$p = P(T(X'_1, \dots, X'_n) \leq t | H_0), t = T(X_1, \dots, X_n)$$

(c) Das Verhalten des p -Wertes kann folgendermaßen untersucht werden:

Lemma 10.10 Falls die Verteilungsfunktion F von X_i stetig und streng monoton steigend ist (die Verteilung von T ist absolut stetig mit zum Beispiel stetiger Dichte), dann ist $p \sim U[0, 1]$.

Beweis Wir zeigen es am speziellen Beispiel des rechtsseitigen Tests.

$$\begin{aligned} P(p \leq \alpha | H_0) &= P(\bar{F}_T(T(X_1, \dots, X_n)) \leq \alpha | H_0) \\ &= P(F_T(T(X_1, \dots, X_n)) \geq 1 - \alpha | H_0) \\ &= P(U \geq 1 - \alpha) = 1 - (1 - \alpha) = \alpha, \quad \alpha \in [0, 1], \end{aligned}$$

da $F_T(T(X_1, \dots, X_n)) \stackrel{d}{=} U \sim U[0, 1]$ und F_T absolut stetig ist. \square

Übungsaufgabe 10.11 Zeigen Sie, dass für eine beliebige Zufallsvariable X mit absolut stetiger Verteilung und streng monoton steigender Verteilungsfunktion F_X gilt:

$$F_X(X) \sim U[0, 1]$$

Falls die Verteilung von T diskret ist, mit dem Wertebereich $\{t_1, \dots, t_n\}$, $t_i < t_j$ für $i < j$, so ist auch die Verteilung von p diskret, somit gilt nicht $p \sim U[0, 1]$. In diesem Fall ist $F_T(x)$ eine Treppenfunktion, die die Gerade $y = u$ in den Punkten $u = \sum_{i=1}^k P(T(X_1, \dots, X_n) = t_i)$, $k = 1, \dots, n$ berührt (vgl. Abbildung 10.4).

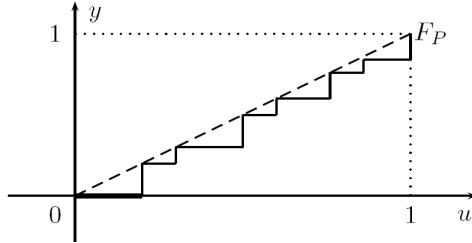
Definition 10.12 1. Falls die Macht $G_n(\cdot)$ eines Tests φ zum Niveau α die Ungleichung

$$G_n(F) \geq \alpha, \quad F \in \Lambda_1$$

erfüllt, dann heißt der Test *unverfälscht*.

2. Es seien φ und φ^* zwei Tests zum Niveau α mit Gütfunktionen $G_n(\cdot)$ und $G_n^*(\cdot)$. Man sagt, dass der Test φ *besser* als φ^* ist, falls er eine größere Macht besitzt:

$$G_n(F) \geq G_n^*(F) \quad \forall F \in \Lambda_1$$

Abbildung 10.4: Verteilung von p für diskrete T 

3. Der Test φ heißt konsistent, falls $G_n(F) \xrightarrow{n \rightarrow \infty} 1$ für alle $F \in \Lambda_1$.

Bemerkung 10.13 1. Die einseitigen Tests haben oft eine größere Macht als ihre zweiseitigen Versionen.

Beispiel 10.14 Betrachten wir zum Beispiel den Gauß-Test des Erwartungswertes der Normalverteilung bei bekannter Varianz. Beim zweiseitigen Test

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0.$$

erhalten wir die Gütfunktion

$$G_n(\mu) = \Phi\left(-z_{1-\alpha/2} + \sqrt{n}\frac{\mu - \mu_0}{\sigma}\right) + \Phi\left(-z_{1-\alpha/2} - \sqrt{n}\frac{\mu - \mu_0}{\sigma}\right).$$

Beim einseitigen Test φ^* der Hypothesen

$$H_0^* : \mu \leq \mu_0 \text{ vs. } H_1^* : \mu > \mu_0$$

ist seine Gütfunktion gleich

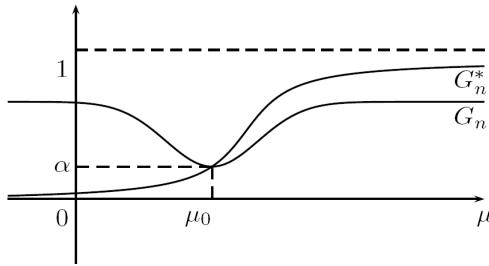
$$G_n^*(\mu) = \Phi\left(-z_{1-\alpha} + \sqrt{n}\frac{\mu - \mu_0}{\sigma}\right)$$

Beide Tests sind offensichtlich konsistent, denn $G_n(\mu) \xrightarrow{n \rightarrow \infty} 1$, $G_n^*(\mu) \xrightarrow{n \rightarrow \infty} 1$.

1. Dabei ist φ^* besser als φ . Beide Tests sind unverfälscht (vergleiche Abbildung 10.5).

2. Beim Testen einer Intervallhypothese $H_0 : \theta \in [a, b]$ vs. $H_1 : \theta \notin [a, b]$ zum Niveau α kann man wie folgt vorgehen: Teste

Abbildung 10.5: Gütfunktionen eines ein- bzw. zweiseitigen Tests der Erwartungswertes einer Normalverteilung



- (a) $H_0^a : \theta \geq a$ vs. $H_1^a : \theta < a$ zum Niveau $\alpha/2$.
- (b) $H_0^b : \theta \leq b$ vs. $H_1^b : \theta > b$ zum Niveau $\alpha/2$.

H_0 wird nicht abgelehnt, falls H_0^a und H_0^b nicht abgelehnt werden. Die Wahrscheinlichkeit des Fehlers 1. Art ist hier $\leq \alpha$. Die Macht dieses Tests ist im Allgemeinen schlecht.

3. Je mehr Parameter für den Aufbau der Teststatistik T geschätzt werden müssen, desto kleiner wird in der Regel die Macht.

10.2 Nichtrandomisierte Tests

10.2.1 Parametrische Signifikanztests

In diesem Abschnitt geben wir Beispiele einiger Tests, die meistens aus den entsprechenden Konfidenzintervallen für die Parameter von Verteilungen entstehen. Deshalb werden wir sie nur kurz behandeln.

1. Tests für die Parameter der Normalverteilung $N(\mu, \sigma^2)$

(a) Test von μ bei unbekannter Varianz

- Hypothesen: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$.
- Teststatistik:

$$T(X_1, \dots, X_n) = \frac{\bar{X}_n - \mu_0}{S_n} \sim t_{n-1} \quad | H_0$$

- Entscheidungsregel:

$$\varphi(X_1, \dots, X_n) = 1, \text{ falls } |T(X_1, \dots, X_n)| > t_{n-1, 1-\alpha/2}.$$

(b) **Test von σ^2 bei unbekanntem μ**

- Hypothesen: $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$.
- Teststatistik:

$$T(X_1, \dots, X_n) = \frac{(n-1)S_n^2}{\sigma_0^2} \sim \chi_{n-1}^2 \quad | \quad H_0,$$

$$\text{wobei } S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

- Entscheidungsregel:

$$\varphi(X_1, \dots, X_n) = 1, \text{ falls } T(X_1, \dots, X_n) \notin [\chi_{n-1, \alpha/2}^2, \chi_{n-1, 1-\alpha/2}^2].$$

Übungsaufgabe 10.15 i. Finden Sie $G_n(\cdot)$ für die einseitige Version der obigen Tests.

ii. Zeigen Sie, dass diese einseitigen Tests unverfälscht sind, die zweiseitigen aber nicht.

2. Asymptotische Tests

Bei asymptotischen Tests ist die Verteilung der Teststatistik nur näherungsweise (für große n) bekannt. Ebenso asymptotisch wird das Konfidenzniveau α erreicht. Ihre Konstruktion basiert meistens auf Verwendung der Grenzwertsätze.

Die allgemeine Vorgehensweise wird im sogenannten *Wald-Test* (genannt nach dem Statistiker Abraham Wald (1902-1980)) fixiert:

- Sei (X_1, \dots, X_n) eine Zufallsstichprobe, X_i seien unabhängig und identisch verteilt für $i = 1, \dots, n$, mit $X_i \sim F_\theta$, $\theta \in \Theta \subseteq \mathbb{R}$.
- Wir testen $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. Es sei $\hat{\theta}_n = \hat{\theta}(X_1, \dots, X_n)$ ein erwartungstreuer, asymptotisch normalverteilter Schätzer für θ .

$$\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1) \quad | \quad H_0,$$

wobei $\hat{\sigma}_n^2$ ein konsistenter Schätzer für die Varianz von $\hat{\theta}_n$ sei.

Die Teststatistik ist

$$T(X_1, \dots, X_n) = \frac{\hat{\theta}_n(X_1, \dots, X_n) - \theta_0}{\hat{\sigma}_n}.$$

- Die Entscheidungsregel lautet: H_0 wird abgelehnt, wenn $|T(X_1, \dots, X_n)| > z_{1-\alpha/2}$, wobei $z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$. Diese Entscheidungsregel soll nur bei großen n verwendet werden. Die Wahrscheinlichkeit des Fehlers 1. Art ist asymptotisch gleich α , denn $P(|T(X_1, \dots, X_n)| > z_{1-\alpha/2} \mid H_0) \xrightarrow{n \rightarrow \infty} \alpha$ wegen der asymptotischen Normalverteilung von T .

Die Gütfunktion des Tests ist asymptotisch gleich

$$\lim_{n \rightarrow \infty} G_n(\theta) = 1 - \Phi\left(z_{1-\alpha/2} + \frac{\theta_0 - \theta}{\sigma}\right) + \Phi\left(-z_{1-\alpha/2} + \frac{\theta_0 - \theta}{\sigma}\right),$$

wobei $\hat{\sigma}_n^2 \xrightarrow[n \rightarrow \infty]{P} \sigma^2$.

Spezialfälle des Wald-Tests sind asymptotische Tests der Erwartungswerte bei einer Poisson- oder Bernoulliverteilten Stichprobe.

Beispiel 10.16 (a) Bernoulliverteilung

Es seien $X_i \sim \text{Bernoulli}(p)$, $p \in [0, 1]$ unabhängige, identisch verteilte Zufallsvariablen.

- Hypothesen: $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$.
- Teststatistik:

$$T(X_1, \dots, X_n) = \begin{cases} \sqrt{n} \frac{\bar{X}_n - p_0}{\sqrt{\bar{X}_n(1 - \bar{X}_n)}}, & \text{falls } \bar{X}_n \neq 0, 1, \\ 0, & \text{sonst.} \end{cases}$$

Unter H_0 gilt: $T(X_1, \dots, X_n) \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1)$.

(b) Poissonverteilung

Es seien $X_i \sim \text{Poisson}(\lambda)$, $\lambda > 0$ unabhängige, identisch verteilte Zufallsvariablen.

- Hypothesen: $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$
- Teststatistik:

$$T(X_1, \dots, X_n) = \begin{cases} \sqrt{n} \frac{\bar{X}_n - \lambda_0}{\sqrt{\bar{X}_n}}, & \text{falls } \bar{X}_n > 0, \\ 0, & \text{sonst.} \end{cases}$$

Unter H_0 gilt: $T(X_1, \dots, X_n) \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, 1)$

10.3 Randomisierte Tests

In diesem Abschnitt werden wir klassische Ergebnisse von Neyman-Pearson über die besten Tests präsentieren. Dabei werden randomisierte Tests eine wichtige Rolle spielen.

10.3.1 Grundlagen

Gegeben sei eine Zufallsstichprobe (X_1, \dots, X_n) von unabhängigen und identisch verteilten Zufallsvariablen X_i mit konkreter Ausprägung (x_1, \dots, x_n) . Sei unser Stichprobenraum (B, \mathcal{B}) entweder $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ oder $(\mathbb{N}_0^n, \mathcal{B}_{\mathbb{N}_0^n})$, je nachdem, ob die Stichprobenvariablen $X_i, i = 1, \dots, n$ absolut stetig oder diskret verteilt sind.

Hier wird zur Einfachheit im Falle einer diskret verteilten Zufallsvariable X_i ihr diskreter Wertebereich mit $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ gleichgesetzt. Der Wertebereich sei mit einem Maß μ versehen, wobei

$$\mu = \begin{cases} \text{Lebesgue-Maß auf } \mathbb{R}, & \text{falls } X_i \text{ als stetig verteilt} \\ \text{Zählmaß auf } \mathbb{N}_0, & \text{falls } X_i \text{ diskret verteilt.} \end{cases}$$

Dementsprechend gilt

$$\int g(x) \mu(dx) = \begin{cases} \int_{\mathbb{R}} g(x) dx, & \text{im absolut stetigen Fall,} \\ \sum_{x \in \mathbb{N}_0} g(x), & \text{im diskreten Fall.} \end{cases}$$

Es sei zusätzlich $X_i \sim F_{\theta}, \theta \in \Theta \subseteq \mathbb{R}^m, i = 1, \dots, n$ (parametrisches Modell). Für $\Theta = \Theta_0 \cup \Theta_1, \Theta_0 \cap \Theta_1 = \emptyset$ formulieren wir die Hypothesen $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$, die mit Hilfe eines randomisierten Tests

$$\varphi(x) = \begin{cases} 1, & x \in K_1, \\ \gamma \in (0, 1), & x \in K_{01} \quad x = (x_1, \dots, x_n), \\ 0, & x \in K_0 \end{cases}$$

getestet werden.

Im Falle $x \in K_{01}$ wird mit Hilfe einer Zufallsvariable $Y \sim \text{Bernoulli}(\varphi(x))$ entschieden, ob H_0 verworfen wird ($Y = 1$) oder nicht ($Y = 0$).

Definition 10.17 1. Die *Gütfunktion* eines randomisierten Tests φ sei

$$G_n(\theta) = G_n(\varphi, \theta) = \mathbb{E}_{\theta} \varphi(X_1, \dots, X_n), \theta \in \Theta.$$

2. Der Test φ hat das *Signifikanzniveau* $\alpha \in [0, 1]$, falls $G_n(\varphi, \theta) \leq \alpha, \forall \theta \in \Theta_0$ ist. Die Zahl

$$\sup_{\theta \in \Theta_0} G_n(\varphi, \theta)$$

wird *Umfang* des Tests φ genannt. Offensichtlich ist der Umfang eines Niveau- α -Tests kleiner gleich α .

3. Sei $\Psi(\alpha)$ die Menge aller Tests zum Niveau α . Der Test $\varphi_1 \in \Psi(\alpha)$ ist (*gleichmäßig*) besser als Test $\varphi_2 \in \Psi(\alpha)$, falls $G_n(\varphi_1, \theta) \geq G_n(\varphi_2, \theta), \theta \in \Theta_1$, also falls φ_1 eine größere Macht besitzt.

4. Ein Test $\varphi^* \in \Psi(\alpha)$ ist (*gleichmäßig*) *bester Test* in $\Psi(\alpha)$, falls

$$G_n(\varphi^*, \theta) \geq G_n(\varphi, \theta), \text{ für alle Tests } \varphi \in \Psi(\alpha), \theta \in \Theta_1.$$

Bemerkung 10.18 1. Definition 10.17 1) ist eine offensichtliche Verallgemeinerung der Definition 10.4 der Gütefunktion eines nicht-randomisierten Tests φ . Nämlich, für $\varphi(x) = \mathbb{1}(x \in K_1)$ gilt:

$$\begin{aligned} G_n(\varphi, \theta) &= \mathbb{E}_\theta \varphi(X_1, \dots, X_n) \\ &= P_\theta((X_1, \dots, X_n) \in K_1) \\ &= P_\theta(H_0 \text{ ablehnen}), \theta \in \Theta. \end{aligned}$$

2. Ein bester Test φ^* in $\Psi(\alpha)$ existiert nicht immer, sondern nur unter gewissen Voraussetzungen an $P_\theta, \Theta_0, \Theta_1$ und $\Psi(\alpha)$.

10.3.2 Neyman-Pearson-Tests bei einfachen Hypothesen

In diesem Abschnitt betrachten wir einfache Hypothesen

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1 \tag{10.1}$$

wobei $\theta_0, \theta_1 \in \Theta, \theta_1 \neq \theta_0$.

Dementsprechend sind $\Theta_0 = \{\theta_0\}, \Theta_1 = \{\theta_1\}$. Wir setzen voraus, dass F_{θ_i} eine Dichte $g_i(x)$ bezüglich μ besitzt, $i = 0, 1$. Führen wir einige abkürzende Bezeichnungen $P_0 = P_{\theta_0}, P_1 = P_{\theta_1}, \mathbb{E}_0 = \mathbb{E}_{\theta_0}, \mathbb{E}_1 = \mathbb{E}_{\theta_1}$ ein. Sei $f_i(x) = \prod_{j=1}^n g_i(x_j)$, $x = (x_1, \dots, x_n)$, $i = 0, 1$ die Dichte der Stichprobe unter H_0 bzw. H_1 .

Definition 10.19 Ein *Neyman-Pearson-Test (NP-Test)* der einfachen Hypothesen in (10.1) ist gegeben durch die Regel

$$\varphi(x) = \varphi_K(x) = \begin{cases} 1, & \text{falls } f_1(x) > K f_0(x), \\ \gamma, & \text{falls } f_1(x) = K f_0(x), \\ 0, & \text{falls } f_1(x) < K f_0(x) \end{cases} \tag{10.2}$$

für Konstanten $K > 0$ und $\gamma \in [0, 1]$.

Bemerkung 10.20 1. Manchmal werden $K = K(x)$ und $\gamma = \gamma(x)$ als Funktionen von x und nicht als Konstanten betrachtet.

2. Der *Ablehnungsbereich* des Neyman-Pearson-Tests φ_K ist

$$K_1 = \{x \in B : f_1(x) > K f_0(x)\}.$$

3. Der *Umfang* des Neyman-Pearson-Tests φ_K ist

$$\begin{aligned} \mathbb{E}_0 \varphi_K(X_1, \dots, X_n) &= P_0(f_1(X_1, \dots, X_n) > K f_0(X_1, \dots, X_n)) \\ &\quad + \gamma P_0(f_1(X_1, \dots, X_n) = K f_0(X_1, \dots, X_n)) \end{aligned}$$

4. Die Definition 10.19 kann man äquivalent folgendermaßen geben: Wir definieren eine Teststatistik

$$T(x) = \begin{cases} \frac{f_1(x)}{f_0(x)}, & x \in B : f_0(x) > 0, \\ \infty, & x \in B : f_0(x) = 0. \end{cases}$$

Dann wird der neue Test

$$\tilde{\varphi}_K(x) = \begin{cases} 1, & \text{falls } T(x) > K, \\ \gamma, & \text{falls } T(x) = K, \\ 0, & \text{falls } T(x) < K \end{cases}$$

eingeführt, der für P_0 - und P_1 - fast alle $x \in B$ äquivalent zu φ_k ist. In der Tat gilt $\varphi_K(x) = \tilde{\varphi}_K(x) \forall x \in B \setminus C$, wobei $C = \{x \in B : f_0(x) = f_1(x) = 0\}$ das P_0 - bzw. P_1 -Maß Null besitzt.

In der neuen Formulierung ist der Umfang von φ bzw. $\tilde{\varphi}_K$ gleich

$$\mathbb{E}_0 \tilde{\varphi}_K = P_0(T(X_1, \dots, X_n) > K) + \gamma \cdot P_0(T(X_1, \dots, X_n) = K).$$

Satz 10.21 Optimalitätssatz

Es sei φ_K ein Neyman-Pearson-Test für ein $K > 0$ und $\gamma \in [0, 1]$. Dann ist φ_K der beste Test zum Niveau $\alpha = \mathbb{E}_0 \varphi_K$ seines Umfangs.

Beweis Sei $\varphi \in \Psi(\alpha)$, also $\mathbb{E}_0(\varphi(X_1, \dots, X_n)) \leq \alpha$. Um zu zeigen, dass φ_K besser als φ ist, genügt es bei einfachen Hypothesen H_0 und H_1 zu zeigen, dass $\mathbb{E}_1 \varphi_K(X_1, \dots, X_n) \geq \mathbb{E}_1 \varphi(X_1, \dots, X_n)$. Wir führen dazu die folgenden Mengen ein:

$$\begin{aligned} M^+ &= \{x \in B : \varphi_K(x) > \varphi(x)\} \\ M^- &= \{x \in B : \varphi_K(x) < \varphi(x)\} \\ M^= &= \{x \in B : \varphi_K(x) = \varphi(x)\} \end{aligned}$$

Es gilt offensichtlich $x \in M^+ \Rightarrow \varphi_K(x) > 0 \Rightarrow f_1(x) \geq K f_0(x)$,

$$x \in M^- \Rightarrow \varphi_K(x) < 1 \Rightarrow f_1(x) \leq K f_0(x) \text{ und } B = M^+ \cup M^- \cup M^=.$$

Als Folgerung erhalten wir

$$\begin{aligned}
\mathbb{E}_1(\varphi_K(X_1, \dots, X_n) - \varphi(X_1, \dots, X_n)) &= \int_B (\varphi_K(x) - \varphi(x)) f_1(x) \mu(dx) \\
&= \left(\int_{M^+} + \int_{M^-} + \int_{M^=} \right) (\varphi_K(x) - \varphi(x)) f_1(x) \mu(dx) \\
&\geq \int_{M^+} (\varphi_K(x) - \varphi(x)) K f_0(x) \mu(dx) \\
&\quad + \int_{M^-} (\varphi_K(x) - \varphi(x)) K f_0(x) \mu(dx) \\
&= \int_B (\varphi_K(x) - \varphi(x)) K f_0(x) \mu(dx) \\
&= K [\mathbb{E}_0 \varphi_K(X_1, \dots, X_n) - \mathbb{E}_0 \varphi(X_1, \dots, X_n)] \\
&\geq K(\alpha - \alpha) = 0,
\end{aligned}$$

weil beide Tests das Niveau α haben. Damit ist die Behauptung bewiesen.

□

Bemerkung 10.22 1. Da im Beweis γ nicht vorkommt, wird derselbe Beweis im Falle von $\gamma(x) \neq \text{const}$ gelten.

2. Aus dem Beweis folgt die Gültigkeit der Ungleichung

$$\int_B (\varphi_K(x) - \varphi(x)) (f_1(x) - K f_0(x)) \mu(dx) \geq 0$$

im Falle des konstanten K , bzw.

$$\mathbb{E}_1(\varphi_K(X_1, \dots, X_n) - \varphi(X_1, \dots, X_n)) \geq \int_B (\varphi_K(x) - \varphi(x)) K(x) f_0(x) \mu(dx)$$

im allgemeinen Fall.

Satz 10.23 (Fundamentallemma von Neyman-Pearson)

1. Zu einem beliebigen $\alpha \in (0, 1)$ gibt es einen Neyman-Pearson-Test φ_K mit Umfang α , der dann nach Satz 10.21 der beste Niveau- α -Test ist.
2. Ist φ ebenfalls bester Test zum Niveau α , so gilt $\varphi(x) = \varphi_K(x)$ für μ -fast alle $x \in K_0 \cup K_1 = \{x \in B : f_1(x) \neq K f_0(x)\}$ und φ_K aus Teil 1).

Beweis 1. Für $\varphi_K(x)$ gilt

$$\varphi_K(x) = \begin{cases} 1, & \text{falls } x \in K_1 = \{x : f_1(x) > K \cdot f_0(x)\}, \\ \gamma, & \text{falls } x \in K_{01} = \{x : f_1(x) = K \cdot f_0(x)\}, \\ 0, & \text{falls } x \in K_0 = \{x : f_1(x) < K \cdot f_0(x)\}. \end{cases}$$

Der Umfang von φ_K ist

$$P_0(T(X_1, \dots, X_n) > K) + \gamma P_0(T(X_1, \dots, X_n) = K) = \alpha, \quad (10.3)$$

wobei

$$T(x_1, \dots, x_n) = \begin{cases} \frac{f_1(x_1, \dots, x_n)}{f_0(x_1, \dots, x_n)}, & \text{falls } f_0(x_1, \dots, x_n) > 0, \\ \infty, & \text{sonst.} \end{cases}$$

Nun suchen wir ein $K > 0$ und ein $\gamma \in [0, 1]$, sodass Gleichung (10.3) stimmt. Es sei $\tilde{F}_0(x) = P_0(T(X_1, \dots, X_n) \leq x)$, $x \in \mathbb{R}$ die Verteilungsfunktion von T . Da $T \geq 0$ ist, gilt $\tilde{F}_0(x) = 0$, falls $x < 0$. Außerdem ist $P_0(T(X_1, \dots, X_n) < \infty) = 1$, das heißt $\tilde{F}^{-1}(\alpha) \in [0, \infty)$, $\alpha \in (0, 1)$. Die Gleichung (10.3) kann dann folgendermaßen umgeschrieben werden:

$$1 - \tilde{F}_0(K) + \gamma (\tilde{F}_0(K) - \tilde{F}_0(K-)) = \alpha, \quad (10.4)$$

wobei $\tilde{F}_0(K-) = \lim_{x \rightarrow K-0} \tilde{F}_0(x)$.

Sei $K = \tilde{F}_0^{-1}(1 - \alpha)$, dann gilt:

- (a) Falls K ein Stetigkeitspunkt von \tilde{F}_0 ist, ist Gleichung (10.4) erfüllt für alle $\gamma \in [0, 1]$, zum Beispiel $\gamma = 0$.
- (b) Falls K kein Stetigkeitspunkt von \tilde{F}_0 ist, dann ist $\tilde{F}_0(K) - \tilde{F}_0(K-) > 0$, woraus folgt

$$\gamma = \frac{\alpha - 1 + \tilde{F}_0(K)}{\tilde{F}_0(K) - \tilde{F}_0(K-)}$$

\Rightarrow es gibt einen Neyman-Pearson-Test zum Niveau α .

2. Wir definieren $M^\neq = \{x \in B : \varphi(x) \neq \varphi_K(x)\}$. Es muss gezeigt werden, dass

$$\mu((K_0 \cup K_1) \cap M^\neq) = 0.$$

Dazu betrachten wir

$$\mathbb{E}_1 \varphi(X_1, \dots, X_n) - \mathbb{E}_1 \varphi_K(X_1, \dots, X_n) = 0 \quad (\varphi \text{ und } \varphi_K \text{ sind beste Tests})$$

$$\mathbb{E}_0 \varphi(X_1, \dots, X_n) - \mathbb{E}_0 \varphi_K(X_1, \dots, X_n) \leq 0 \quad (\varphi \text{ und } \varphi_K \text{ sind } \alpha\text{-Tests})$$

mit Umfang von $\varphi_K = \alpha$)

$$\Rightarrow \int_B (\varphi - \varphi_K) \cdot (f_1 - K \cdot f_0) \mu(dx) \geq 0.$$

In Bemerkung 10.22 wurde bewiesen, dass

$$\begin{aligned} \int_B (\varphi - \varphi_K)(f_1 - K \cdot f_0) d\mu &\leq 0 \\ \Rightarrow \int_B (\varphi - \varphi_K)(f_1 - K \cdot f_0) d\mu &= 0 = \int_{M^\neq \cap (K_0 \cup K_1)} (\varphi - \varphi_K)(f_1 - K f_0) d\mu. \end{aligned}$$

Es gilt $\mu(M^\neq \cap (K_0 \cup K_1)) = 0$, falls der Integrand $(\varphi_K - \varphi)(f_1 - K f_0) > 0$ auf M^\neq ist. Wir zeigen, dass

$$(\varphi_K - \varphi)(f_1 - K f_0) > 0 \text{ für } x \in M^\neq \quad (10.5)$$

ist. Es gilt

$$\begin{aligned} f_1 - K f_0 &> 0 \Rightarrow \varphi_K - \varphi > 0, \\ f_1 - K f_0 &< 0 \Rightarrow \varphi_K - \varphi < 0, \end{aligned}$$

weil

$$\begin{aligned} f_1(x) > K f_0(x) &\Rightarrow \varphi_K(x) = 1 \\ &\quad \text{und mit } \varphi(x) < 1 \Rightarrow \varphi_K(x) - \varphi(x) > 0 \text{ auf } M^\neq. \\ f_1(x) < K f_0(x) &\Rightarrow \varphi_K(x) = 0 \\ &\quad \text{und mit } \varphi(x) > 0 \Rightarrow \varphi_K(x) - \varphi(x) < 0 \text{ auf } M^\neq. \end{aligned}$$

Daraus folgt die Gültigkeit der Ungleichung (10.5) und somit

$$\mu((K_0 \cup K_1) \cap M^\neq) = 0.$$

□

Bemerkung 10.24 Falls φ und φ_K beste α -Tests sind, dann sind sie P_0 - bzw. P_1 - fast sicher gleich.

Beispiel 10.25 (Neyman-Pearson-Test für den Parameter der Poissonverteilung) Es sei (X_1, \dots, X_n) eine Zufallsstichprobe mit $X_i \sim \text{Poisson}(\lambda)$, $\lambda > 0$, wobei X_i unabhängig und identisch verteilt sind für $i = 1, \dots, n$. Wir testen die Hypothesen $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda = \lambda_1$. Dabei ist

$$\begin{aligned} g_i(x) &= e^{-\lambda_i} \frac{\lambda_i^x}{x!}, \quad x \in \mathbb{N}_0, \quad i = 0, 1, \\ f_i(x) &= f_i(x_1, \dots, x_n) = \prod_{j=1}^n g_i(x_j) = \prod_{j=1}^n e^{-\lambda_i} \frac{\lambda_i^{x_j}}{x_j!} = e^{-n\lambda_i} \cdot \frac{\lambda_i^{\sum_{j=1}^n x_j}}{(x_1! \cdot \dots \cdot x_n!)} \end{aligned}$$

für $i = 0, 1$. Die Neyman-Pearson-Teststatistik ist

$$T(x_1, \dots, x_n) = \begin{cases} \frac{f_1(x)}{f_0(x)} = e^{-n(\lambda_1 - \lambda_0)} \cdot (\lambda_1/\lambda_0)^{\sum_{j=1}^n x_j}, & \text{falls } x_1, \dots, x_n \in \mathbb{N}_0, \\ \infty, & \text{sonst.} \end{cases} .$$

Die Neyman-Pearson-Entscheidungsregel lautet

$$\varphi_K(x_1, \dots, x_n) = \begin{cases} 1, & \text{falls } T(x_1, \dots, x_n) > K, \\ \gamma, & \text{falls } T(x_1, \dots, x_n) = K, \\ 0, & \text{falls } T(x_1, \dots, x_n) < K. \end{cases}$$

Wir wählen $K > 0$, $\gamma \in [0, 1]$, sodass φ_K den Umfang α hat. Dazu lösen wir

$$\alpha = P_0(T(X_1, \dots, X_n) > K) + \gamma P_0(T(X_1, \dots, X_n) = K)$$

bezüglich γ und K auf.

$$\begin{aligned} P_0(T(X_1, \dots, X_n) > K) &= P_0(\log T(X_1, \dots, X_n) > \log K) \\ &= P_0\left(-n(\lambda_1 - \lambda_0) + \sum_{j=1}^n X_j \cdot \log\left(\frac{\lambda_1}{\lambda_0}\right) > \log K\right) = P_0\left(\sum_{j=1}^n X_j > A\right) \\ \text{wobei } A &:= \left\lfloor \frac{\log K + n \cdot (\lambda_1 - \lambda_0)}{\log \frac{\lambda_1}{\lambda_0}} \right\rfloor, \end{aligned}$$

falls zum Beispiel $\lambda_1 > \lambda_0$. Im Falle $\lambda_1 < \lambda_0$ ändert sich das $>$ auf $<$ in der Wahrscheinlichkeit.

Wegen der Faltungsstabilität der Poissonverteilung ist unter H_0

$$\sum_{j=1}^n X_i \sim \text{Poisson}(n\lambda_0),$$

also wählen wir K als minimale, nichtnegative Zahl, für die gilt: $P_0\left(\sum_{j=1}^n X_j > A\right) \leq \alpha$, und setzen

$$\gamma = \frac{\alpha - P_0(\sum_{j=1}^n X_j > A)}{P_0(\sum_{j=1}^n X_j = A)},$$

wobei

$$\begin{aligned} P_0\left(\sum_{j=1}^n X_j > A\right) &= 1 - \sum_{j=0}^A e^{-\lambda_0 n} \frac{(\lambda_0 n)^j}{j!}, \\ P_0\left(\sum_{j=1}^n X_j = A\right) &= e^{-\lambda_0 n} \frac{(\lambda_0 n)^A}{A!}. \end{aligned}$$

Somit haben wir die Parameter K und γ gefunden und damit einen Neyman-Pearson-Test φ_K konstruiert.

10.3.3 Einseitige Neyman-Pearson-Tests

Bisher betrachteten wir Neyman-Pearson-Tests für einfache Hypothesen der Form $H_i : \theta = \theta_i, i = 0, 1$. In diesem Abschnitt wollen wir einseitige Neyman-Pearson-Tests einführen, für Hypothesen der Form $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$.

Zunächst konstruieren wir einen Test für diese Hypothesen: Sei (X_1, \dots, X_n) eine Zufallsstichprobe, X_i seien unabhängig und identisch verteilt mit

$$X_i \sim F_\theta \in \Lambda = \{F_\theta : \theta \in \Theta\},$$

wobei $\Theta \subset \mathbb{R}$ offen ist und Λ eindeutig parametrisiert, das heißt

$$\theta \neq \theta' \Rightarrow F_\theta \neq F_{\theta'}.$$

Ferner besitze F_θ eine Dichte g_θ bezüglich des Lebesgue-Maßes (bzw. Zählmäßig) auf \mathbb{R} (bzw. \mathbb{N}_0). Dann ist

$$f_\theta(x) = \prod_{j=1}^n g_\theta(x_j), \quad x = (x_1, \dots, x_n)$$

eine Dichte von (X_1, \dots, X_n) bezüglich μ auf B .

Definition 10.26 Eine Verteilung auf B mit Dichte f_θ gehört zur Klasse von *Verteilungen mit monotonen Dichtekoeffizienten* in T , falls es für alle $\theta < \theta'$ eine Funktion $h : \mathbb{R} \times \Theta^2 \rightarrow \mathbb{R} \cup \infty$, die monoton wachsend in $t \in \mathbb{R}$ ist und eine Statistik $T : B \rightarrow \mathbb{R}$ gibt, mit der Eigenschaft

$$\frac{f_{\theta'}(x)}{f_\theta(x)} = h(T(x), \theta, \theta'),$$

wobei

$$h(T(x), \theta, \theta') = \infty \quad \text{für alle } x \in B : f_\theta(x) = 0, f_{\theta'}(x) > 0.$$

Der Fall $f_\theta(x) = f_{\theta'}(x) = 0$ tritt mit P_Θ - bzw. $P_{\Theta'}$ -Wahrscheinlichkeit 0 auf.

Definition 10.27 Es sei Q_θ eine Verteilung auf (B, \mathcal{B}) mit der Dichte f_θ bzgl. μ . Q_θ gehört zur *einparametrischen Exponentialklasse* ($\theta \in \Theta \subset \mathbb{R}$ offen), falls die Dichte folgende Form hat:

$$f_\theta(x) = \exp \{c(\theta) \cdot T(x) + a(\theta)\} \cdot l(x), \quad x = (x_1, \dots, x_n) \in B,$$

wobei $c(\theta)$ eine monoton steigende Funktion ist, und $\operatorname{Var}_\theta T(X_1, \dots, X_n) > 0, \theta \in \Theta$.

Lemma 10.28 Verteilungen aus der einparametrischen Exponentialfamilie besitzen einen monotonen Dichtekoeffizienten.

Beweis Es sei Q_θ aus der einparametrischen Exponentialfamilie mit der Dichte

$$f_\theta(x) = \exp \{c(\theta) \cdot T(x) + a(\theta)\} \cdot l(x).$$

Für $\theta < \theta'$ ist dann

$$\frac{f_{\theta'}(x)}{f_\theta(x)} = \exp \{(c(\theta') - c(\theta)) \cdot T(x) + a(\theta') - a(\theta)\}$$

monoton bezüglich T , weil $c(\theta') - c(\theta) > 0$ wegen der Monotonie von $c(\theta)$. Also besitzt f_θ einen monotonen Dichtekoeffizienten. \square

Beispiel 10.29 1. Normalverteilte Stichprobenvariablen

Es seien $X_i \sim N(\mu, \sigma_0^2)$, $i = 1, \dots, n$, unabhängige, identisch verteilte Zufallsvariablen, mit unbekanntem Parameter μ und bekannter Varianz σ_0^2 (Hier wird μ für die Bezeichnung des Erwartungswertes von X_i und nicht des Maßes auf \mathbb{R}^n verwendet. (wie früher)). Die Dichte des Zufallsvektors $X = (X_1, \dots, X_n)^\top$ ist gleich

$$\begin{aligned} f_\mu(x) &= \prod_{i=1}^n g_\mu(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma_0^2}} \\ &= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \\ &= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma_0^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + \mu^2 n \right) \right\} \\ &= \underbrace{\exp \left(\underbrace{\frac{\mu}{\sigma_0^2} \cdot \sum_{i=1}^n x_i}_{c(\mu)} - \underbrace{\frac{\mu^2 n}{2\sigma_0^2}}_{a(\mu)} \right)}_{l(x)} \cdot \underbrace{\frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp \left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma_0^2} \right)}_{l(x)}. \end{aligned}$$

Also gehört $N(\mu, \sigma_0^2)$ zur einparametrischen Exponentialklasse mit $c(\mu) = \frac{\mu}{\sigma_0^2}$ und $T(x) = \sum_{i=1}^n x_i$.

2. Binomialverteilte Stichprobenvariablen

Es seien $X_i \sim \text{Bin}(k, p)$ unabhängig und identisch verteilt, $i = 1, \dots, n$. Der Parameter p sei unbekannt. Die Zähldichte des Zufallsvektors $X =$

$(X_1, \dots, X_n)^\top$ ist

$$\begin{aligned} f_p(x) &= P_p(X_i = x_i, i = 1, \dots, n) \\ &= \prod_{i=1}^n \binom{k}{x_i} p^{x_i} (1-p)^{k-x_i} = p^{\sum_{i=1}^n x_i} \cdot \frac{(1-p)^{nk}}{(1-p)^{\sum_{i=1}^n x_i}} \cdot \prod_{i=1}^n \binom{k}{x_i} \\ &= \exp \left\{ \underbrace{\left(\sum_{i=1}^n x_i \right)}_{T(x)} \cdot \underbrace{\log \left(\frac{p}{1-p} \right)}_{c(p)} + \underbrace{nk \cdot \log(1-p)}_{a(p)} \right\} \cdot \underbrace{\prod_{i=1}^n \binom{k}{x_i}}_{l(x)}, \end{aligned}$$

also gehört $\text{Bin}(n, p)$ zur einparametrischen Exponentialklasse mit

$$c(p) = \log \left(\frac{p}{1-p} \right)$$

und

$$T(x) = \sum_{i=1}^n x_i.$$

Lemma 10.30 Falls φ_K der Neyman-Pearson-Test der Hypothesen $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ ist, dann gilt:

$$\mu \underbrace{\left(\{x \in B : f_1(x) \neq K f_0(x)\} \right)}_{K_0 \cup K_1} > 0.$$

Beweis Wegen $\theta_0 \neq \theta_1$ und der eindeutigen Parametrisierung gilt $f_0 \neq f_1$ auf einer Menge mit μ -Maß > 0 .

Nun sei $\mu(K_0 \cup K_1) = 0$. Daraus folgt, dass $f_1(x) = K \cdot f_0(x)$ μ -fast sicher. Das heißt

$$1 = \int_B f_1(x) dx = K \cdot \int_B f_0(x) dx,$$

woraus folgt, dass $K = 1$ und $f_1(x) = f_0(x)$ μ -fast sicher, was aber ein Widerspruch zur eindeutigen Parametrisierung ist. \square

Im Folgenden sei (X_1, \dots, X_n) eine Stichprobe von unabhängigen, identisch verteilten Zufallsvariablen mit $X_i \sim \text{Dichte } g_\theta, i = 1, \dots, n$ und

$$(X_1, \dots, X_n) \sim \text{Dichte } f_\theta(x) = \prod_{i=1}^n g_\theta(x_i)$$

aus der Klasse der Verteilungen mit monotonen Dichtekoeffizienten und einer Statistik $T(X_1, \dots, X_n)$.

Wir betrachten die Hypothesen $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$ und den Neyman-Pearson-Test:

$$\varphi_{K^*}^*(x) = \begin{cases} 1, & \text{falls } T(x) > K^*, \\ \gamma^*, & \text{falls } T(x) = K^*, \\ 0, & \text{falls } T(x) < K^* \end{cases} \quad (10.6)$$

für $K^* \in \mathbb{R}$ und $\gamma^* \in [0, 1]$. Die Gütfunktion von $\varphi_{K^*}^*$ bei θ_0 ist

$$G_n(\theta_0) = \mathbb{E}_0 \varphi_{K^*}^* = P_0(T(X_1, \dots, X_n) > K^*) + \gamma^* \cdot P_0(T(X_1, \dots, X_n) = K^*)$$

Satz 10.31 1. Falls $\alpha = \mathbb{E}_0 \varphi_{K^*}^* > 0$, dann ist der soeben definierte Neyman-Pearson-Test ein bester Test der einseitigen Hypothesen H_0 vs. H_1 zum Niveau α .

2. Zu jedem Konfidenzniveau $\alpha \in (0, 1)$ gibt es ein $K^* \in \mathbb{R}$ und $\gamma^* \in [0, 1]$, sodass $\varphi_{K^*}^*$ ein bester Test zum Umfang α ist.
3. Die Gütfunktion $G_n(\theta)$ von $\varphi_{K^*}^*(\theta)$ ist monoton wachsend in θ . Falls $0 < G_n(\theta) < 1$, dann ist sie sogar streng monoton wachsend.

Beweis 1. Wähle $\theta_1 > \theta_0$ und betrachte die einfachen Hypothesen $H'_0 : \theta = \theta_0$ und $H'_1 : \theta = \theta_1$. Sei

$$\varphi_K(x) = \begin{cases} 1, & f_1(x) > Kf_0(x), \\ \gamma, & f_1(x) = Kf_0(x), \\ 0, & f_1(x) < Kf_0(x) \end{cases}$$

der Neyman-Pearson-Test für H'_0, H'_1 mit $K > 0$. Da f_θ den monotonen Dichtekoeffizienten mit Statistik T besitzt,

$$\frac{f_1(x)}{f_0(x)} = h(T(x), \theta_0, \theta_1),$$

existiert ein $K > 0$, so dass

$$\left\{ x : \frac{f_1(x)}{f_0(x)} > K \right\} \subset \left\{ T(x) > K^* \right\} \quad \text{mit } K = h(K^*, \theta_0, \theta_1).$$

φ_K ist ein bester Neyman-Pearson-Test zum Niveau $\alpha = \mathbb{E}_0 \varphi_K = \mathbb{E}_0 \varphi_{K^*}^*$. Aus $\alpha > 0$ folgt $K < \infty$, denn aus $K = \infty$ würde folgen

$$\begin{aligned} 0 < \alpha = \mathbb{E}_0 \varphi_K &\leq P_0 \left(\frac{f_1(X_1, \dots, X_n)}{f_0(X_1, \dots, X_n)} \geq K^* \right) \leq P_0 \left(\frac{f_1(X_1, \dots, X_n)}{f_0(X_1, \dots, X_n)} = \infty \right) \\ &= P_0(f_1(X_1, \dots, X_n) > 0, f_0(X_1, \dots, X_n) = 0) \\ &= \int_B \mathbb{1}(f_1(x) > 0, f_0(x) = 0) \cdot f_0(x) \mu(dx) = 0. \end{aligned}$$

Für den Test $\varphi_{K^*}^*$ aus (10.6) gilt dann

$$\varphi_{K^*}^*(x) = \begin{cases} 1, & \text{falls } f_1(x)/f_0(x) > K, \\ \gamma^*(x), & \text{falls } f_1(x)/f_0(x) = K, \\ 0, & \text{falls } f_1(x)/f_0(x) < K, \end{cases}$$

wobei $\gamma^*(x) \in \{\gamma^*, 0, 1\}$. Daraus folgt, dass $\varphi_{K^*}^*$ ein bester Neyman-Pearson-Test ist für H'_0 vs. H'_1 (vergleiche Bemerkung 10.20, 1.) und Bemerkung 10.22 für beliebige $\theta_1 > \theta_0$. Deshalb ist $\varphi_{K^*}^*$ ein bester Neyman-Pearson-Test für $H''_0 : \theta = \theta_0$ vs. $H''_1 : \theta > \theta_0$ ist.

Dieselbe Behauptung erhalten wir aus dem Teil 3. des Satzes für $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$, weil dann $G_n(\theta) \leq G_n(\theta_0) = \alpha$ für alle $\theta < \theta_0$.

2. Siehe Beweis zu Satz 10.23, 1.).
3. Wir müssen zeigen, dass $G_n(\theta)$ monoton ist. Dazu wählen wir $\theta_1 < \theta_2$ und zeigen, dass $\alpha_1 = G_n(\theta_1) \leq G_n(\theta_2)$. Wir betrachten die neuen, einfachen Hypothesen $H''_0 : \theta = \theta_1$ vs. $H''_1 : \theta = \theta_2$. Der Test $\varphi_{K^*}^*$ kann genauso wie in 1. als Neyman-Pearson-Test dargestellt werden (für die Hypothesen H''_0 und H''_1), der ein bester Test zum Niveau α_1 ist. Betrachten wir einen weiteren konstanten Test $\varphi(x) = \alpha_1$. Dann ist $\alpha_1 = \mathbb{E}_{\theta_2} \varphi \leq \mathbb{E}_{\theta_2} \varphi_{K^*}^* = G_n(\theta_2)$. Daraus folgt, dass $G_n(\theta_1) \leq G_n(\theta_2)$. Nun zeigen wir, dass für $G_n(\theta) \in (0, 1)$ gilt: $G_n(\theta_1) < G_n(\theta_2)$. Wir nehmen an, dass $\alpha_1 = G_n(\theta_1) = G_n(\theta_2)$ und $\theta_1 < \theta_2$ für $\alpha \in (0, 1)$. Es folgt, dass $\varphi(x) = \alpha_1$ auch ein bester Test für H''_0 und H''_1 ist. Aus Satz 10.23, 2.) folgt

$$\mu(\{x \in B : \underbrace{\varphi(x)}_{=\alpha_1} \neq \varphi_{K^*}^*(x)\}) = 0 \text{ auf } K_0 \cup K_1 = \{f_1(x) \neq K f_0(x)\},$$

was ein Widerspruch zur Bauart des Tests φ_{K^*} ist, der auf $K_0 \cup K_1$ nicht gleich $\alpha_1 \in (0, 1)$ sein kann. \square

Bemerkung 10.32 1. Der Satz 10.31 ist genauso auf Neyman-Pearson-Tests der einseitigen Hypothesen

$$H_0 : \theta \geq \theta_0 \text{ vs. } H_1 : \theta < \theta_0$$

anwendbar, mit dem entsprechenden Unterschied

$$\begin{aligned} \theta &\mapsto -\theta \\ T &\mapsto -T \end{aligned}$$

Somit existiert der beste α -Test auch in diesem Fall.

2. Man kann zeigen, dass die Gütefunktion $G_n(\varphi_{K^*}^*, \theta)$ des besten Neyman-Pearson-Tests auf $\Theta_0 = (-\infty, \theta_0)$ folgende Minimalitätseigenschaft besitzt:

$$G_n(\varphi_{K^*}^*, \theta) \leq G_n(\varphi, \theta) \quad \forall \varphi \in \Psi(\alpha), \theta \leq \theta_0$$

Beispiel 10.33 Wir betrachten eine normalverteilte Stichprobe (X_1, \dots, X_n) von unabhängigen und identisch verteilten Zufallsvariablen X_i , wobei $X_i \sim N(\mu, \sigma_0^2)$ und σ_0^2 sei bekannt. Es werden die Hypothesen

$$H_0 : \mu \leq \mu_0 \text{ vs. } H_1 : \mu > \mu_0,$$

getestet. Aus Beispiel 10.7 kennen wir die Testgröße

$$T(X_1, \dots, X_n) = \sqrt{n} \frac{\bar{X}_n - \mu_0}{\sigma_0},$$

wobei unter H_0 gilt: $T(X_1, \dots, X_n) \sim N(0, 1)$. H_0 wird verworfen, falls

$$T(X_1, \dots, X_n) > z_{1-\alpha}, \quad \text{wobei } \alpha \in (0, 1).$$

Wir zeigen jetzt, dass dieser Test der beste Neyman-Pearson-Test zum Niveau α ist. Aus Beispiel 10.29 ist bekannt, dass die Dichte f_n von (X_1, \dots, X_n) aus der einparametrischen Exponentialklasse ist, mit

$$\tilde{T}(X_1, \dots, X_n) = \sum_{i=1}^n X_i.$$

Dann gehört f_μ von (x_1, \dots, x_n) zur einparametrischen Exponentialklasse auch bezüglich der Statistik

$$T(X_1, \dots, X_n) = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma_0}$$

Es gilt nämlich

$$\begin{aligned} f_\mu(x) &= \exp \left(\underbrace{\frac{\mu}{\sigma_0^2}}_{c(\mu)} \cdot \underbrace{\sum_{i=1}^n x_i}_{\tilde{T}} - \underbrace{\frac{\mu^2 n}{2\sigma_0^2}}_{\tilde{a}(\mu)} \right) \cdot l(x) \\ &= \exp \left(\underbrace{\frac{\mu \sqrt{n}}{\sigma_0}}_{c(\mu)} \cdot \underbrace{\sqrt{n} \frac{\bar{x}_n - \mu}{\sigma_0}}_T + \underbrace{\frac{\mu^2 n}{2\sigma_0^2}}_{a(\mu)} \right) \cdot l(x). \end{aligned}$$

Die Statistik T kann also in der Konstruktion des Neyman-Pearson-Tests (Gleichung (10.6)) verwendet werden:

$$\varphi_{K^*}(x) = \begin{cases} 1, & \text{falls } T(x) > z_{1-\alpha}, \\ 0, & \text{falls } T(x) = z_{1-\alpha}, \\ 0, & \text{falls } T(x) < z_{1-\alpha} \end{cases}$$

(mit $K^* = z_{1-\alpha}$ und $\gamma^* = 0$). Nach Satz 10.31 ist dieser Test der beste Neyman-Pearson-Test zum Niveau α für unsere Hypothesen:

$$\begin{aligned} G_n(\varphi_{K^*}, \mu_0) &= P_0(T(X_1, \dots, X_n) > z_{1-\alpha}) + 0 \cdot P_0(T(X_1, \dots, X_n) \leq z_{1-\alpha}) \\ &= 1 - \Phi(z_{1-\alpha}) = 1 - (1 - \alpha) = \alpha. \end{aligned}$$

10.3.4 Unverfälschte zweiseitige Tests

Es sei (X_1, \dots, X_n) eine Stichprobe von unabhängigen und identisch verteilten Zufallsvariablen mit der Dichte

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

Es wird ein zweiseitiger Test der Hypothesen

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

betrachtet. Für alle $\alpha \in (0, 1)$ kann es jedoch keinen besten Neyman-Pearson-Test φ zum Niveau α für H_0 vs. H_1 geben. Denn, nehmen wir an, φ wäre der beste Test zum Niveau α für H_0 vs. H_1 , dann wäre φ der beste Test für die Hypothesen

1. $H'_0 : \theta = \theta_0$ vs. $H'_1 : \theta > \theta_0$
2. $H''_0 : \theta = \theta_0$ vs. $H''_1 : \theta < \theta_0$.

Dann ist nach Satz 10.31, 3. die Gütfunktion

1. $G_n(\varphi, \theta) < \alpha$ auf $\theta < \theta_0$, bzw.
2. $G_n(\varphi, \theta) > \alpha$ auf $\theta < \theta_0$,

was ein Widerspruch ist!

Darum werden wir die Klasse aller möglichen Tests auf unverfälschte Niveau- α -Tests (Definition 10.12) eingrenzen. Der Niveau- α -Test φ ist unverfälscht genau dann, wenn

$$\begin{aligned} G_n(\varphi, \theta) &\leq \alpha \text{ für } \theta \in \Theta_0 \\ G_n(\varphi, \theta) &\geq \alpha \text{ für } \theta \in \Theta_1 \end{aligned}$$

Beispiel 10.34 1. $\varphi(x) \equiv \alpha$ ist unverfälscht.

2. Der zweiseitige Gauß-Test ist unverfälscht, vergleiche Beispiel 10.7:
 $G_n(\varphi, \mu) \geq \alpha$ für alle $\mu \in \mathbb{R}$.

Im Folgenden seien X_i unabhängig und identisch verteilt. Die Dichte f_θ von (X_1, \dots, X_n) gehöre zur einparametrischen Exponentialklasse:

$$f_\theta(x) = \exp \{c(\theta) \cdot T(x) + a(\theta)\} \cdot l(x), \quad (10.7)$$

wobei $c(\theta)$ und $a(\theta)$ stetig differenzierbar auf Θ sein sollen, mit

$$c'(\theta) > 0 \quad \text{und} \quad \operatorname{Var}_\theta T(X_1, \dots, X_n) > 0$$

für alle $\theta \in \Theta$. Sei $f_\Phi(x)$ stetig in (x, Θ) auf $B \times \Theta$.

Übungsaufgabe 10.35 Zeigen Sie, dass folgende Relation gilt:

$$a'(\theta) = -c'(\theta) \mathbb{E}_\theta T(X_1, \dots, X_n).$$

Lemma 10.36 Es sei φ ein unverfälschter Test zum Niveau α für

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0.$$

Dann gilt:

1. $\alpha = \mathbb{E}_0 \varphi(X_1, \dots, X_n) = G_n(\varphi, \theta_0)$
2. $\mathbb{E}_0 [T(X_1, \dots, X_n) \varphi(X_1, \dots, X_n)] = \alpha \cdot \mathbb{E}_0 T(X_1, \dots, X_n)$

Beweis 1. Die Gütfunktion von φ ist

$$G_n(\varphi, \theta) = \int_B \varphi(x) f_\theta(x) \mu(dx)$$

Da f_θ aus der einparametrischen Exponentialklasse ist, ist $G_n(\varphi, \theta)$ differenzierbar (unter dem Integral) bezüglich θ . Wegen der Unverfälschtheit von φ gilt

$$G_n(\varphi, \theta_0) \leq \alpha, \quad G_n(\varphi, \theta) \geq \alpha, \quad \theta \neq \theta_0$$

und daraus folgt $G_n(\varphi, \theta_0) = \alpha$ und θ_0 ist ein Minimumpunkt von G_n . Somit ist 1) bewiesen.

2. Da θ_0 der Minimumpunkt von G_n ist, gilt

$$\begin{aligned} 0 &= G'_n(\varphi, \theta_0) = \int_B \varphi(x) (c'(\theta_0) T(x) + a'(\theta_0)) f_\theta(x) \mu(dx) \\ &= c'(\theta_0) \cdot \mathbb{E}_0 [\varphi(X_1, \dots, X_n) T(X_1, \dots, X_n)] + a'(\theta_0) \cdot G_n(\varphi, \theta_0) \\ &= c'(\theta_0) \cdot \mathbb{E}_0 [\varphi(X_1, \dots, X_n) T(X_1, \dots, X_n)] + \alpha a'(\theta_0) \\ &\stackrel{\text{(Übung 10.35)}}{=} c'(\theta_0) (\mathbb{E}_0 (\varphi \cdot T) - \alpha \mathbb{E}_0 T) \end{aligned}$$

Daraus folgt $\mathbb{E}_0 (\varphi T) = \alpha \mathbb{E}_0 T$ und damit ist das Lemma bewiesen. \square

Wir definieren jetzt die modifizierten Neyman-Pearson-Tests für einfache Hypothesen

$$H_0 : \theta = \theta_0 \text{ vs. } H'_1 : \theta = \theta_1, \quad \theta_1 \neq \theta_0.$$

Für $\lambda, K \in \mathbb{R}$, $\gamma : B \rightarrow [0, 1]$ definieren wir

$$\varphi_{K,\lambda}(x) = \begin{cases} 1, & \text{falls } f_1(x) > (K + \lambda T(x))f_0(x), \\ \gamma(x), & \text{falls } f_1(x) = (K + \lambda T(x))f_0(x), \\ 0, & \text{falls } f_1(x) < (K + \lambda T(x))f_0(x), \end{cases} \quad (10.8)$$

wobei $T(x)$ die Statistik aus der Darstellung (10.7) ist.

Es sei $\tilde{\Psi}(\alpha)$ die Klasse aller Tests, die Aussagen 1) und 2) des Lemmas 10.36 erfüllen. Aus Lemma 10.36 folgt dann, dass die Menge der unverfälschten Tests zum Niveau α eine Teilmenge von $\tilde{\Psi}(\alpha)$ ist.

Satz 10.37 Der modifizierte Neyman-Pearson-Test $\varphi_{K,\lambda}$ ist der beste α -Test in $\tilde{\Psi}(\alpha)$ für Hypothesen H_0 vs. H'_1 zum Niveau $\alpha = \mathbb{E}_0 \varphi_{K,\lambda}$, falls $\varphi_{K,\lambda} \in \tilde{\Psi}(\alpha)$.

Beweis Es ist zu zeigen, dass $\mathbb{E}_1 \varphi_{K,\lambda} \geq \mathbb{E}_1 \varphi$ für alle $\varphi \in \tilde{\Psi}(\alpha)$, bzw. $\mathbb{E}_1 (\varphi_{K,\lambda} - \varphi) \geq 0$. Es gilt

$$\begin{aligned} \mathbb{E}_1 (\varphi_{K,\lambda} - \varphi) &= \int_B (\varphi_{K,\lambda}(x) - \varphi(x)) f_1(x) \mu(dx) \\ &\stackrel{(\text{Bem. 10.22, 2.)})}{\geq} \int_B (\varphi_{K,\lambda}(x) - \varphi(x)) (K + \lambda T(x)) f_0(x) \mu(dx) \\ &= K \left(\underbrace{\mathbb{E}_0 \varphi_{K,\lambda}}_{=\alpha} - \underbrace{\mathbb{E}_0 \varphi}_{=\alpha} \right) + \lambda \left(\underbrace{\mathbb{E}_0 (\varphi_{K,\lambda} \cdot T)}_{\alpha \mathbb{E}_0 T} - \underbrace{\mathbb{E}_0 (\varphi \cdot T)}_{=\alpha \cdot \mathbb{E}_0 T} \right) \\ &= 0, \end{aligned}$$

weil $\varphi, \varphi_{K,\lambda} \in \tilde{\Psi}(\alpha)$. \square

Wir definieren folgende Entscheidungsregel, die später zum Testen der zweiseitigen Hypothesen

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

verwendet wird:

$$\varphi_c(x) = \begin{cases} 1, & \text{falls } T(x) \notin [c_1, c_2], \\ \gamma_1, & \text{falls } T(x) = c_1, \\ \gamma_2, & \text{falls } T(x) = c_2, \\ 0, & \text{falls } T(x) \in (c_1, c_2), \end{cases} \quad (10.9)$$

für $c_1 \leq c_2 \in \mathbb{R}$, $\gamma_1, \gamma_2 \in [0, 1]$ und die Statistik $T(x)$, $x = (x_1, \dots, x_n) \in B$, die in der Dichte (10.7) vorkommt. Zeigen wir, dass φ_c sich als modifizierter Neyman-Pearson-Test schreiben lässt.

Für die Dichte

$$f_\theta(x) = \exp\{c(\theta)T(x) + a(\theta)\} \cdot l(x)$$

wird (wie immer) vorausgesetzt, dass $l(x) > 0$, $c'(x) > 0$ und $a'(\theta)$ existiert für $\theta \in \Theta$.

Lemma 10.38 Es sei (X_1, \dots, X_n) eine Stichprobe von unabhängigen, identisch verteilten Zufallsvariablen mit gemeinsamer Dichte $f_\theta(x), x \in B$, die zur einparametrischen Exponentialfamilie gehört. Sei $T(x)$ die dazugehörige Statistik, die im Exponenten der Dichte f_θ vorkommt. Für beliebige reelle Zahlen $c_1 \leq c_2$, $\gamma_1, \gamma_2 \in [0, 1]$ und Parameterwerte $\theta_0, \theta_1 \in \Theta : \theta_0 \neq \theta_1$ lässt sich der Test φ_c aus (10.9) als modifizierter Neyman-Pearson-Test $\varphi_{K,\lambda}$ aus (10.8) mit gegebenen $K, \lambda \in \mathbb{R}$, $\gamma(x) \in [0, 1]$ schreiben.

Beweis Falls wir die Bezeichnung

$$f_{\theta_i}(x) = f_i(x), \quad i = 0, 1$$

verwenden, dann gilt

$$\frac{f_1(x)}{f_0(x)} = \exp \left\{ \underbrace{(c(\theta_1) - c(\theta_0))}_{c} T(x) + \underbrace{a(\theta_1) - a(\theta_0)}_{a} \right\},$$

und somit

$$\{x \in B : f_1(x) > (K + \lambda T(x)) f_0(x)\} = \{x \in B : \exp(cT(x) + a) > K + \lambda T(x)\}.$$

Finden wir solche K und λ aus \mathbb{R} , für die die Gerade $K + \lambda t$, $t \in \mathbb{R}$ die konvexe Kurve $\exp\{ct + a\}$ genau an den Stellen c_1 und c_2 schneidet (falls $c_1 \neq c_2$) bzw. an der Stelle $t = c_1$ berührt (falls $c_1 = c_2$). Dies ist immer möglich, siehe Abbildung 10.6.

Ferner setzen wir $\gamma(x) = \gamma_i$ für $\{x \in B : T(x) = c_i\}$. Insgesamt gilt dann

$$\{x : \exp(cT(x) + a) > K + \lambda T(x)\} = \{x : T(x) \notin [c_1, c_2]\}$$

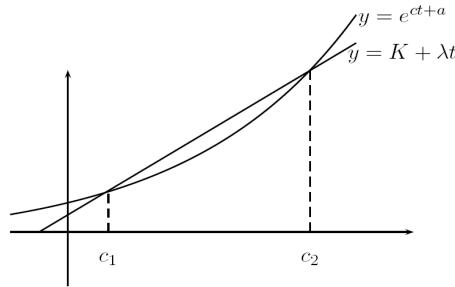
und

$$\{x : \exp(cT(x) + a) < K + \lambda T(x)\} = \{x : T(x) \in (c_1, c_2)\}.$$

Damit ist das Lemma bewiesen. □

Bemerkung 10.39 1. Die Umkehrung des Lemmas stimmt nicht, denn bei vorgegebenen Kurven $y = K + \lambda t$ und $y = \exp\{ct + a\}$ muss es die Schnittpunkte c_1 und c_2 nicht unbedingt geben. So kann die Gerade vollständig unter der Kurve $y = \exp\{ct + a\}$ liegen.

Abbildung 10.6:



2. Der Test φ_c macht von den Werten θ_0 und θ_1 nicht explizit Gebrauch. Dies unterscheidet ihn vom Test $\varphi_{K,\lambda}$, für den die Dichten f_0 und f_1 gebraucht werden.

Jetzt sind wir bereit, den Hauptsatz über zweiseitige Tests zum Prüfen der Hypothesen

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

zu formulieren und zu beweisen.

Satz 10.40 (Hauptsatz über zweiseitige Tests)

Unter den Voraussetzungen des Lemmas 10.38 sei φ_c ein Test aus (10.9), für den $\varphi_c \in \tilde{\Psi}(\alpha)$ gilt. Dann ist φ_c bester unverfälschter Test zum Niveau α (und dadurch bester Test in $\tilde{\Psi}(\alpha)$) der Hypothesen

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0.$$

Beweis Wählen wir ein beliebiges $\theta_1 \in \Theta$, $\theta_1 \neq \theta_0$. Nach Lemma 10.38 ist φ_c ein modifizierter Neyman-Pearson-Test $\varphi_{K,\lambda}$ für eine spezielle Wahl von K und $\lambda \in \mathbb{R}$. $\varphi_{K,\lambda}$ ist aber nach Satz 10.37 bester Test in $\tilde{\Psi}(\alpha)$ für $H_0 : \theta = \theta_0$ vs. $H'_1 : \theta = \theta_1$. Da φ_c nicht von θ_1 abhängt, ist es bester Test in $\tilde{\Psi}(\alpha)$ für $H_1 : \theta \neq \theta_0$. Da unverfälschte Niveau- α -Tests in $\tilde{\Psi}(\alpha)$ liegen, müssen wir nur zeigen, dass φ_c unverfälscht ist. Da φ_c der beste Test ist, ist er nicht schlechter als der konstante unverfälschte Test $\varphi = \alpha$, das heißt

$$G_n(\varphi_c, \theta) \geq G_n(\varphi, \theta) = \alpha, \quad \theta \neq \theta_0.$$

Somit ist auch φ_c unverfälscht. Der Beweis ist beendet. \square

Bemerkung 10.41 Wir haben gezeigt, dass φ_c der beste Test seines Umfangs ist. Es wäre jedoch noch zu zeigen, dass für beliebiges $\alpha \in (0, 1)$ Konstanten $c_1, c_2, \gamma_1, \gamma_2$ gefunden werden, für die $\mathbb{E}_0 \varphi_c = \alpha$ gilt. Da der Beweis schwierig ist, wird er hier ausgelassen. Im folgenden Beispiel jedoch wird es klar, wie die Parameter $c_1, c_2, \gamma_1, \gamma_2$ zu wählen sind.

Beispiel 10.42 (Zweiseitiger Gauß-Test) Im Beispiel 10.7 haben wir folgenden Test des Erwartungswertes einer normalverteilten Stichprobe (X_1, \dots, X_n) mit unabhängigen und identisch verteilten X_i und $X_i \sim N(\mu, \sigma_0^2)$ bei bekannten Varianzen σ_0^2 betrachtet. Getestet werden die Hypothesen

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0.$$

Der Test $\varphi(x)$ lautet

$$\varphi(x) = \mathbb{1}_{\{x \in \mathbb{R}^n : |T(x)| > z_{1-\alpha/2}\}},$$

wobei

$$T(x) = \sqrt{n} \frac{\bar{x}_n - \mu_0}{\sigma_0}.$$

Zeigen wir, dass φ der beste Test zum Niveau α in $\tilde{\Psi}(\alpha)$ (und somit bester unverfälschter Test) ist. Nach Satz 10.40 müssen wir lediglich prüfen, dass φ als φ_c mit (10.9) dargestellt werden kann, weil die n-dimensionale Normalverteilung mit Dichte f_μ (siehe Beispiel 10.33) zu der einparametrischen Exponentialfamilie mit Statistik

$$T(x) = \sqrt{n} \frac{\bar{x}_n - \mu}{\sigma_0}$$

gehört. Setzen wir $c_1 = z_{1-\alpha/2}$, $c_2 = -z_{1-\alpha/2}$, $\gamma_1 = \gamma_2 = 0$. Damit ist

$$\varphi(x) = \varphi_c(x) = \begin{cases} 1, & \text{falls } |T(x)| > z_{1-\alpha/2}, \\ 0, & \text{falls } |T(x)| \leq z_{1-\alpha/2}. \end{cases}$$

und die Behauptung ist bewiesen, weil aus der in Beispiel 10.7 ermittelten Gütefunktion $G_n(\varphi, \theta)$ von φ ersichtlich ist, dass φ ein unverfälschter Test zum Niveau α ist (und somit $\varphi \in \tilde{\Psi}(\alpha)$).

Bemerkung 10.43 Bisher haben wir immer vorausgesetzt, dass nur *ein* Parameter der Verteilung der Stichprobe (X_1, \dots, X_n) unbekannt ist, um die Theorie des Abschnittes 1.3 über die besten (Neyman-Pearson-) Tests im Fall der einparametrischen Exponentialfamilie aufzustellen zu können. Um jedoch den Fall weiterer unbekannter Parameter betrachten zu können (wie im Beispiel der zweiseitigen Tests des Erwartungswertes der normalverteilten Stichprobe bei unbekannter Varianz (der sog. *t*-Test, vergleiche Abschnitt 1.2.1, 1 (a))), bedarf es einer tiefergehenderen Theorie, die aus Zeitgründen in dieser Vorlesung nicht behandelt wird. Der interessierte Leser findet das Material dann im Buch [?].

10.4 Anpassungstests

Sei eine Stichprobe von unabhängigen, identisch verteilten Zufallsvariablen (X_1, \dots, X_n) gegeben mit $X_i \sim F$ (Verteilungsfunktion) für $i = 1, \dots, n$. Bei den Anpassungstests wird die Hypothese

$$H_0 : F = F_0 \text{ vs. } H_1 : F \neq F_0$$

überprüft, wobei F_0 eine vorgegebene Verteilungsfunktion ist.

Einen Test aus dieser Klasse haben wir bereits in der Vorlesung Stochastik I kennengelernt: den Kolmogorow-Smirnov-Test (vergleiche Bemerkung 3.3.8. 3), Vorlesungsskript Stochastik I).

Jetzt werden weitere nichtparametrische Anpassungstests eingeführt. Der erste ist der χ^2 -Anpassungs-test von K. Pearson.

10.4.1 χ^2 -Anpassungstest

Der Test von Kolmogorov-Smirnov basierte auf dem Abstand

$$D_n = \sup_{x \in \mathbb{R}} | \hat{F}_n(x) - F_0(x) |$$

zwischen der empirischen Verteilungsfunktion der Stichprobe (X_1, \dots, X_n) und der Verteilungsfunktion F_0 . In der Praxis jedoch erscheint dieser Test zu feinfühlig, denn er ist zu sensibel gegenüber Unregelmäßigkeiten in den Stichproben und verwirft H_0 zu oft. Einen Ausweg aus dieser Situation stellt die Vergrößerung der Haupthypothese H_0 dar, auf welcher der folgende χ^2 -Anpassungstest beruht.

Man zerlegt den Wertebereich der Stichprobenvariablen X_i in r Klassen $(a_j, b_j]$, $j = 1, \dots, r$ mit der Eigenschaft

$$-\infty \leq a_1 < b_1 = a_2 < b_2 = \dots = a_r < b_r \leq \infty.$$

Anstelle von X_i , $i = 1, \dots, n$ betrachten wir die sogenannten *Klassenstärken* Z_j , $j = 1, \dots, r$, wobei

$$Z_j = \#\{i : a_j < X_i \leq b_j, 1 \leq i \leq n\}.$$

Lemma 10.44 Der Zufallsvektor $Z = (Z_1, \dots, Z_r)^\top$ ist *multinomialverteilt* mit Parametervektor

$$p = (p_1, \dots, p_{r-1})^\top \in [0, 1]^{r-1},$$

wobei

$$p_j = P(a_j < X_1 \leq b_j) = F(b_j) - F(a_j), \quad j = 1, \dots, r-1, \quad p_r = 1 - \sum_{j=1}^{r-1} p_j.$$

Schreibweise:

$$Z \sim M_{r-1}(n, p)$$

Beweis Es ist zu zeigen, dass für alle Zahlen $k_1, \dots, k_r \in \mathbb{N}_0$ mit $k_1 + \dots + k_r = n$ gilt:

$$P(Z_i = k_i, i = 1, \dots, r) = \frac{n!}{k_1! \cdot \dots \cdot k_r!} p_1^{k_1} \cdot \dots \cdot p_r^{k_r}. \quad (10.10)$$

Da X_i unabhängig und identisch verteilt sind, gilt

$$P\left(X_j \in (a_{i_j}, b_{i_j}], j = 1, \dots, n\right) = \prod_{j=1}^n P\left(a_{i_j} < X_1 \leq b_{i_j}\right) = p_1^{k_1} \cdot \dots \cdot p_r^{k_r},$$

falls die Folge von Intervallen $(a_{i_j}, b_{i_j}]_{j=1, \dots, n}$ das Intervall $(a_i, b_i]$ k_i Mal enthält, $i = 1, \dots, r$. Die Formel (10.10) ergibt sich aus dem Satz der totalen Wahrscheinlichkeit als Summe über die Permutationen von Folgen $(a_{i_j}, b_{i_j}]_{j=1, \dots, n}$ dieser Art. \square

Im Sinne des Lemmas 10.44 werden neue Hypothesen über die Beschaffenheit von F geprüft.

$$H_0 : p = p_0 \text{ vs. } H_1 : p \neq p_0,$$

wobei $p = (p_1, \dots, p_{r-1})^\top$ der Parametervektor der Multinomialverteilung von Z ist, und $p_0 = (p_{01}, \dots, p_{0,r-1})^\top \in (0, 1)^{r-1}$ mit $\sum_{i=1}^{r-1} p_{0i} < 1$. In diesem Fall ist

$$\Lambda_0 = \{F \in \Lambda : F(b_j) - F(a_j) = p_{0j}, \quad j = 1, \dots, r-1\},$$

$\Lambda_1 = \Lambda \setminus \Lambda_0$, wobei Λ die Menge aller Verteilungsfunktionen ist. Um H_0 vs. H_1 zu testen, führen wir die *Pearson-Teststatistik*

$$T_n(x) = \sum_{j=1}^r \frac{(z_j - np_{0j})^2}{np_{0j}}$$

ein, wobei $x = (x_1, \dots, x_n)$ eine konkrete Stichprobe der Daten ist und z_j , $j = 1, \dots, r$ ihre Klassenstärken sind.

Unter H_0 gilt

$$\mathbb{E} Z_j = np_{0j}, \quad j = 1, \dots, r,$$

somit soll H_0 abgelehnt werden, falls $T_n(x)$ ungewöhnlich große Werte annimmt.

Im nächsten Satz zeigen wir, dass $T(X_1, \dots, X_n)$ asymptotisch (für $n \rightarrow \infty$) χ_{r-1}^2 -verteilt ist, was zu folgendem Anpassungstest (χ^2 -Anpassungstest) führt:

$$H_0 \text{ wird verworfen, falls } T_n(x_1, \dots, x_n) > \chi_{r-1, 1-\alpha}^2.$$

Dieser Test ist nach seinem Entdecker *Karl Pearson* (1857-1936) benannt worden.

Satz 10.45 Unter H_0 gilt

$$\lim_{n \rightarrow \infty} P_{p_0} \left(T_n(X_1, \dots, X_n) > \chi^2_{r-1, 1-\alpha} \right) = \alpha, \quad \alpha \in (0, 1),$$

das heißt, der χ^2 -Pearson-Test ist ein asymptotischer Test zum Niveau α .

Beweis Führen wir die Bezeichnung $Z_{nj} = Z_j(X_1, \dots, X_n)$ der Klassenstärken ein, die aus der Stichprobe (X_1, \dots, X_n) entstehen. Nach Lemma 10.44 ist

$$Z_n = (Z_{n1}, \dots, Z_{nr}) \sim M_{r-1}(n, p_0) \text{ unter } H_0.$$

Insbesondere soll $\mathbb{E} Z_{nj} = np_{0j}$ und

$$\text{Cov}(Z_{ni}, Z_{nj}) = \begin{cases} np_{0j}(1 - p_{0j}), & i = j, \\ -np_{0i}p_{0j}, & i \neq j \end{cases}$$

für alle $i, j = 1, \dots, r$ gelten. Da

$$Z_{nj} = \sum_{i=1}^n \mathbb{1}(a_j < X_i \leq b_j), \quad j = 1, \dots, r,$$

ist $Z_n = (Z_{n1}, \dots, Z_{n,r-1})$ eine Summe von n unabhängigen und identisch verteilten Zufallsvektoren $Y_i \in \mathbb{R}^{r-1}$ mit Koordinaten $Y_{ij} = \mathbb{1}(a_j < X_i \leq b_j)$, $j = 1, \dots, r-1$. Daher gilt nach dem multivariaten Grenzwertsatz (der in Lemma 10.46 bewiesen wird), dass

$$Z'_n = \frac{Z_n - \mathbb{E} Z_n}{\sqrt{n}} = \frac{\sum_{i=1}^n Y_i - n\mathbb{E} Y_1}{\sqrt{n}} \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, K),$$

mit $N(0, K)$ eine $(r-1)$ -dimensionale multivariate Normalverteilung (vergleiche Vorlesungsskript WR, Beispiel 3.4.1. 3.) mit Erwartungswertvektor Null und Kovarianzmatrix $K = (\sigma_{ij}^2)$, wobei

$$\sigma_{ij}^2 = \begin{cases} -p_{0i}p_{0j}, & i \neq j, \\ p_{0i}(1 - p_{0j}), & i = j \end{cases}$$

für $i, j = 1, \dots, r-1$ ist. Diese Matrix K ist invertierbar mit $K^{-1} = A = (a_{ij})$,

$$a_{ij} = \begin{cases} \frac{1}{p_{0r}}, & i \neq j, \\ \frac{1}{p_{0i}} + \frac{1}{p_{0r}}, & i = j. \end{cases}$$

Außerdem ist K (als Kovarianzmatrix) symmetrisch und positiv definit. Aus der linearen Algebra ist bekannt, dass es eine invertierbare $(r-1) \times (r-1)$ -Matrix $A^{1/2}$ gibt, mit der Eigenschaft $A = A^{1/2}(A^{1/2})^\top$. Daraus folgt,

$$K = A^{-1} = ((A^{1/2})^\top)^{-1} \cdot (A^{1/2})^{-1}.$$

Wenn wir $(A^{1/2})^\top$ auf Z'_n anwenden, so bekommen wir

$$(A^{1/2})^\top \cdot Z'_n \xrightarrow[n \rightarrow \infty]{d} (A^{1/2})^\top \cdot Y,$$

wobei

$$(A^{1/2})^\top \cdot Y \sim N\left(0, (A^{1/2})^\top \cdot K \cdot A^{1/2}\right) = N(0, I_{r-1})$$

nach der Eigenschaft der multivariaten Normalverteilung, die im Kapitel 2, Satz 11.11 behandelt wird. Des Weiteren wurde hier der Stetigkeitssatz aus der Wahrscheinlichkeitsrechnung benutzt, dass

$$Y_n \xrightarrow[n \rightarrow \infty]{d} Y \implies \varphi(Y_n) \xrightarrow[n \rightarrow \infty]{d} \varphi(Y)$$

für beliebige Zufallsvektoren $\{Y_n\}$, $Y \in \mathbb{R}^m$ und stetige Abbildungen $\varphi : \mathbb{R} \rightarrow \mathbb{R}$. Diesen Satz haben wir in WR für Zufallsvariablen bewiesen (Satz 6.4.3, Vorlesungsskript WR). Die erneute Anwendung des Stetigkeitssatzes ergibt

$$\left| (A^{1/2})^\top Z'_n \right|^2 \xrightarrow[n \rightarrow \infty]{d} |Y|^2 = R \sim \chi^2_{r-1}.$$

Zeigen wir, dass

$$T_n(X_1, \dots, X_n) = \left| (A^{1/2})^\top Z'_n \right|^2.$$

Es gilt:

$$\begin{aligned} \left| (A^{1/2})^\top Z'_n \right|^2 &= ((A^{1/2})^\top Z'_n)^\top ((A^{1/2})^\top Z'_n) = Z'^\top \cdot \underbrace{A^{1/2} \cdot (A^{1/2})^\top}_A Z'_n = Z'^\top A Z'_n \\ &= n \sum_{j=1}^{r-1} \frac{1}{p_{0j}} \left(\frac{Z_{nj}}{n} - p_{0j} \right)^2 + \frac{n}{p_{0r}} \sum_{i=1}^{r-1} \sum_{j=1}^{r-1} \left(\frac{Z_{ni}}{n} - p_{0i} \right) \left(\frac{Z_{nj}}{n} - p_{0j} \right) \\ &= \sum_{j=1}^{r-1} \frac{(Z_{nj} - np_{0j})^2}{np_{0j}} + \frac{n}{p_{0r}} \left(\sum_{j=1}^{r-1} \left(\frac{Z_{nj}}{n} - p_{0j} \right) \right)^2 \\ &= \sum_{j=1}^{r-1} \frac{(Z_{nj} - np_{0j})^2}{np_{0j}} + \frac{n}{p_{0r}} \left(\frac{Z_{nr}}{n} - p_{0r} \right)^2 \\ &= \sum_{j=1}^r \frac{(Z_{nj} - np_{0j})^2}{np_{0j}} = T_n(X_1, \dots, X_n), \end{aligned}$$

weil

$$\begin{aligned}\sum_{j=1}^{r-1} Z_{nj} &= n - Z_{nr}, \\ \sum_{j=1}^{r-1} p_{0j} &= 1 - p_{0r}.\end{aligned}$$

□

Lemma 10.46 (Multivariater zentraler Grenzwertsatz) Sei $\{Y_n\}_{n \in \mathbb{N}}$ eine Folge von unabhängigen und identisch verteilten Zufallsvektoren, mit $\mathbb{E} Y_1 = \mu$ und Kovarianzmatrix K. Dann gilt

$$\frac{\sum_{i=1}^n Y_i - n\mu}{\sqrt{n}} \xrightarrow[n \rightarrow \infty]{d} Y \sim N(0, K). \quad (10.11)$$

Beweis Sei $Y_j = (Y_{j1}, \dots, Y_{jm})^\top$. Nach dem Stetigkeitssatz für charakteristische Funktionen ist die Konvergenz (10.11) äquivalent zu

$$\varphi_n(t) \xrightarrow[n \rightarrow \infty]{} \varphi(t) \quad t \in \mathbb{R}^m, \quad (10.12)$$

wobei

$$\varphi_n(t) = \mathbb{E} e^{itS_n} = \mathbb{E} \exp \left\{ i \sum_{j=1}^m t_j \frac{Y_{1j} + \dots + Y_{nj} - n\mu_j}{\sqrt{n}} \right\}$$

die charakteristische Funktion vom Zufallsvektor

$$S_n = \frac{\sum_{i=1}^n Y_i - n\mu}{\sqrt{n}}$$

und

$$\varphi(t) = e^{-t^\top K t / 2}$$

die charakteristische Funktion der $N(0, K)$ -Verteilung ist. Die Funktion $\varphi_n(t)$ kann in der Form

$$\varphi_n(t) = \mathbb{E} \exp \left\{ i \sum_{i=1}^n \frac{\sum_{j=1}^m t_j (Y_{ij} - \mu_j)}{\sqrt{n}} \right\}, \quad t = (t_1, \dots, t_m)^\top \in \mathbb{R}^m$$

umgeschrieben werden, wobei für die Zufallsvariable

$$L_i := \sum_{j=1}^m t_j (Y_{ij} - \mu_j)$$

gilt:

$$\begin{aligned}\mathbb{E} L_i &= 0, \\ \text{Var } L_i &= \mathbb{E} \left[\sum_{k,j=1}^m t_j(Y_{ij} - \mu_j)(Y_{ik-\mu_k})t_k \right] = t^\top Kt, \quad i \in \mathbb{N}.\end{aligned}$$

Falls $t^\top Kt = 0$, dann gilt $L_i = 0$ fast sicher, für alle $i \in \mathbb{N}$. Hieraus folgt $\varphi_n(t) = \varphi(t) = 1$, also gilt die Konvergenz 10.11.

Falls jedoch $t^\top Kt > 0$, dann kann $\varphi_n(t)$ als charakteristische Funktion der Zufallsvariablen

$$\sum_{i=1}^n L_i / \sqrt{n}$$

an Stelle 1, und $\varphi(t)$ als charakteristische Funktion der eindimensionalen Normalverteilung $N(0, t^\top Kt)$ an Stelle 1 interpretiert werden. Aus dem zentralen Grenzwertsatz für eindimensionale Zufallsvariablen (vergleiche Satz 7.2.1, Vorlesungsskript WR) gilt

$$\sum_{i=1}^n \frac{L_i}{\sqrt{n}} \xrightarrow[n \rightarrow \infty]{d} L \sim N(0, t^\top Kt)$$

und somit

$$\varphi_n(t) = \varphi\left(\sum_{i=1}^n L_i / \sqrt{n}\right)(1) \xrightarrow[n \rightarrow \infty]{} \varphi_L(1) = \varphi(t).$$

Somit ist die Konvergenz (10.11) bewiesen. \square

Bemerkung 10.47 1. Die im letzten Beweis verwendete Methode der Reduktion einer mehrdimensionalen Konvergenz auf den eindimensionalen Fall mit Hilfe von Linearkombinationen von Zufallsvariablen trägt den Namen von *Cramér-Wold*.

2. Der χ^2 -Pearson-Test ist asymptotisch, also für große Stichprobenumfänge, anzuwenden. Aber welches n ist groß genug? Als „Faustregel“ gilt: np_0j soll größer gleich a sein, $a \in (2, \infty)$. Für eine größere Klassenanzahl $r \geq 10$ kann sogar $a = 1$ verwendet werden. Wir zeigen jetzt, dass der χ^2 -Anpassungstest konsistent ist.

Lemma 10.48 Der χ^2 -Pearson-Test ist konsistent, das heißt

$$\forall p \in [0, 1]^{r-1}, p \neq p_0 \text{ gilt: } \lim_{n \rightarrow \infty} P_p \left(T_n(X_1, \dots, X_n) > \chi^2_{r-1, 1-\alpha} \right) = 1$$

Beweis Unter H_1 gilt

$$Z_{nj}/n = \frac{\sum_{i=1}^n \mathbb{1}(a_j < X_i \leq b_j)}{n} \xrightarrow[n \rightarrow \infty]{f.s.} \underbrace{\mathbb{E} \mathbb{1}(a_j < X_1 \leq b_j)}_{=p_j}$$

nach dem starken Gesetz der großen Zahlen. Wir wählen j so, dass $p_j \neq p_{0j}$. Es gilt

$$T_n(X_1, \dots, X_n) \geq \frac{(Z_{nj} - np_{0j})^2}{np_{0j}} \geq n \underbrace{\left(\frac{Z_{nj}}{n} - p_{0j} \right)^2}_{\sim n(p_j - p_{0j})^2} \xrightarrow[n \rightarrow \infty]{f.s.} \infty.$$

Somit ist auch

$$P_p \left(T_n(X_1, \dots, X_n) > \chi^2_{r-1, 1-\alpha} \right) \xrightarrow[n \rightarrow \infty]{f.s.} 1.$$

□

Kapitel 11

Lineare Regression

In Stochastik I betrachteten wir die einfache lineare Regression der Form

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

In Matrix-Form schreiben wir $Y = X\beta + \varepsilon$, wobei $Y = (Y_1, \dots, Y_n)^\top$ der Vektor der Zielzufallsvariablen ist,

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

eine $(n \times 2)$ -Matrix, die die Ausgangsvariablen $x_i, i = 1, \dots, n$ enthält und deshalb *Design-Matrix* genannt wird, $\beta = (\beta_0, \beta_1)^\top$ der Parametervektor und $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$ der Vektor der Störgrößen. Bisher waren oft $\varepsilon_i \sim N(0, \sigma^2)$ für $i = 1, \dots, n$ und $\varepsilon \sim N(0, \cdot \sigma^2)$ multivariat normalverteilt.

Die multivariate (das bedeutet, nicht einfache) lineare Regression lässt eine beliebige $(n \times m)$ -Design-Matrix

$$X = (x_{ij})_{\substack{i=1, \dots, n \\ j=1, \dots, m}}$$

und einen m -dimensionalen Parametervektor $\beta = (\beta_1, \dots, \beta_m)^\top$ zu, für $m \geq 2$. Das heißt, es gilt

$$Y = X\beta + \varepsilon, \tag{11.1}$$

wobei $\varepsilon \sim N(0, K)$ ein multivariat normalverteilter Zufallsvektor der Störgrößen mit Kovarianzmatrix K ist, die im Allgemeinen nicht unabhängig voneinander sind:

$$K \neq \text{diag}(\sigma_1^2, \dots, \sigma_n^2).$$

Das Ziel dieses Kapitels ist es, Schätzer und Tests für β zu entwickeln. Zuvor müssen jedoch die Eigenschaften der multivariaten Normalverteilung untersucht werden.

11.1 Multivariate Normalverteilung

Im Vorlesungsskript Wahrscheinlichkeitsrechnung wurde die multivariate Normalverteilung in Beispiel 3.4.1 folgendermaßen eingeführt:

Definition 11.1 Es sei $X = (X_1, \dots, X_n)^\top$ ein n -dimensionaler Zufallsvektor, $\mu \in \mathbb{R}^n$, K eine symmetrische, positiv definite $(n \times n)$ -Matrix. X ist *multivariat normalverteilt* mit den Parametern μ und K ($X \sim N(\mu, K)$), falls X absolut stetig verteilt ist mit der Dichte

$$f_X(x) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{\det(K)}} \exp \left\{ -\frac{1}{2} (x - \mu)^\top K^{-1} (x - \mu) \right\}, \quad x = (x_1, \dots, x_n)^\top \in \mathbb{R}^n.$$

Wir geben drei weitere Definitionen von $N(\mu, K)$ an und wollen die Zusammenhänge zwischen ihnen untersuchen:

Definition 11.2 Der Zufallsvektor $X = (X_1, \dots, X_n)^\top$ ist multivariat normalverteilt ($X \sim N(\mu, K)$) mit Parametern $\mu \in \mathbb{R}^n$ und K (eine symmetrische, nicht-negativ definite $(n \times n)$ -Matrix), falls die charakteristische Funktion $\varphi_X(t) = \mathbb{E} e^{i(t, X)}$, $t \in \mathbb{R}^n$, gegeben ist durch

$$\varphi_X(t) = \exp \left\{ it^\top \mu - \frac{1}{2} t^\top K t \right\}, \quad t \in \mathbb{R}^n.$$

Definition 11.3 Der Zufallsvektor $X = (X_1, \dots, X_n)^\top$ ist multivariat normalverteilt ($X \sim N(\mu, K)$) mit Parametern $\mu \in \mathbb{R}^n$ und einer symmetrischen, nicht negativ definiten $(n \times n)$ -Matrix K , falls

$$\forall a \in \mathbb{R}^n : \text{die Zufallsvariable } (a, X) = a^\top X \sim N(a^\top \mu, a^\top K a)$$

eindimensional normalverteilt ist.

Definition 11.4 Es sei $\mu \in \mathbb{R}^n$, K eine nicht-negativ definite, symmetrische $(n \times n)$ -Matrix. Ein Zufallsvektor $X = (X_1, \dots, X_n)^\top$ ist multivariat normalverteilt mit Parametern μ und K ($X \sim N(\mu, K)$), falls

$$X \stackrel{d}{=} \mu + C \cdot Y,$$

wobei C eine $(n \times m)$ - Matrix mit $\text{rang}(C) = m$, $K = C \cdot C^\top$ und $Y \sim N(0, I) \in \mathbb{R}^m$ ein m -dimensionaler Zufallsvektor mit unabhängigen und identisch verteilten Koordinaten $Y_j \sim N(0, 1)$ ist, $j = 1, \dots, m$.

Bemerkung: Dies ist das Analogon im eindimensionalen Fall: $Y \sim N(\mu, \sigma^2) \Leftrightarrow Y \stackrel{d}{=} \mu + \sigma X$ mit $X \sim N(0, 1)$.

Übungsaufgabe 11.5 Prüfen Sie, daß die in Definition 11.1 angegebene Dichte

$$f_X(x) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{\det(K)}} \exp \left\{ -\frac{1}{2} (x - \mu)^\top K^{-1} (x - \mu) \right\}, \quad x \in \mathbb{R}^n$$

tatsächlich eine Verteilungsdichte darstellt.

Lemma 11.6 Es seien X und Y n -dimensionale Zufallsvektoren mit charakteristischen Funktionen

$$\begin{aligned}\varphi_X(t) &= \mathbb{E} e^{it^\top X} = \mathbb{E} e^{it^\top X} \\ \varphi_Y(t) &= \mathbb{E} e^{it^\top Y} = \mathbb{E} e^{it^\top Y}\end{aligned}$$

für $t \in \mathbb{R}^n$. Es gelten folgende Eigenschaften:

1. *Eindeutigkeitssatz:*

$$X \stackrel{d}{=} Y \Leftrightarrow \varphi_X(t) = \varphi_Y(t), \quad t \in \mathbb{R}^n$$

2. Falls X und Y unabhängig sind, dann gilt:

$$\varphi_{X+Y}(t) = \varphi_X(t) \cdot \varphi_Y(t), \quad t \in \mathbb{R}^n.$$

ohne Beweis: vergleiche den Beweis des Satzes 5.1.1 (5), Folgerung 5.1.1, Vorlesungsskript WR.

Satz 11.7 1. Die Definitionen 11.2 - 11.4 der multivariaten Normalverteilung sind äquivalent.

2. Die Definitionen 11.1 und 11.4 sind im Falle $n = m$ äquivalent.

Bemerkung 11.8 1. Falls die Matrix K in Definition 11.4 den vollen Rang n besitzt, so besitzt sie die Dichte aus Definition 11.1. Sie wird in dem Fall *regulär* genannt.

2. Falls $\text{Rang}(K) = m < n$, dann ist die Verteilung $N(\mu, K)$ laut Definition 11.4 auf dem m -dimensionalen linearen Unterraum

$$\{y \in \mathbb{R}^n : y = \mu + Cx, x \in \mathbb{R}^m\}$$

konzentriert. $N(\mu, K)$ ist in diesem Fall offensichtlich nicht absolutstetig verteilt und wird daher *singulär* genannt.

Beweis Wir beweisen: Definition 11.3 \Leftrightarrow 11.2 \Leftrightarrow 11.4.

1. (a) Wir zeigen: Die Definitionen 11.2 und 11.3 sind äquivalent. Dazu ist zu zeigen: Für die Zufallsvariable X mit der charakteristischen Funktion

$$\varphi_X(t) = \exp\{it^\top \mu - \frac{1}{2}t^\top Kt\} \Leftrightarrow \forall a \in \mathbb{R}^n : a^\top X \sim N(a^\top \mu, a^\top K a).$$

Es gilt:

$$\varphi_{t^\top X}(1) = \mathbb{E} e^{it^\top X \cdot 1} \stackrel{\varphi_N(\mu, \sigma^2)}{=} \exp\{it^\top \mu - \frac{1}{2}t^\top Kt\} = \varphi_X(t) \quad \forall t \in \mathbb{R}^n.$$

(Dies nennt man das *Verfahren von Cramér-Wold*, vergleiche den multivariaten zentralen Grenzwertsatz).

- (b) Wir zeigen: Die Definitionen 11.2 und 11.4 sind äquivalent. Dazu ist zu zeigen: $X = \mu + C \cdot Y$ (mit μ , C , und Y wie in Definition 11.4) $\Leftrightarrow \varphi_X(t) = \exp\{it^\top \mu - \frac{1}{2}t^\top Kt\}$, wobei $K = C \cdot C^\top$. Es gilt:

$$\begin{aligned} \varphi_{\mu + CY}(t) &= \mathbb{E} e^{i(t, \mu + CY)} = \mathbb{E} e^{it^\top \mu + it^\top CY} = e^{it^\top \mu} \cdot \mathbb{E} e^{i(C^\top \overset{y}{\overbrace{t, Y}})} \\ &\stackrel{Y \sim N(0, I)}{=} e^{it^\top \mu} \cdot \exp\left(-\frac{1}{2}y^\top \cdot y\right) = \exp\left\{it^\top \mu - \frac{1}{2}t^\top C \cdot C^\top t\right\} \\ &= \exp\left\{it^\top \mu - \frac{1}{2}t^\top Kt\right\}, \quad t \in \mathbb{R}^n. \end{aligned}$$

2. Zu zeigen ist: Aus $X \sim N(\mu, K)$ im Sinne von Definition 11.4, $Y \sim N(\mu, K)$ im Sinne der Definition 11.1, $\text{Rang}(K) = n$ folgt, daß $\varphi_X = \varphi_Y$.

Aus der Definition 11.2 (die äquivalent zu Definition 11.4 ist) folgt, daß

$$\begin{aligned} \varphi_X(t) &= \exp\left\{it^\top \mu - \frac{1}{2}t^\top Kt\right\}, \quad t \in \mathbb{R}^n, \\ \varphi_Y(t) &= \mathbb{E} e^{it^\top Y} = \int_{\mathbb{R}^n} e^{it^\top y} \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{\det K}} \cdot \exp\left\{-\frac{1}{2}\overset{x}{\overbrace{(y - \mu)^\top K^{-1} \overbrace{(y - \mu)}}}\right\} dy \\ &= e^{it^\top \mu} \cdot \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{n/2} \sqrt{\det K}} \cdot \exp\left\{it^\top x - \frac{1}{2}x^\top K^{-1} x\right\} dx \end{aligned}$$

Wir diagonalisieren K : \exists orthogonale $(n \times n)$ -Matrix V : $V^\top = V^{-1}$ und $V^\top K V = \text{diag}(\lambda_1, \dots, \lambda_n)$, wobei $\lambda_i > 0$, $i = 1, \dots, n$. Mit der

neuen Substitution: $x = Vz$, $t = Vs$ erhalten wir:

$$\begin{aligned}
\varphi_Y(t) &= \frac{e^{it^\top \mu}}{(2\pi)^{n/2} \sqrt{\det K}} \cdot \int_{\mathbb{R}^n} \exp \left\{ is^\top V^\top V z - \frac{1}{2} z^\top V^\top K^{-1} V z \right\} dz \\
&= \frac{e^{it^\top \mu}}{\sqrt{(2\pi)^n \lambda_1 \cdots \lambda_n}} \cdot \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \exp \left\{ is^\top z - \frac{1}{2} \sum_{i=1}^n \frac{z_i^2}{\lambda_i} \right\} dz_1 \cdots dz_n \\
&= e^{it^\top \mu} \prod_{i=1}^n \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\lambda_i}} e^{is_i z_i - \frac{z_i^2}{2\lambda_i}} dz_i = e^{it^\top \mu} \cdot \prod_{i=1}^n \varphi_{N(0, \lambda_i)}(s_i) = e^{it^\top \mu} \prod_{i=1}^n e^{-\frac{s_i^2 \lambda_i}{2}} \\
&= \exp \left\{ it^\top \mu - \frac{1}{2} s^\top \text{diag}(\lambda_1, \dots, \lambda_n) s \right\} = \exp \left\{ it^\top \mu - \frac{1}{2} (V^\top t)^\top V^\top K V V^\top t \right\} \\
&= \exp \left\{ it^\top \mu - \frac{1}{2} t^\top \underbrace{V V^\top}_{} K \underbrace{V V^\top}_{} t \right\} = \exp \left\{ it^\top \mu - \frac{1}{2} t^\top K t \right\}, t \in \mathbb{R}^n.
\end{aligned}$$

□

11.1.1 Eigenschaften der multivariaten Normalverteilung

Satz 11.9 Es sei $X = (X_1, \dots, X_n) \sim N(\mu, K)$, $\mu \in \mathbb{R}^n$, K symmetrisch und nicht-negativ definit. Dann gelten folgende Eigenschaften:

1. μ ist der *Erwartungswertvektor* von X :

$$\mathbb{E} X = \mu, \quad \text{das heißt: } \mathbb{E} X_i = \mu_i, i = 1, \dots, n.$$

K ist die *Kovarianzmatrix* von X :

$$K = (k_{ij}), \text{ mit } k_{ij} = \text{Cov}(X_i, X_j).$$

2. Jeder Teilvektor $X' = (X_{i_1}, \dots, X_{i_k})^\top$ ($1 \leq i_1 < \dots < i_k \leq n$) von X ist ebenso multivariat normalverteilt, $X' \sim N(\mu', K')$, wobei $\mu' = (\mu_{i_1}, \dots, \mu_{i_k})^\top$, $K' = (k'_{jl}) = (\text{Cov}(X_{i_j}, X_{i_l}))$, $j, l = 1, \dots, k$. Insbesondere sind $X_i \sim N(\mu_i, k_{ii})$, wobei $k_{ii} = \text{Var } X_i$, $i = 1, \dots, n$.
3. Zwei Teilvektoren von X sind unabhängig genau dann, wenn entsprechende Elemente k_{ij} von K , die ihre Kreuzkovarianzen darstellen, Null sind, das heißt: $X' = (X_1, \dots, X_k)^\top$, $X'' = (X_{k+1}, \dots, X_n)$ unabhängig (wobei die Reihenfolge nur wegen der Einfachheit so gewählt wurde, aber unerheblich ist) $\Leftrightarrow k_{ij} = 0$ für $1 \leq i \leq k$, $j > k$ oder $i > k$, $1 \leq j \leq k$.

$$K = \left(\begin{array}{c|c} K' & 0 \\ \hline 0 & K'' \end{array} \right)$$

K' und K'' sind Kovarianzmatrizen von X' bzw. X'' .

4. *Faltungsstabilität*: Falls X und Y unabhängige, n -dimensionale Zufallsvektoren mit $X \sim N(\mu_1, K_1)$ und $Y \sim N(\mu_2, K_2)$ sind, dann ist

$$X + Y \sim N(\mu_1 + \mu_2, K_1 + K_2).$$

Übungsaufgabe 11.10

Beweisen Sie Satz 11.9.

Satz 11.11 (Lineare Transformation von $N(\mu, K)$) Sei $X \sim N(\mu, K)$ ein n -dimensionaler Zufallsvektor, A eine $(m \times n)$ -Matrix mit $\text{Rang}(A) = m \leq n$, $b \in \mathbb{R}^m$. Dann ist der Zufallsvektor $Y = AX + b$ multivariat normalverteilt:

$$Y \sim N(A\mu + b, AKA^\top).$$

Beweis Ohne Beschränkung der Allgemeinheit setzen wir $\mu = 0$ und $b = 0$, weil $\varphi_{Y-a}(t) = e^{-it^\top a} \cdot \varphi_Y(t)$, für $a = A\mu + b$. Es ist zu zeigen:

$$Y = AX, X \sim N(0, K) \Rightarrow Y \sim N(0, AKA^\top)$$

Es ist

$$\begin{aligned}
 \varphi_Y(t) &= \varphi_{AX}(t) = \mathbb{E} e^{it^\top AX} = \mathbb{E} e^{\overbrace{i(X, A^\top t)}^{:=s}} \\
 &\stackrel{(\text{Def. 11.2})}{=} \exp\left\{-\frac{1}{2}s^\top Ks\right\} = \exp\left\{-\frac{1}{2}t^\top AKA^\top t\right\}, t \in \mathbb{R}^n \\
 &\Rightarrow Y \sim N(0, AKA^\top).
 \end{aligned}$$

□

11.1.2 Lineare und quadratische Formen von normalverteilten Zufallsvariablen

Definition 11.12 Seien $X = (X_1, \dots, X_n)^\top$ und $Y = (Y_1, \dots, Y_n)^\top$ Zufallsvektoren auf (Ω, \mathcal{F}, P) , A eine $(n \times n)$ -Matrix aus \mathbb{R}^{n^2} , die symmetrisch ist.

1. $Z = AX$ heißt *lineare Form* von X mit Matrix A .
2. $Z = Y^\top AX$ heißt *bilineare Form* von X und Y mit Matrix A ,

$$Z = \sum_{i=1}^n \sum_{j=1}^n a_{ij} X_j Y_i.$$

3. Die Zufallsvariable $Z = X^\top AX$ (die eine bilineare Form aus 2. mit $Y = X$ ist) heißt *quadratische Form* von X mit Matrix A .

Satz 11.13 Sei $Z = Y^\top AX$ eine bilineare Form von Zufallsvektoren $X, Y \in \mathbb{R}^n$ bzgl. der symmetrischen Matrix A . Falls $\mu_X = \mathbb{E} X$, $\mu_Y = \mathbb{E} Y$ und $K_{XY} = (\text{Cov}(X_i, Y_j))_{i,j=1,\dots,n}$ die Kreuzkovarianzmatrix von X und Y ist, dann gilt:

$$\mathbb{E} Z = \mu_Y^\top A \mu_X + \text{Spur}(AK_{XY}).$$

Beweis

$$\begin{aligned}
 \mathbb{E} Z &= \mathbb{E} \text{Spur}(Z) = \mathbb{E} \text{Spur}(Y^\top AX) \quad (\text{wegen } \text{Spur}(AB) = \text{Spur}(BA)) \\
 &= \mathbb{E} \text{Spur}(AXY^\top) = \text{Spur}(A\mathbb{E}(XY^\top)) \quad (\text{wobei } XY^\top = (X_i Y_j)_{i,j=1,\dots,n}) \\
 &= \text{Spur}\left(A\mathbb{E}\left((X - \mu_X) \cdot (Y - \mu_Y)^\top + \mu_X Y^\top + X \mu_Y^\top - \mu_X \mu_Y^\top\right)\right) \\
 &= \text{Spur}\left(A(K_{XY} + \mu_X \mu_Y^\top + \mu_X \mu_Y^\top - \mu_X \mu_Y^\top)\right) = \text{Spur}\left(AK_{XY} + A\mu_X \mu_Y^\top\right) \\
 &= \text{Spur}(AK_{XY}) + \text{Spur}\left(A\mu_X \cdot \mu_Y^\top\right) \\
 &= \text{Spur}\left(\mu_Y^\top A \mu_X\right) + \text{Spur}(AK_{XY}) = \mu_Y^\top A \mu_X + \text{Spur}(AK_{XY}).
 \end{aligned}$$

□

Folgerung 11.14 Für quadratische Formen gilt

$$\mathbb{E}(X^\top AX) = \mu_X^\top A \mu_X + \text{Spur}(A \cdot K),$$

wobei $\mu_X = \mathbb{E} X$ und K die Kovarianzmatrix von X ist.

Satz 11.15 (Kovarianz quadratischer Formen) Es sei $X \sim N(\mu, K)$ ein n -dimensionaler Zufallsvektor und $A, B \in \mathbb{R}^{n^2}$ zwei symmetrische $(n \times n)$ -Matrizen. Dann gilt Folgendes:

$$\text{Cov}\left(X^\top AX, X^\top BX\right) = 4\mu^\top AKB\mu + 2 \cdot \text{Spur}(AKBK).$$

Lemma 11.16 (gemischte Momente) Es sei $Y = (Y_1, \dots, Y_n)^\top \sim N(0, K)$ ein Zufallsvektor. Dann gilt Folgendes:

$$\mathbb{E}(Y_i Y_j Y_k) = 0,$$

$$\mathbb{E}(Y_i Y_j Y_k Y_l) = k_{ij} \cdot k_{kl} + k_{ik} \cdot k_{jl} + k_{jk} \cdot k_{il}, \quad 1 \leq i, j, k, l \leq n,$$

wobei $K = (k_{ij})_{i,j=1,\dots,n}$ die Kovarianzmatrix von Y ist.

Übungsaufgabe 11.17 Beweisen Sie dieses Lemma.

Beweis von Satz 11.15.

$$\begin{aligned}
\text{Cov}(X^\top AX, X^\top BX) &= \mathbb{E}(X^\top AX \cdot X^\top BX) - \mathbb{E}(X^\top AX) \cdot \mathbb{E}(X^\top BX) \\
&\stackrel{\text{(Folgerung 11.14)}}{=} \mathbb{E}\left(\underbrace{(X-\mu)}_{:=Y} + \mu)^\top A \underbrace{(X-\mu)}_{=:Y} + \mu) \cdot \underbrace{(X-\mu)}_{=:Y} + \mu)^\top B \underbrace{(X-\mu)}_{=:Y} + \mu)\right) \\
&\quad - (\mu^\top A\mu + \text{Spur}(AK))(\mu^\top B\mu + \text{Spur}(BK)) \\
&= \mathbb{E}\left[\left(Y^\top AY + 2\mu^\top AY + \mu^\top A\mu\right)\left(Y^\top BY + 2\mu^\top BY + \mu^\top B\mu\right)\right] \\
&\quad - \mu^\top A\mu \cdot \mu^\top B\mu - \mu^\top A\mu \cdot \text{Spur}(BK) - \mu^\top B\mu \cdot \text{Spur}(AK) \\
&\quad - \text{Spur}(AK) \cdot \text{Spur}(BK) \\
&= \mathbb{E}(Y^\top AY \cdot Y^\top BY) + 2\mathbb{E}(Y^\top AY \cdot \mu^\top BY) + \mathbb{E}(Y^\top AY) \cdot \mu^\top B\mu \\
&\quad + 2\mathbb{E}(\mu^\top AY \cdot Y^\top BY) + 4\mathbb{E}(\mu^\top AY \cdot \mu^\top BY) + 2\underbrace{\mathbb{E}(\mu^\top AY)}_{=0} \mu^\top B\mu \\
&\quad + \mu^\top A\mu \cdot \mathbb{E}(Y^\top BY) + 2\mu^\top A\mu \cdot \underbrace{\mathbb{E}\mu^\top BY}_{=0} + \mu^\top A\mu \cdot \mu^\top B\mu - \mu^\top A\mu \cdot \mu^\top B\mu \\
&\quad - \mu^\top A\mu \cdot \text{Spur}(BK) - \mu^\top B\mu \cdot \text{Spur}(AK) - \text{Spur}(AK) \cdot \text{Spur}(BK) \\
&\stackrel{=0 \text{ (Lemma 11.16)}}{=} \mathbb{E}(Y^\top AY \cdot Y^\top BY) + 2\mu^\top B \underbrace{\mathbb{E}(Y \cdot Y^\top AY)}_{=0} + \mu^\top B\mu \cdot \text{Spur}(AK) \\
&\quad + 2\mu^\top A \underbrace{\mathbb{E}(Y \cdot Y^\top BY)}_{=0} + 4\mu^\top A \underbrace{\mathbb{E}(YY^\top)}_{=K} B\mu + \mu^\top A\mu \cdot \text{Spur}(BK) \\
&\quad - \mu^\top A\mu \cdot \text{Spur}(BK) - \mu^\top B\mu \cdot \text{Spur}(AK) - \text{Spur}(AK) \cdot \text{Spur}(BK) \\
&= \mathbb{E}(Y^\top AY \cdot Y^\top BY) + 4\mu^\top AKB\mu - \text{Spur}(AK) \cdot \text{Spur}(BK).
\end{aligned}$$

Wegen

$$\begin{aligned}
\mathbb{E}(Y^\top AY \cdot Y^\top BY) &= \mathbb{E}\left(\sum_{i,j=1}^n a_{ij} Y_i Y_j \cdot \sum_{k,l=1}^n b_{kl} Y_k Y_l\right) = \sum_{i,j,k,l=1}^n a_{ij} b_{kl} \mathbb{E}(Y_i Y_j Y_k Y_l) \\
&\stackrel{\text{(Lemma 11.16)}}{=} \sum_{i,j,k,l=1}^n a_{ij} b_{kl} (k_{ij} \cdot k_{kl} + k_{ik} \cdot k_{jl} + k_{jk} \cdot k_{il}) \\
&= \sum_{i,j=1}^n a_{ij} k_{ij} \cdot \sum_{k,l=1}^n b_{kl} \cdot k_{kl} + 2 \sum_{i,j,k,l=1}^n a_{ij} \cdot k_{jl} \cdot b_{lk} \cdot k_{ki} \\
&= 2 \cdot \text{Spur}(AKBK) + \text{Spur}(AK) \cdot \text{Spur}(BK)
\end{aligned}$$

folgt:

$$\begin{aligned}\text{Cov} & \left(X^\top AX, X^\top BX \right) \\ & = 2 \cdot \text{Spur}(AKBK) + \text{Spur}(AK) \cdot \text{Spur}(BK) + 4\mu^\top AKB\mu \\ & \quad - \text{Spur}(AK) \cdot \text{Spur}(BK) = 4\mu^\top AKB\mu + 2 \cdot \text{Spur}(AKBK).\end{aligned}$$

□

Folgerung 11.18

$$\text{Var} \left(X^\top AX \right) = 4\mu^\top AKA\mu + 2 \cdot \text{Spur} \left((AK)^2 \right)$$

Satz 11.19 Es seien $X \sim N(\mu, K)$ und $A, B \in \mathbb{R}^{n^2}$ zwei symmetrische Matrizen. Dann gilt:

$$\text{Cov} (BX, X^\top AX) = 2BKA\mu$$

Beweis

$$\begin{aligned}\text{Cov} (BX, X^\top AX) & \stackrel{\text{(Folgerung 11.14)}}{=} \mathbb{E} \left[(BX - B\mu)(X^\top AX - \mu^\top A\mu - \text{Spur}(AK)) \right] \\ & = \mathbb{E} \left[B(X - \mu) \left((X - \mu)^\top A(X - \mu) + 2\mu^\top AX - 2\mu^\top A\mu - \text{Spur}(AK) \right) \right],\end{aligned}$$

denn

$$(X - \mu)^\top A(X - \mu) = X^\top AX - \mu^\top AX - X^\top A\mu + \mu^\top A\mu$$

und mit der Substitution $Z = X - \mu$ (und damit $\mathbb{E} Z = 0$)

$$\begin{aligned}\text{Cov} (BX, X^\top AX) & = \mathbb{E} \left[BZ(Z^\top AZ + 2\mu^\top AZ - \text{Spur}(AK)) \right] \\ & \stackrel{=B\mathbb{E} Z=0}{=} \mathbb{E} (BZ \cdot Z^\top AZ) + 2\mathbb{E} (BZ \cdot \mu^\top AZ) - \text{Spur}(AK) \cdot \overbrace{\mathbb{E} (BZ)}^{\text{Cov } X=K} \\ & = 2\mathbb{E} (BZ \cdot Z^\top A\mu) + \mathbb{E} (BZZ^\top AZ) = 2B \underbrace{\mathbb{E} (ZZ^\top)}_{\text{Cov } X=K} A\mu \\ & \quad + B \cdot \underbrace{\mathbb{E} (ZZ^\top AZ)}_{=0} = 2BKA\mu,\end{aligned}$$

wegen $Z \sim N(0, K)$ und Lemma 11.16 und dem Beweis von Satz 11.15. □

Definition 11.20 Es seien $X_i \sim N(\mu_i, 1)$, $i = 1, \dots, n$ unabhängig. Dann besitzt die Zufallsvariable

$$Y = X_1^2 + \dots + X_n^2$$

die sogenannte *nicht-zentrale* $\chi_{n,\mu}^2$ -Verteilung mit n Freiheitsgraden und dem Nichtzentralitätsparameter

$$\mu = \sum_{i=1}^n \mu_i^2.$$

(in Stochastik I betrachteten wir den Spezialfall der zentralen χ_n^2 -Verteilung mit $\mu = 0$).

In Bemerkung 5.2.1, Vorlesungsskript WR, haben wir momenterzeugende Funktionen von Zufallsvariablen eingeführt. Jetzt benötigen wir für den Beweis des Satzes 11.22 folgenden Eindeutigkeitssatz:

Lemma 11.21 (*Eindeutigkeitssatz für momenterzeugende Funktionen*) Es seien X_1 und X_2 zwei absolutstetige Zufallsvariablen mit momenterzeugenden Funktionen

$$M_{X_i}(t) = \mathbb{E} e^{tX_i}, \quad i = 1, 2,$$

die auf einem Intervall (a, b) definiert sind. Falls f_1 und f_2 die Dichten der Verteilung von X_1 und X_2 sind, dann gilt

$$f_1(x) = f_2(x) \text{ für fast alle } x \in \mathbb{R} \Leftrightarrow M_{X_1}(t) = M_{X_2}(t), t \in (a, b).$$

Ohne Beweis.

Satz 11.22 Die Dichte einer $\chi_{n,\mu}^2$ -verteilten Zufallsvariable X (mit $n \in \mathbb{N}$ und $\mu > 0$) ist gegeben durch die Mischung der Dichten von χ_{n+2J}^2 -Verteilungen mit Mischungsvariable $J \sim \text{Poisson}(\mu/2)$:

$$f_X(x) = \begin{cases} \sum_{j=0}^{\infty} e^{-\mu/2} \frac{(\mu/2)^j}{j!} \cdot \frac{e^{-x/2} x^{\frac{n+2j}{2}-1}}{\Gamma(\frac{n+2j}{2}) \cdot 2^{\frac{n+2j}{2}}}, & x \geq 0, \\ 0, & x < 0. \end{cases} \quad (11.2)$$

Beweis 1. Wir berechnen zuerst $M_X(t)$, $X \sim \chi_{n,\mu}^2$:

$$\begin{aligned} M_X(t) &= \mathbb{E}(e^{tX}) = \mathbb{E} \exp \left\{ t \sum_{i=1}^n X_i^2 \right\} \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} e^{tx_i^2} \cdot e^{-\frac{(x_i - \mu_i)^2}{2}} dx_i \quad \left(t < \frac{1}{2}, X_i \sim N(\mu_i, 1) \right) \end{aligned}$$

Es gilt:

$$\begin{aligned}
tx_i^2 - \frac{(x_i - \mu_i)^2}{2} &= \frac{1}{2}(2tx_i^2 - x_i^2 + 2x_i\mu_i - \mu_i^2) \\
&= -\frac{1}{2} \left(x_i^2(1-2t) - 2x_i\mu_i + \frac{\mu_i^2}{(1-2t)} - \frac{\mu_i^2}{(1-2t)} + \mu_i^2 \right) \\
&= -\frac{1}{2} \left(\left(x_i \cdot \sqrt{1-2t} - \frac{\mu_i}{\sqrt{1-2t}} \right)^2 + \mu_i^2 \left(1 - \frac{1}{1-2t} \right) \right) \\
&= -\frac{1}{2} \left(\frac{(x_i(1-2t) - \mu_i)^2}{1-2t} - \mu_i^2 \cdot \frac{2t}{1-2t} \right)
\end{aligned}$$

Wir substituieren

$$y_i = \frac{(x_i \cdot (1-2t) - \mu_i)}{\sqrt{1-2t}}$$

und erhalten

$$\begin{aligned}
M_X(t) &= (1-2t)^{-\frac{n}{2}} \prod_{i=1}^n \exp \left\{ \mu_i^2 \cdot \left(\frac{t}{1-2t} \right) \right\} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y_i^2}{2}} dy_i}_{=1} \\
&= (1-2t)^{-\frac{n}{2}} \cdot \exp \left\{ \frac{t}{1-2t} \cdot \sum_{i=1}^n \mu_i^2 \right\} = \frac{1}{(1-2t)^{n/2}} \cdot \exp \left\{ \frac{\mu t}{1-2t} \right\}, \quad t < \frac{1}{2}.
\end{aligned}$$

2. Es sei Y eine Zufallsvariable mit der Dichte (11.2). Wir berechnen $M_Y(t)$:

$$\begin{aligned}
M_Y(t) &= \sum_{j=0}^{\infty} e^{-\frac{\mu}{2}} \frac{(\mu/2)^j}{j!} \cdot \underbrace{\int_0^{\infty} e^{xt} \cdot \frac{e^{-\frac{x}{2}} \cdot x^{\frac{n+2j}{2}-1}}{\Gamma\left(\frac{n+2j}{2}\right) \cdot \frac{n+2j}{2}} dx}_{= M_{X_{n+2j}}(t) = \frac{1}{(1-2t)^{(n+2j)/2}} \text{ (Stochastik I, Satz 3.2.1)}} \\
&= \frac{e^{-\frac{\mu}{2}}}{(1-2t)^{\frac{n}{2}}} \cdot \sum_{j=1}^{\infty} \left(\frac{\mu}{2(1-2t)} \right)^j \cdot \frac{1}{j!} \\
&= \frac{1}{(1-2t)^{\frac{n}{2}}} \cdot \exp \left\{ -\frac{\mu}{2} + \frac{\mu}{2(1-2t)} \right\} = \frac{1}{(1-2t)^{\frac{n}{2}}} \cdot \exp \left\{ \frac{\mu \cdot (1-(1-2t))}{2 \cdot (1-2t)} \right\} \\
&= (1-2t)^{-\frac{n}{2}} \cdot \exp \left\{ \frac{\mu t}{1-2t} \right\} \\
\implies M_X(t) &= M_Y(t), \quad t < \frac{1}{2}
\end{aligned}$$

Nach Lemma 11.21 gilt dann, $f_X(x) = f_Y(x)$ für fast alle $x \in \mathbb{R}$.

□

Bemerkung 11.23 1. Die Definition 11.20 kann in folgender Form umgeschrieben werden:

Falls $X \sim N(\vec{\mu},), \vec{\mu} = (\mu_1, \dots, \mu_n)^\top$, dann gilt $|X|^2 = X^\top X \sim \chi_{n,\mu}^2$, wobei $\mu = |\vec{\mu}|^2$.

2. Die obige Eigenschaft kann auf $X \sim N(\vec{\mu}, K)$, mit einer symmetrischen, positiv definiten $(n \times n)$ -Matrix K verallgemeinert werden:

$$X^\top K^{-1} X \sim \chi_{n,\tilde{\mu}}^2, \quad \text{wobei } \tilde{\mu} = \vec{\mu}^\top K^{-1} \vec{\mu},$$

denn weil K positiv definit ist, gibt es ein $K^{\frac{1}{2}}$, sodaß $K = K^{\frac{1}{2}} K^{\frac{1}{2}\top}$. Dann gilt

$$Y = K^{-\frac{1}{2}} X \sim N(K^{-\frac{1}{2}} \mu,), \quad \text{weil } K^{-\frac{1}{2}} K K^{-\frac{1}{2}\top} = K^{-\frac{1}{2}} \cdot K^{\frac{1}{2}} \cdot K^{\frac{1}{2}\top} \cdot K^{-\frac{1}{2}\top} =$$

und daher

$$\begin{aligned} X^\top K^{-1} X &= Y^\top Y \stackrel{\text{Punkt 1}}{\sim} \chi_{n,\tilde{\mu}}^2, \quad \text{mit } \tilde{\mu} = \left(K^{-\frac{1}{2}} \vec{\mu} \right)^\top K^{-\frac{1}{2}} \vec{\mu} = \vec{\mu}^\top K^{-\frac{1}{2}\top} K^{-\frac{1}{2}} \vec{\mu} \\ &= \vec{\mu}^\top K^{-1} \vec{\mu}. \end{aligned}$$

Satz 11.24 Es sei $X \sim N(\mu, K)$, wobei K eine symmetrische, positiv definite $(n \times n)$ -Matrix ist, und sei A eine weitere symmetrische $(n \times n)$ -Matrix mit der Eigenschaft $AK = (AK)^2$ (Idempotenz) und $\text{Rang}(A) = r \leq n$. Dann gilt:

$$X^\top A X \sim \chi_{r,\tilde{\mu}}^2, \quad \text{wobei } \tilde{\mu} = \mu^\top A \mu.$$

Beweis Wir zeigen, daß A nicht negativ definit ist.

$$\begin{aligned} AK &= (AK)^2 = AK \cdot AK \quad | \quad K^{-1} \\ &\implies A = AKA \Rightarrow \forall x \in \mathbb{R}^n : x^\top A x = x^\top AKAx \\ &= (\underbrace{Ax}_y)^\top K (\underbrace{Ax}_y) \geq 0 \text{ wegen der positiven Definitheit von } K. \\ &\implies A \text{ ist nicht negativ definit.} \\ &\implies \exists H : \text{eine } (n \times r)\text{-Matrix mit } \text{Rang}(H) = r : A = HH^\top \end{aligned}$$

Somit gilt

$$X^\top A X = X^\top H \cdot H^\top X = (\underbrace{H^\top X}_Y)^\top \cdot H^\top X = Y^\top Y$$

Es gilt: $Y \sim N(H^\top \mu_{r,r},)$, denn nach Satz 11.11 ist $Y \sim N(H^\top \mu, H^\top KH)$ und $\text{Rang}(H) = r$. Das heißt, $H^\top H$ ist eine invertierbare $(r \times r)$ -Matrix,

und

$$\begin{aligned}
 H^\top K H &= (H^\top H)^{-1} \underbrace{(H^\top H \cdot H^\top K H \cdot (H^\top H))}_{=AKA=A} (H^\top H)^{-1} \\
 &= (H^\top H)^{-1} H^\top \cdot \underbrace{A}_{=HH^T} \cdot H (H^\top H)^{-1} \\
 &\stackrel{=} r
 \end{aligned}$$

Dann ist

$$X^\top A X = |Y|^2 \sim \chi_{r,\tilde{\mu}}^2 \text{ mit } \tilde{\mu} = (H^\top \mu)^2 = \mu^\top H \cdot H^\top \mu = \mu^\top A \mu.$$

□

Satz 11.25 (Unabhängigkeit) Es sei $X \sim N(\mu, K)$ und K eine symmetrische, nicht-negativ definite $(n \times n)$ -Matrix.

1. Es seien A, B $(r_1 \times n)$ bzw. $(r_2 \times n)$ -Matrizen, $r_1, r_2 \leq n$ mit $AKB^\top = 0$. Dann sind die Vektoren AX und BX unabhängig.
2. Sei ferner C eine symmetrische, nicht-negativ definite $(n \times n)$ -Matrix mit der Eigenschaft $AKC = 0$. Dann sind AX und $X^\top CX$ unabhängig.

Beweis 1. Nach Satz 11.9, 3) gilt: AX und BX sind unabhängig $\iff \varphi_{(AX,BX)}(t) = \varphi_{AX}(t) \cdot \varphi_{BX}(t)$, $t = (t_1, t_2)^\top \in \mathbb{R}^{r_1+r_2}$, $t_1 \in \mathbb{R}^{r_1}$, $t_2 \in \mathbb{R}^{r_2}$. Es ist zu zeigen:

$$\varphi_{(AX,BX)}(t) = \mathbb{E} e^{(it_1^\top A + t_2^\top B) \cdot X} \stackrel{!}{=} \mathbb{E} e^{it_1^\top AX} \cdot \mathbb{E} e^{it_2^\top BX}.$$

Es gilt

$$\varphi_{(AX,BX)}(t) = \mathbb{E} e^{i(t_1^\top A + t_2^\top B) \cdot X} \stackrel{(Def. 11.2)}{=} e^{i(t_1^\top A + t_2^\top B) \cdot \mu - \frac{1}{2} \cdot (t_1^\top A + t_2^\top B) \cdot K \cdot (t_1^\top A + t_2^\top B)^\top},$$

und mit

$$\begin{aligned}
 &(t_1^\top A + t_2^\top B) \cdot K \cdot (t_1^\top A + t_2^\top B)^\top \\
 &= (t_1^\top A) K (t_1^\top A)^\top + (t_1^\top A)^\top K (t_2^\top B) + (t_2^\top B) K (t_1^\top A)^\top + (t_2^\top B)^\top K (t_2^\top B)^\top \\
 &= t_1^\top A K A^\top t_1 + t_1^\top \underbrace{A K B^\top}_{=0} \cdot t_2 + t_2^\top \underbrace{B K A^\top}_{=(A K B^\top)^\top = 0} \cdot t_1 + t_2^\top B K B^\top t_2
 \end{aligned}$$

ist

$$\begin{aligned}
 \varphi_{(AX,BX)}(t) &= e^{it_1^\top A - \frac{1}{2} t_1^\top A K A^\top t_1} \cdot e^{it_2^\top B - \frac{1}{2} t_2^\top B K B^\top t_2} \\
 &= \varphi_{AX}(t_1) \cdot \varphi_{BX}(t_2), \quad t_1 \in \mathbb{R}^{r_1}, t_2 \in \mathbb{R}^{r_2}
 \end{aligned}$$

2. C ist symmetrisch, nicht-negativ definit \Rightarrow Es gibt eine $(n \times r)$ -Matrix H mit $\text{Rang}(H) = r \leq n$ und $C = HH^\top$, $\Rightarrow H^\top H$ hat Rang r und ist somit invertierbar. Dann gilt:

$$X^\top CX = X^\top HH^\top X = (H^\top X)^\top \cdot H^\top X = |H^\top X|^2.$$

Falls AX und $H^\top X$ unabhängig sind, dann sind auch AX und $X^\top CX = |H^\top X|^2$ unabhängig, nach dem Transformationssatz für Zufallsvektoren. Nach 1) sind AX und $H^\top X$ unabhängig, falls $AK(H^\top)^\top = AKH = 0$. Da nach Voraussetzung

$$AKC = AKH \cdot H^\top = 0 \Rightarrow AKH \cdot H^\top H = 0,$$

da aber $\exists (H^\top H)^{-1}$, folgt, daß

$$\begin{aligned} 0 &= AKH \cdot H^\top H \cdot (H^\top H)^{-1} = AKH \Rightarrow AKH = 0 \\ &\Rightarrow AX \text{ und } H^\top X \text{ sind unabhängig} \\ &\Rightarrow AX \text{ und } X^\top CX \text{ sind unabhängig.} \end{aligned}$$

□

11.2 Multivariate lineare Regressionsmodelle mit vollem Rang

Die *multivariate lineare Regression* hat die Form

$$Y = X\beta + \varepsilon,$$

wobei $Y = (Y_1, \dots, Y_n)^\top$ der Zufallsvektor der Zielvariablen ist,

$$X = (x_{ij})_{\substack{i=1, \dots, n \\ j=1, \dots, m}}$$

ist eine deterministische *Design-Matrix* mit vollem Rang, $\text{Rang}(X) = r = m \leq n$, $\beta = (\beta_1, \dots, \beta_m)^\top$ ist der *Parametervektor* und $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$ ist der Zufallsvektor der *Störgrößen*, mit $\mathbb{E} \varepsilon_i = 0$, $\text{Var } \varepsilon_i = \sigma^2 > 0$. Das Ziel dieses Abschnittes wird sein, β und σ^2 geeignet zu schätzen.

11.2.1 Methode der kleinsten Quadrate

Sei $X = (X_1, \dots, X_m)$, wobei die deterministischen Vektoren $X_j = (x_{1j}, x_{2j}, \dots, x_{nj})^\top$, $j = 1, \dots, m$ einen m -dimensionalen linearen Unterraum $L_X = \langle X_1, \dots, X_m \rangle$ aufspannen. Sei

$$e(\beta) = \frac{1}{n} |Y - X\beta|^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - x_{i1}\beta_1 - \dots - x_{im}\beta_m)^2$$

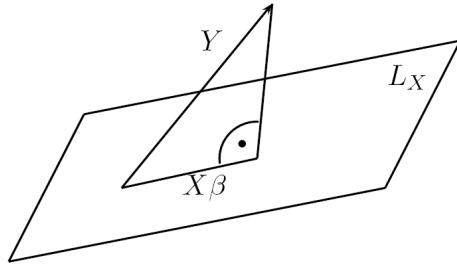
die mittlere quadratische Abweichung zwischen Y und $X\beta$.

Der MKQ-Schätzer $\hat{\beta}$ für β ist definiert durch

$$\hat{\beta} = \operatorname{argmin}(e(\beta)). \quad (11.3)$$

Warum existiert eine Lösung $\beta \in \mathbb{R}^m$ des quadratischen Optimierungsproblems (11.3)? Geometrisch kann $X\hat{\beta}$ als die orthogonale Projektion des Datenvektors Y auf den linearen Unterraum L_X interpretiert werden. Formal zeigen wir die Existenz der Lösung mit folgendem Satz.

Abbildung 11.1: Projektion auf den linearen Unterraum L_X



Satz 11.26 Unter den obigen Voraussetzungen existiert der eindeutig bestimmte MKQ-Schätzer $\hat{\beta}$, der die Lösung der sogenannten *Normalengleichung* ist:

$$X^\top X\beta = X^\top Y. \quad (11.4)$$

Daher gilt:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

Beweis Die notwendige Bedingung für die Existenz des Minimums ist $e'(\beta) = 0$, das heißt

$$e'(\beta) = \left(\frac{\partial e(\beta)}{\partial \beta_1}, \dots, \frac{\partial e(\beta)}{\partial \beta_m} \right)^\top = 0.$$

Es gilt:

$$e'(\beta) = \frac{2}{n} (X^\top X\beta - X^\top Y)$$

$\implies \hat{\beta}$ ist eine Lösung der Normalengleichung $X^\top X\beta = X^\top Y$. Wir zeigen die hinreichende Bedingung des Minimums:

$$e''(\beta) = \left(\frac{\partial^2 e(\beta)}{\partial \beta_i \partial \beta_j} \right)_{i,j=1,\dots,m} = \frac{2}{n} X^\top X.$$

$X^\top X$ ist symmetrisch und positiv definit, weil X einen vollen Rang hat:

$$\forall y \neq 0, y \in \mathbb{R}^m : \quad y^\top X^\top X y = (Xy)^\top X y = |Xy|^2 > 0$$

und aus $y \neq 0 \implies Xy \neq 0$, folgt, daß $e''(\beta)$ positiv definit ist. Also ist $X^\top X$ invertierbar. Das heißt, $\hat{\beta}$ ist der Minimumspunkt von $e(\beta)$. Den Schätzer $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ bekommt man, indem man die Normalengleichung $X^\top X\beta = X^\top Y$ von links mit $(X^\top X)^{-1}$ multipliziert. \square

Beispiel 11.27 1. *Einfache lineare Regression*

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad m = 2, \beta = (\beta_1, \beta_2)^\top, Y = X\beta + \varepsilon$$

$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ ergibt den MKQ-Schätzer aus der Stochastik I

$$\hat{\beta}_2 = \frac{S_{XY}^2}{S_{XX}^2}, \quad \hat{\beta}_1 = \bar{Y}_n - \bar{x}_n \hat{\beta}_2,$$

wobei

$$\begin{aligned} \bar{x}_n &= \frac{1}{n} \sum_{i=1}^n x_i, & \bar{Y}_n &= \frac{1}{n} \sum_{i=1}^n Y_i \\ S_{XY}^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n) \\ S_{XX}^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \end{aligned}$$

Übungsaufgabe 11.28 Beweisen Sie dies!

2. *Multiple lineare Regression*

$Y = X\beta + \varepsilon$ mit Designmatrix

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} \text{ für } \beta = (\beta_0, \beta_1, \dots, \beta_m)^\top.$$

Der MKQ-Schätzer $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ ist offensichtlich ein linearer Schätzer bezüglich Y .

Wir werden jetzt zeigen, daß $\hat{\beta}$ der *beste lineare, erwartungstreue Schätzer* von β (im Englischen *BLUE = best linear unbiased estimator*) in der Klasse

$$\mathcal{L} = \left\{ \tilde{\beta} = AY + b : \mathbb{E} \tilde{\beta} = \beta \right\}$$

aller linearen erwartungstreuen Schätzer ist.

Satz 11.29 (*Güteeigenschaften des MKQ-Schätzers $\hat{\beta}$*) Es sei $Y = X\beta + \varepsilon$ ein multivariates lineares Regressionsmodell mit vollem Rang m und Störgrößen $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$, die folgende Voraussetzungen erfüllen:

$$\mathbb{E} \varepsilon = 0, \quad \text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij}, \quad i, j = 1, \dots, n \text{ für ein } \sigma^2 \in (0, \infty).$$

Dann gilt Folgendes:

1. Der MKQ-Schätzer $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ ist erwartungstreu: $\mathbb{E} \hat{\beta} = \beta$.
2. $\text{Cov}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}$
3. $\hat{\beta}$ besitzt die minimale Varianz:

$$\forall \tilde{\beta} \in \mathcal{L} : \quad \text{Var } \tilde{\beta}_j \geq \text{Var } \hat{\beta}_j, \quad j = 1, \dots, m.$$

Beweis 1. Es gilt:

$$\begin{aligned} \mathbb{E} \hat{\beta} &= \mathbb{E} \left[(X^\top X)^{-1} X^\top (X\beta + \varepsilon) \right] \\ &= (X^\top X)^{-1} \cdot X^\top X \cdot \beta + (X^\top X)^{-1} X^\top \cdot \underbrace{\mathbb{E} \varepsilon}_{=0} \\ &= \beta \quad \forall \beta \in \mathbb{R}^m. \end{aligned}$$

2. Für alle $\tilde{\beta} = AY + b \in \mathcal{L}$ gilt:

$$\begin{aligned} \beta &= \mathbb{E} \tilde{\beta} = A\mathbb{E} Y + b = AX\beta + b \quad \forall \beta \in \mathbb{R}^m. \\ \implies b &= 0, \quad AX = . \\ \implies \tilde{\beta} &= AY = A(X\beta + \varepsilon) = AX\beta + A\varepsilon \\ &= \beta + A\varepsilon. \end{aligned}$$

Für

$$\hat{\beta} = \underbrace{\left(X^\top X \right)^{-1}}_{=A} X^\top Y$$

gilt:

$$\begin{aligned} \text{Cov } \hat{\beta} &= \left(\mathbb{E} \left((\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j) \right) \right)_{i,j=1,\dots,m} \\ &= \mathbb{E} \left(A\varepsilon \cdot (A\varepsilon)^\top \right) = \mathbb{E} \left(A\varepsilon \varepsilon^\top A^\top \right) = A \mathbb{E} \left(\varepsilon \varepsilon^\top \right) \cdot A^\top \\ &= A \cdot \sigma^2 A^\top = \sigma^2 A A^\top = \sigma^2 \left(X^\top X \right)^{-1} X^\top \left(\left(X^\top X \right)^{-1} X^\top \right)^\top \\ &= \sigma^2 \left(X^\top X \right)^{-1} X^\top X \left(X^\top X \right)^{-1} = \sigma^2 \left(X^\top X \right)^{-1}. \end{aligned}$$

3. Sei $\tilde{\beta} \in \mathcal{L}$, $\tilde{\beta} = \beta + A\varepsilon$. Zu zeigen ist, daß

$$\left(\text{Cov } (\tilde{\beta}) \right)_{ii} = \sigma^2 (AA^\top)_{ii} \geq \left(\text{Cov } (\hat{\beta}) \right)_{ii} = \sigma^2 (X^\top X)_{ii}^{-1}, \quad i = 1, \dots, m.$$

Sei $D = A - (X^\top X)^{-1} X^\top$, dann folgt: $A = D + (X^\top X)^{-1} X^\top$,

$$\begin{aligned} AA^\top &= \left(D + \left(X^\top X \right)^{-1} X^\top \right) \left(D^\top + X \left(X^\top X \right)^{-1 \top} \right) \\ &= DD^\top + \left(X^\top X \right)^{-1}, \text{ weil} \end{aligned}$$

$$\begin{aligned} DX \left(X^\top X \right)^{-1} &= \left(\underbrace{AX}_{=} - \underbrace{\left(X^\top X \right)^{-1} X^\top X}_{=} \right) \left(X^\top X \right)^{-1} = 0 \\ \left(X^\top X \right)^{-1} X^\top D^\top &= \left(X^\top X \right)^{-1} X^\top \left(A^\top - X \left(X^\top X \right)^{-1 \top} \right) \\ &= \left(X^\top X \right)^{-1} \left(\underbrace{\left(AX \right)^\top}_{=} - \underbrace{X^\top X \left(X^\top X \right)^{-1}}_{=} \right) = 0. \end{aligned}$$

$$\begin{aligned} \implies (AA^\top)_{ii} &= \underbrace{(DD^\top)_{ii}}_{\geq 0} + \left(X^\top X \right)^{-1}_{ii} \geq \left(X^\top X \right)^{-1}_{ii} \\ \implies \text{Var } \hat{\beta}_i &\leq \text{Var } \tilde{\beta}_i, \quad i = 1, \dots, m. \end{aligned}$$

□

Satz 11.30 Es sei $\hat{\beta}_n$ der MKQ-Schätzer im oben eingeführten multivariaten linearen Regressionsmodell. Sei $\{a_n\}_{n \in \mathbb{N}}$ eine Zahlenfolge mit $a_n \neq 0$,

$n \in \mathbb{N}$, $a_n \rightarrow 0$ ($n \rightarrow \infty$). Es wird vorausgesetzt, daß eine invertierbare $(m \times m)$ -Matrix Q existiert mit

$$Q = \lim_{n \rightarrow \infty} a_n (X_n^\top X_n)^{-1}.$$

Dann ist $\hat{\beta}_n$ schwach konsistent:

$$\hat{\beta}_n \xrightarrow[n \rightarrow \infty]{p} \beta.$$

Beweis

$$\hat{\beta}_n \xrightarrow[n \rightarrow \infty]{p} \beta \iff P(|\hat{\beta}_n - \beta| > \varepsilon) \xrightarrow{n \rightarrow \infty} 0 \quad \forall \varepsilon > 0.$$

$$\begin{aligned} P(|\hat{\beta}_n - \beta| > \varepsilon) &= P(|\hat{\beta}_n - \beta|^2 > \varepsilon^2) \\ &= P\left(\sum_{i=1}^m |\hat{\beta}_{in} - \beta_i|^2 > \varepsilon^2\right) \leq P\left(\bigcup_{i=1}^m \left\{|\hat{\beta}_{in} - \beta_i|^2 > \frac{\varepsilon^2}{m}\right\}\right) \\ &\leq \sum_{i=1}^m P\left(|\hat{\beta}_{in} - \beta_i| > \frac{\varepsilon}{\sqrt{m}}\right) \\ &\leq m \sum_{i=1}^m \frac{\text{Var } \hat{\beta}_{in}}{\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0, \quad (\text{aus der Ungleichung von Tschebyschew}) \\ &\text{falls } \text{Var } \hat{\beta}_{in} \xrightarrow{n \rightarrow \infty} 0, \quad i = 1, \dots, m. \end{aligned}$$

$\text{Var } \hat{\beta}_{in}$ ist ein Diagonaleintrag von der Matrix

$$\text{Cov } \hat{\beta}_n \stackrel{(Satz 11.29)}{=} \sigma^2 (X_n^\top X_n)^{-1}.$$

Wenn wir zeigen, daß $\text{Cov } \hat{\beta}_n \xrightarrow{n \rightarrow \infty} 0$, ist der Satz bewiesen.

Es existiert

$$Q^{-1} = \lim_{n \rightarrow \infty} \frac{1}{a_n} (X_n^\top X_n)^{-1}$$

und damit gilt:

$$\begin{aligned} \lim_{n \rightarrow \infty} \text{Cov } \hat{\beta}_n &= \sigma^2 \lim_{n \rightarrow \infty} (X_n^\top X_n)^{-1} = \sigma^2 \lim_{n \rightarrow \infty} a_n \cdot \frac{1}{a_n} (X_n^\top X_n)^{-1} \\ &= 0 \cdot Q^{-1} \cdot \sigma^2 = 0. \end{aligned}$$

□

11.2.2 Schätzer der Varianz σ^2

Wir führen den Schätzer $\hat{\sigma}^2$ für die Varianz σ^2 der Störgrößen ε_i folgendermaßen ein:

$$\hat{\sigma}^2 = \frac{1}{n-m} |Y - X\hat{\beta}|^2. \quad (11.5)$$

Dies ist eine verallgemeinerte Version des Varianzschätzers aus der einfachen linearen Regression, die wir bereits in Stochastik I kennengelernten. Dabei ist $\hat{Y} = Y - X\hat{\beta}$ der Vektor der Residuen.

Satz 11.31 (Erwartungstreue) Der Varianzschätzer

$$\hat{\sigma}^2 = \frac{1}{n-m} |Y - X\hat{\beta}|^2$$

ist erwartungstreu. Das heißt,

$$\mathbb{E} \hat{\sigma}^2 = \sigma^2.$$

Beweis

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-m} (Y - X\hat{\beta})^\top (Y - X\hat{\beta}) \\ &= \frac{1}{n-m} (Y - X(X^\top X)^{-1}X^\top Y)^\top (Y - X(X^\top X)^{-1}X^\top Y) \\ &= \frac{1}{n-m} (DY)^\top DY \end{aligned}$$

wobei $D = -X(X^\top X)^{-1}X^\top$ eine $(n \times n)$ -Matrix ist. Dann ist

$$\hat{\sigma}^2 = \frac{1}{n-m} Y^\top D^\top DY = \frac{1}{n-m} Y^\top D^2 Y = \frac{1}{n-m} Y^\top DY, \text{ falls}$$

$D^\top = D$ und $D^2 = D$ (das heißt, daß D symmetrisch und idempotent ist). Tatsächlich gilt:

$$\begin{aligned} D^\top &= - (X^\top)^\top (X^\top X)^{-1} X^\top = -X (X^\top X)^{-1} X^\top = D. \\ D^2 &= (-X(X^\top X)^{-1}X^\top) (-X(X^\top X)^{-1}X^\top) \\ &= -2X(X^\top X)^{-1}X^\top + X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top \\ &= -X(X^\top X)^{-1}X^\top = D. \end{aligned}$$

Weiterhin gilt:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-m} \cdot \text{Spur}(Y^\top DY) = \frac{1}{n-m} \cdot \text{Spur}(DYY^\top) \\ \implies \mathbb{E} \hat{\sigma}^2 &= \frac{1}{n-m} \cdot \text{Spur}(D\mathbb{E}(YY^\top)) = \frac{\sigma^2}{n-m} \cdot \text{Spur}(D), \end{aligned}$$

denn

$$\begin{aligned} \text{Spur} & \left(D \cdot \mathbb{E} (YY^\top) \right) \\ &= \text{Spur} \left(D(X\beta)(X\beta)^\top + DX\beta \underbrace{\mathbb{E}\varepsilon^\top}_{=0} + D \underbrace{\mathbb{E}\varepsilon}_{=0} (X\beta)^\top + D \cdot \underbrace{\mathbb{E}\varepsilon\varepsilon^\top}_{=\text{Cov } \varepsilon = \sigma^2} \right) \\ &= \text{Cov } \varepsilon = \sigma^2. \end{aligned}$$

und

$$\begin{aligned} DX &= \left(-X (X^\top X)^{-1} X^T \right) X \\ &= X - X (X^\top X)^{-1} X^\top X = X - X = 0. \end{aligned}$$

Es bleibt zu zeigen, daß $\text{Spur}(D) = n - m$:

$$\begin{aligned} \text{Spur}(D) &= \text{Spur} \left(-X (X^\top X)^{-1} X^\top \right) = \text{Spur}() - \text{Spur} \left(X (X^\top X)^{-1} X^\top \right) \\ &= n - \underbrace{\text{Spur} \left(X^\top X \cdot (X^\top X)^{-1} \right)}_{\text{eine } (m \times m)\text{-Matrix}} = n - m. \end{aligned}$$

□

11.2.3 Maximum-Likelihood-Schätzer für β und σ^2

Um Maximum-Likelihood-Schätzer für β und σ^2 bzw. Verteilungseigenschaften der MKQ-Schätzer $\hat{\beta}$ und $\hat{\sigma}^2$ herleiten zu können, muß die Verteilung von ε bzw. Y präzisiert werden. Wir werden ab sofort normalverteilte Störgrößen betrachten, die unabhängig und identisch verteilt sind:

$$\varepsilon \sim N(0, \sigma^2), \quad \sigma^2 > 0.$$

Daraus folgt:

$$Y \sim N(X\beta, \sigma^2).$$

Wie sieht die Verteilung der MKQ-Schätzer $\hat{\beta}$ und $\hat{\sigma}^2$ aus? Da $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ linear von Y abhängt, erwartungstreu ist und die $\text{Cov } \hat{\beta} = \hat{\sigma}^2 (X^\top X)^{-1}$ besitzt, gilt:

$$\hat{\beta} \sim N \left(\beta, \sigma^2 (X^\top X)^{-1} \right)$$

Berechnen wir nun Maximum-Likelihood-Schätzer für β und σ^2 , und zwar $\tilde{\beta}$ und $\tilde{\sigma}^2$. Dann zeigen wir, daß sie im Wesentlichen mit den MKQ-Schätzern übereinstimmen.

$$\begin{aligned} \tilde{\beta} &= \hat{\beta}, \\ \tilde{\sigma}^2 &= \frac{n-m}{n} \hat{\sigma}^2. \end{aligned}$$

Betrachten wir zunächst die Likelihood-Funktion von Y :

$$L(y, \beta, \sigma^2) = f_Y(y) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \cdot \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)^\top (y - X\beta) \right\}$$

und die Log-Likelihood-Funktion

$$\log L(y, \beta, \sigma^2) = -\frac{n}{2} \log(2\pi) - \underbrace{\frac{n}{2} \log(\sigma^2)}_{:=g} - \frac{1}{2\sigma^2} |y - X\beta|^2.$$

Die Maximum-Likelihood-Schätzer sind dann

$$(\tilde{\beta}, \tilde{\sigma}^2) = \underset{\beta \in \mathbb{R}^m, \sigma^2 > 0}{\operatorname{argmax}} \log L(y, \beta, \sigma^2),$$

sofern sie existieren.

Satz 11.32 (*Maximum-Likelihood-Schätzung von $\tilde{\beta}$ und $\tilde{\sigma}^2$*) Es existieren eindeutig bestimmte Maximum-Likelihood-Schätzer für β und σ^2 , die folgendermaßen aussehen:

$$\begin{aligned} \tilde{\beta} &= \hat{\beta} = (X^\top X)^{-1} X^\top Y \\ \tilde{\sigma}^2 &= \frac{n-m}{n} \hat{\sigma}^2 = \frac{1}{n} |Y - X\tilde{\beta}|^2. \end{aligned}$$

Beweis Wir fixieren $\sigma^2 > 0$ und suchen

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^m}{\operatorname{argmax}} \log L(Y, \beta, \sigma^2) = \underset{\beta \in \mathbb{R}^m}{\operatorname{argmin}} |Y - X\beta|^2,$$

woraus folgt, daß $\tilde{\beta}$ mit dem bekannten MKQ-Schätzer $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ identisch ist, der nicht von σ^2 abhängt. Berechnen wir jetzt

$$\tilde{\sigma}^2 = \underset{\sigma^2 > 0}{\operatorname{argmax}} \log L(Y, \tilde{\beta}, \sigma^2) = \underset{\sigma^2 > 0}{\operatorname{argmax}} g(\sigma^2).$$

Es gilt

$$g(\sigma^2) \xrightarrow{\sigma^2 \rightarrow +\infty} -\infty, \quad g(\sigma^2) \xrightarrow{\sigma^2 \rightarrow 0} -\infty,$$

weil $|Y - X\beta|^2 \neq 0$, dadurch, daß $Y \sim N(X\beta, \sigma^2) \in \{Xy : y \in \mathbb{R}^m\}$ mit Wahrscheinlichkeit Null. Da

$$g'(\sigma^2) = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{|Y - X\beta|}{2(\sigma^2)^2} = 0, \text{ ist } \tilde{\sigma}^2 = \frac{1}{n} |Y - X\tilde{\beta}|^2$$

ein Maximumpunkt von $g(\sigma^2)$, das heißt, $\tilde{\sigma}^2$ ist ein Maximum-Likelihood-Schätzer für σ^2 . \square

Satz 11.33 Unter den obigen Voraussetzungen gilt:

1. $\mathbb{E} \tilde{\sigma}^2 = \frac{n-m}{n} \sigma^2$, das heißt, $\tilde{\sigma}^2$ ist nicht erwartungstreu; allerdings ist er asymptotisch unverzerrt.
2. $\frac{n}{\sigma^2} \tilde{\sigma}^2 \sim \chi_{n-m}^2$, $\frac{n-m}{\sigma^2} \hat{\sigma}^2 \sim \chi_{n-m}^2$.

Beweis 1. Trivial (vergleiche den Beweis von Satz 11.31)

2. Wir zeigen den Satz nur für $\hat{\sigma}^2$.

$$\begin{aligned} \frac{n-m}{\sigma^2} \hat{\sigma}^2 &= \frac{1}{\sigma^2} |Y - X\hat{\beta}|^2 \\ &= \frac{1}{\sigma^2} Y^\top \underbrace{D}_{=D^2} Y \quad (\text{nach dem Beweis von Satz 11.31}) \\ &= \frac{1}{\sigma^2} (DY)^\top DY = \frac{1}{\sigma^2} (D(X\beta + \varepsilon))^\top \cdot D(X\beta + \varepsilon) \\ &= \frac{1}{\sigma^2} (D\varepsilon)^\top D\varepsilon = \left(\frac{\varepsilon^\top}{\sigma} \right) D \left(\frac{\varepsilon}{\sigma} \right), \end{aligned}$$

wobei

$$\left(\frac{\varepsilon}{\sigma} \right) \sim N(0,).$$

Nach Satz 11.24 gilt

$$\frac{\varepsilon^\top}{\sigma} D \frac{\varepsilon}{\sigma} \sim \chi_r^2,$$

wobei $r = \text{Rang}(D)$, weil $D = D$ idempotent ist. Falls $r = n - m$, dann ist $\frac{n-m}{\hat{\sigma}^2} \sim \chi_{n-m}^2$. Zeigen wir, daß $\text{Rang}(D) = r = n - m$. Aus der linearen Algebra ist bekannt, daß $\text{Rang}(D) = n - \dim(\text{Kern}(D))$. Wir zeigen, daß $\text{Kern}(D) = \{Xx : x \in \mathbb{R}^m\}$ und damit $\dim(\text{Kern}(D)) = m$, weil $\text{Rang}(X) = m$. Es ist $\{Xx : x \in \mathbb{R}^n\} \subseteq \text{Kern}(D)$, da

$$DX = (-X(X^\top X)^{-1}X^\top)X = X - (X^\top X)^{-1}X^\top X = 0.$$

und $\text{Kern}(D) \subseteq \{Xx : x \in \mathbb{R}^m\}$, weil

$$\begin{aligned} \forall y \in \text{Kern}(D) : \quad Dy &= 0 \iff (-X(X^\top X)^{-1}X^\top)y = 0 \\ \iff y &= X \cdot \underbrace{(X^\top X)^{-1}X^\top Y}_x = Xx \in \{Xx : x \in \mathbb{R}^m\}. \end{aligned}$$

□

Satz 11.34 Sei $Y = X\beta + \varepsilon$ ein multivariates lineares Regressionsmodell mit $Y = (Y_1, \dots, Y_n)^\top$, Designmatrix X mit $\text{Rang}(X) = m$, $\beta = (\beta_1, \dots, \beta_m)^\top$, $\varepsilon \sim N(0, \sigma^2)$. Dann sind die Schätzer $\hat{\beta} = (X^\top X)^{-1}X^\top Y$ für β bzw. $\hat{\sigma}^2 = \frac{1}{n-m}|Y - X\hat{\beta}|^2$ für σ^2 unabhängig voneinander.

Beweis In diesem Beweis verwenden wir den Satz 11.25, für dessen Anwendung wir $\hat{\beta}$ als lineare und $\hat{\sigma}^2$ als quadratische Form von ε darstellen. Es ist in den Beweisen der Sätze 11.29 und 11.33 gezeigt worden, daß

$$\begin{aligned}\hat{\beta} &= \beta + \underbrace{(X^\top X)^{-1} X^\top}_{=:A} \varepsilon, \\ \hat{\sigma}^2 &= \frac{1}{n-m} \varepsilon^\top D \varepsilon, \text{ wobei } D = -X(X^\top X)^{-1} X^\top.\end{aligned}$$

Zusätzlich gilt $AD = 0$, weil nach dem Beweis des Satzes 11.31

$$(AD)^\top = D^\top A^\top = \underbrace{D \cdot X}_{=0} ((X^\top X)^{-1})^\top = 0.$$

Da $\varepsilon \sim N(0, \sigma^2)$, folgt daraus

$$A\sigma^2 D = 0.$$

Deshalb sind die Voraussetzungen des Satzes 11.25 erfüllt, und $\hat{\beta}$ und $\hat{\sigma}^2$ sind unabhängig. \square

11.2.4 Tests für Regressionsparameter

In diesem Abschnitt wird zunächst die Hypothese

$$H_0 : \beta = \beta_0 \text{ vs. } H_1 : \beta \neq \beta_0$$

für ein $\beta_0 \in \mathbb{R}^m$ getestet. Dafür definieren wir die Testgröße

$$T = \frac{(\hat{\beta} - \beta_0)^\top X^\top X (\hat{\beta} - \beta_0)}{m\hat{\sigma}^2}.$$

Man kann zeigen (vergleiche Satz 11.36), daß unter H_0 gilt:

$$T \sim F_{m, n-m}.$$

Daraus folgt, daß H_0 abgelehnt werden soll, falls $T > F_{m, n-m, 1-\alpha}$, wobei $F_{m, n-m, 1-\alpha}$ das $(1 - \alpha)$ -Quantil der $F_{m, n-m}$ -Verteilung darstellt. Dies ist ein Test zum Niveau $\alpha \in (0, 1)$.

Spezialfall: Der Fall $\beta_0 = 0$ beschreibt einen *Test auf Zusammenhang*; das heißt, man testet, ob die Parameter β_1, \dots, β_m für die Beschreibung der Daten Y relevant sind.

Bemerkung 11.35 1. Wie kann man verstehen, daß die Testgröße T tatsächlich H_0 von H_1 unterscheiden soll? Führen wir die Bezeichnung

$$\tilde{Y} = Y - \underbrace{X\hat{\beta}}_{:=\hat{Y}}$$

ein; dabei gilt:

$$\hat{\sigma}^2 = \frac{1}{n-m} |\tilde{Y}|^2$$

und \tilde{Y} ist der Vektor der *Residuen*.

Ohne Beschränkung der Allgemeinheit setzen wir $\beta_0 = 0$. Falls H_0 nicht gelten soll, dann ist $\beta \neq 0$, und somit

$$|X\beta|^2 = (X\beta)^\top X\beta = \beta^\top X^\top X\beta > 0,$$

weil X den vollen Rang hat. Daraus folgt, daß H_0 abgelehnt werden soll, falls

$$|\hat{Y}|^2 = |X\hat{\beta}|^2 = \hat{\beta}^\top X^\top X\hat{\beta} \gg 0.$$

In der Testgröße $|X\hat{\beta}|^2$ sind allerdings die Schwankungen der Schätzung von β nicht berücksichtigt. Deswegen teilt man $|X\hat{\beta}|^2$ durch $\hat{\sigma}^2$:

$$T = \frac{\hat{\beta}^\top X^\top X\hat{\beta}}{m \cdot \hat{\sigma}^2} = \frac{|\hat{Y}|^2}{\frac{m}{n-m} |Y - \hat{Y}|^2}.$$

Der Satz von Pythagoras liefert

$$|Y|^2 = |\tilde{Y}|^2 + |\hat{Y}|^2,$$

wobei unter H_0

$$\mathbb{E} |\hat{Y}|^2 = \mathbb{E} |Y|^2 - \mathbb{E} |Y - \hat{Y}|^2 = n\sigma^2 - \mathbb{E} |\tilde{Y}|^2 \quad \text{gilt, und somit}$$

$$\frac{\mathbb{E} |\hat{Y}|^2}{\mathbb{E} \left(\frac{m}{n-m} |\tilde{Y}|^2 \right)} \stackrel{(H_0)}{=} \frac{n\sigma^2 - \mathbb{E} |\tilde{Y}|^2}{\frac{m}{n-m} \mathbb{E} |\tilde{Y}|^2} = \frac{n-m}{m} \left(\frac{n\sigma^2}{\mathbb{E} |\tilde{Y}|^2} - 1 \right),$$

$$\text{weil } \mathbb{E} |Y|^2 = \mathbb{E} (Y^\top Y) = \sigma^2 \cdot n, \quad \text{wegen } Y \sim N(0, \sigma^2).$$

\implies Die Testgröße T ist sensibel gegenüber Abweichungen von H_0 .

2. Die Größe

$$|\tilde{Y}|^2 = |Y - \hat{Y}|^2$$

wird *Reststreuung* genannt. Mit deren Hilfe kann der Begriff des *Bestimmtheitsmaßes* R^2 aus der Stochastik I wie folgt verallgemeinert werden:

$$R^2 = 1 - \frac{|\tilde{Y}|^2}{|Y - \bar{Y}_n \cdot e|^2},$$

wobei $e = (1, \dots, 1)^\top$, $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$.

Satz 11.36 Unter $H_0 : \beta = \beta_0$ gilt

$$T = \frac{(\hat{\beta} - \beta_0)^\top X^\top X (\hat{\beta} - \beta_0)}{m\hat{\sigma}^2} \sim F_{m,n-m}.$$

Beweis Es gilt

$$\begin{aligned}\hat{\beta} &\sim N\left(\beta_0, \sigma^2 (X^\top X)^{-1}\right) \\ \implies \hat{\beta} - \beta_0 &\sim N\left(0, \underbrace{\sigma^2 (X^\top X)^{-1}}_{:= K}\right).\end{aligned}$$

Falls $A = \frac{X^\top X}{\sigma^2}$, dann ist $AK = \text{idempotent}$. Dann gilt nach Satz 11.24

$$(\hat{\beta} - \beta_0)^\top A (\hat{\beta} - \beta_0) \stackrel{H_0}{\sim} \chi_m^2$$

(Zur Information: Unter H_1 wäre $(\hat{\beta} - \beta_0)^\top A (\hat{\beta} - \beta_0)$ nicht-zentral χ^2 -verteilt).

Es gilt zusätzlich:

$$\frac{n-m}{\sigma^2} \hat{\sigma}^2 \sim \chi_{n-m}^2.$$

Aus Satz 11.34 folgt die Unabhängigkeit von $(\hat{\beta} - \beta_0)^\top A (\hat{\beta} - \beta_0)$ und $\frac{n-m}{\sigma^2} \hat{\sigma}^2$.

$$\implies T = \frac{(\hat{\beta} - \beta_0)^\top (X^\top X) (\hat{\beta} - \beta_0)/m}{(n-m)\hat{\sigma}^2/(n-m)} \sim F_{m,n-m}$$

nach der Definition der F -Verteilung. □

Jetzt wird die Relevanz der einzelnen Parameter β_j getestet:

$$H_0 : \beta_j = \beta_{0j} \text{ vs. } H_1 : \beta_j \neq \beta_{0j}.$$

Satz 11.37 Unter $H_0 : \beta_j = \beta_{0j}$ gilt:

$$\begin{aligned}T_j &= \frac{\hat{\beta}_j - \beta_{0j}}{\hat{\sigma} \sqrt{x^{jj}}} \sim t_{n-m}, \text{ wobei} \\ (X^\top X)^{-1} &= (x^{ij})_{i,j=1,\dots,m}.\end{aligned}$$

Beweis Aus $\hat{\beta} \stackrel{H_0}{\sim} N(\beta_0, \sigma^2 (X^\top X)^{-1})$ folgt $\hat{\beta}_j \stackrel{H_0}{\sim} N(\beta_{0j}, \sigma^2 x^{jj})$ und somit $\hat{\beta}_j - \beta_{0j} \sim N(0, \sigma^2 x^{jj})$. Dann ist $\frac{\hat{\beta}_j - \beta_{0j}}{\sigma \sqrt{x^{jj}}} \sim N(0, 1)$. Zusätzlich gilt: $\frac{(n-m)\hat{\sigma}^2}{\sigma^2} \stackrel{H_0}{\sim} \chi_{n-m}^2$, und nach Satz 11.34 sind beide Größen unabhängig. Daraus folgt:

$$T_j = \frac{\hat{\beta}_j - \beta_{0j}}{\sqrt{\frac{(n-m)\hat{\sigma}^2}{(n-m)\sigma^2}}} \sim t_{n-m}.$$

□

Somit wird $H_0 : \beta_j = \beta_{j0}$ abgelehnt, falls $|T| > t_{n-m, 1-\alpha/2}$. Dies ist ein Test von H_0 vs. H_1 zum Niveau α .

Sei nun

$$H_0 : \beta_{j1} = \beta_{0j1}, \dots, \beta_{jl} = \beta_{0jl} \text{ vs. } H_1 : \exists i \in \{1, \dots, l\} : \beta_{ji} \neq \beta_{0ji}$$

die zu testende Hypothese.

Übungsaufgabe 11.38 Zeigen Sie, daß unter H_0 folgende Verteilungsaussage gilt:

$$T = \frac{(\hat{\beta}' - \beta'_0)^\top K'(\hat{\beta}' - \beta'_0)}{l\hat{\sigma}^2} \sim F_{l, n-m},$$

wobei

$$\begin{aligned}\hat{\beta}' &= (\hat{\beta}_{j1}, \dots, \hat{\beta}_{jl}), \\ \beta'_0 &= (\beta_{0j1}, \dots, \beta_{0jl}), \\ K' &= \begin{pmatrix} x^{j_1 j_1} & \dots & x^{j_1 j_l} \\ \vdots & \vdots & \vdots \\ x^{j_l j_1} & \dots & x^{j_l j_l} \end{pmatrix}^{-1}.\end{aligned}$$

Konstruieren Sie den dazugehörigen F -Test!

Test auf Linearkombination von Parametern

Sei nun

$$H_0 : H\beta = c \text{ vs. } H_1 : H\beta \neq c,$$

wobei H eine $(r \times m)$ -Matrix und $c \in \mathbb{R}^r$ sind, $r \leq m$.

Satz 11.39 Unter H_0 gilt

$$T = \frac{(H\hat{\beta} - c)^\top (H(X^\top X)^{-1}H^\top)^{-1}(H\hat{\beta} - c)}{r\hat{\sigma}^2} \sim F_{r, n-m}.$$

Deshalb wird $H_0 : H\beta = c$ abgelehnt, falls $T > F_{r, n-m, 1-\alpha}$.

Übungsaufgabe 11.40 Beweisen Sie Satz 11.39!

11.2.5 Konfidenzbereiche

1. Konfidenzintervall für β_j

Im Satz 11.37 haben wir gezeigt, daß

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \cdot \sqrt{x^{jj}}} \sim t_{n-m},$$

wobei $(X^\top X)^{-1} = (x^{ij})_{i,j=1,\dots,m}$. Daraus kann mit den üblichen Überlegungen folgendes Konfidenzintervall für β_j zum Niveau $1 - \alpha$ abgeleitet werden:

$$P(\hat{\beta}_j - t_{n-m,1-\alpha/2} \cdot \hat{\sigma} \sqrt{x^{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{n-m,1-\alpha/2} \cdot \hat{\sigma} \sqrt{x^{jj}}) = 1 - \alpha.$$

2. Simultaner Konfidenzbereich für $\beta = (\beta_1, \dots, \beta_m)^\top$

Falls A_j wie unten definiert ist, dann erhält man mit Hilfe folgender Bonferroni-Ungleichung

$$P\left(\bigcap_{j=1}^m A_j\right) \geq \sum_{j=1}^m P(A_j) - (m-1),$$

daß

$$\begin{aligned} & P\left(\underbrace{\hat{\beta}_j - t_{n-m,1-\alpha/(2m)} \cdot \hat{\sigma} \sqrt{x^{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{n-m,1-\alpha/(2m)} \cdot \hat{\sigma} \sqrt{x^{jj}}}_{:=A_j}, \quad j = 1, \dots, m\right) \\ & \stackrel{\text{(Bonferroni)}}{\geq} \sum_{j=1}^m P(A_j) - (m-1) = m \cdot \left(1 - \frac{\alpha}{m}\right) - m + 1 = 1 - \alpha. \end{aligned}$$

Daraus folgt, daß

$$\{\beta = (\beta_1, \dots, \beta_m)^\top : \beta_j \in [\hat{\beta}_j - t_{n-m,1-\alpha/(2m)} \cdot \hat{\sigma} \sqrt{x^{jj}}, \hat{\beta}_j + t_{n-m,1-\alpha/(2m)} \cdot \hat{\sigma} \sqrt{x^{jj}}]\}$$

ein simultaner Konfidenzbereich für β zum Niveau $1 - \alpha$ ist.

3. Konfidenzellipsoid für β .

In Satz 11.36 haben wir bewiesen, daß

$$T = \frac{(\hat{\beta} - \beta)^\top (X^\top X)(\hat{\beta} - \beta)}{m\hat{\sigma}^2} \sim F_{m,n-m}.$$

Daraus folgt, daß

$$\begin{aligned} & P(T \leq F_{m,n-m,1-\alpha}) = 1 - \alpha \quad \text{und} \\ & \mathcal{E} = \left\{ \beta \in \mathbb{R}^m : \frac{(\hat{\beta} - \beta)^\top (X^\top X)(\hat{\beta} - \beta)}{m\hat{\sigma}^2} \leq F_{m,n-m,1-\alpha} \right\} \end{aligned}$$

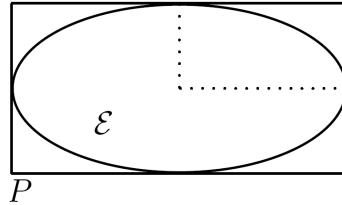
ein Konfidenzellipsoid zum Niveau $1 - \alpha$ ist, siehe Abbildung 11.2.

Da ein Ellipsoid in das minimale Parallelepiped P eingebettet werden kann, sodaß die Seitenlängen von P gleich $2 \times$ der Halbachsenlängen von \mathcal{E} sind, ergibt sich folgender simultaner Konfidenzbereich für $\beta = (\beta_1, \dots, \beta_m)^\top$:

$$P = \left\{ \beta : \hat{\beta}_j - \hat{\sigma} \sqrt{mx^{jj} F_{m,n-m,1-\alpha}} \leq \beta_j \leq \hat{\beta}_j + \hat{\sigma} \sqrt{mx^{jj} F_{m,n-m,1-\alpha}} \right\}$$

$$j = 1, \dots, m.$$

Abbildung 11.2: Konfidenzellipsoid



4. *Konfidenzintervall für den erwarteten Zielwert $x_{01}\beta_1 + \dots + x_{0m}\beta_m$.*

Sei $Y_0 = x_{01}\beta_1 + \dots + x_{0m}\beta_m + \varepsilon_0$ eine neue Zielvariable mit $\mathbb{E} \varepsilon_0 = 0$. Dann ist

$$\mathbb{E} Y_0 = \sum_{i=1}^n x_{0i}\beta_i.$$

Wir konstruieren ein Konfidenzintervall für $\mathbb{E} Y_0$. Dazu verwenden wir die Beweisidee des Satzes 11.37 kombiniert mit Satz 11.39 mit $H = (x_{01}, \dots, x_{0m}) = x_0^\top$, $r = 1$. Dann ist

$$T = \frac{\sum_{i=1}^m \hat{\beta}_i x_{0i} - \sum_{i=1}^m \beta_i x_{0i}}{\hat{\sigma} \sqrt{x_0^\top (X^\top X)^{-1} x_0}} \sim t_{n-m}.$$

Darum ist

$$\begin{aligned} \left\{ \beta = (\beta_1, \dots, \beta_m)^\top : \sum_{i=1}^m x_{0i} \hat{\beta}_i - \hat{\sigma} \sqrt{x_0^\top (X^\top X)^{-1} x_0} \cdot t_{n-m, 1-\alpha/2} \right. \\ \left. \leq \sum_{i=1}^m x_{0i} \beta_i \leq \sum_{i=1}^m x_{0i} \hat{\beta}_i + \hat{\sigma} \sqrt{x_0^\top (X^\top X)^{-1} x_0} \cdot t_{n-m, 1-\alpha/2} \right\} \end{aligned}$$

ein Konfidenzintervall für $\sum_{i=1}^m x_{0i} \beta_i$ zum Niveau $1 - \alpha$.

5. *Prognoseintervall für die Zielvariable Y_0 .*

Für $Y_0 = \sum_{i=1}^m x_{0i} \beta_i + \varepsilon_0$ mit $\varepsilon_0 \sim N(0, \sigma^2)$, ε_0 unabhängig von $\varepsilon_1, \dots, \varepsilon_n$,

gilt:

$$\begin{aligned} x_0^\top \hat{\beta} - Y_0 &\sim N(0, \sigma^2(1 + x_0^\top (X^\top X)^{-1} x_0)) \\ \implies \frac{x_0^\top \hat{\beta} - Y_0}{\sigma \sqrt{1 + x_0^\top (X^\top X)^{-1} x_0}} &\sim N(0, 1) \\ \implies \frac{x_0^\top \hat{\beta} - Y_0}{\hat{\sigma} \sqrt{1 + x_0^\top (X^\top X)^{-1} x_0}} &\sim t_{n-m} \end{aligned}$$

Also ist

$$(x_0^\top \hat{\beta} + c, x_0^\top \hat{\beta} - c)$$

$$\text{mit } c = \hat{\sigma} \sqrt{1 + x_0^\top (X^\top X)^{-1} \cdot x_0} \cdot t_{n-m, 1-\alpha/2}$$

ein Prognoseintervall für die Zielvariable Y_0 zum Niveau $1 - \alpha$.

6. Konfidenzband für die Regressionsebene $y = \beta_1 + \sum_{i=2}^m x_i \beta_i$ im multiplen Regressionsmodell.

Es sei $Y = X\beta + \varepsilon$, wobei

$$X = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1m} \\ 1 & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{nm} \end{pmatrix} \quad \text{und } \varepsilon \sim N(0, \sigma^2 \cdot).$$

Wir wollen ein zufälliges Konfidenzband $B(x)$ für y angeben. Es gilt

$$\begin{aligned} P \left(y = \beta_1 + \sum_{i=2}^m \beta_i x_i \in B(x) \right) &= 1 - \alpha \quad \forall x \in \mathbb{R}_1^{m-1}, \quad \text{wobei} \\ R_1^{m-1} &= \left\{ (1, x_2, \dots, x_m)^\top \in \mathbb{R}^m \right\}. \end{aligned}$$

Satz 11.41 Es gilt:

$$P \left(\max_{x \in \mathbb{R}_1^{m-1}} \frac{\overbrace{(x^\top \hat{\beta} - (\beta_1 + \sum_{i=2}^m \beta_i x_i))^2}^{=y}}{\hat{\sigma}^2 x^\top (X^\top X)^{-1} x} \leq m \cdot F_{m, n-m, 1-\alpha} \right) = 1 - \alpha.$$

ohne Beweis.

11.3 Multivariate lineare Regression mit $\text{Rang}(X) < m$

Es sei $Y = X\beta + \varepsilon$, $Y \in \mathbb{R}^n$, wobei X eine $(n \times m)$ -Matrix mit $\text{Rang}(X) = r < m$ ist, $\beta = (\beta_1, \dots, \beta_m)^\top$, $\varepsilon \in \mathbb{R}^n$, $\mathbb{E}\varepsilon = 0$, $\mathbb{E}(\varepsilon_i\varepsilon_j) = \delta_{ij}\sigma^2$, $i, j = 1, \dots, n$, $\sigma^2 > 0$.

Der MKQ-Schätzer $\hat{\beta}$ ist nach wie vor eine Lösung der Normalengleichung

$$(X^\top X)\beta = X^\top Y.$$

$X^\top X$ ist aber nicht mehr invertierbar, weil

$$\text{Rang}(X^\top X) \leq \min \left\{ \text{Rang}(X), \text{Rang}(X^\top) \right\} = r < m.$$

Um $\hat{\beta}$ aus der Normalengleichung zu gewinnen, sollen beide Seiten der Gleichung mit der sogenannten *verallgemeinerten Inversen* von $X^\top X$ multipliziert werden.

11.3.1 Verallgemeinerte Inverse

Definition 11.42 Sei A eine $(n \times m)$ -Matrix. Eine $(m \times n)$ -Matrix A^- heißt *verallgemeinerte Inverse* von A , falls

$$AA^-A = A \quad \text{gilt.}$$

Die Matrix A^- ist nicht eindeutig bestimmt, was die folgenden Hilfssätze zeigen.

Lemma 11.43 Sei A eine $(n \times m)$ -Matrix, $m \leq n$ mit $\text{Rang}(A) = r \leq m$. Es existieren invertierbare Matrizen P ($n \times n$) und Q ($m \times m$), sodaß

$$PAQ = \begin{pmatrix} r & 0 \\ 0 & 0 \end{pmatrix}, \quad \text{wobei } I_r = \text{diag}(\underbrace{1, \dots, 1}_{r \text{ Mal}}). \quad (11.6)$$

Folgerung 11.44 Für eine beliebige $(n \times m)$ -Matrix A mit $n \geq m$, $\text{Rang}(A) = r \leq m$ gilt

$$A^- = Q \begin{pmatrix} r & A_2 \\ A_1 & A_3 \end{pmatrix} P, \quad (11.7)$$

wobei P und Q Matrizen aus der Darstellung (11.6) sind, $r = \text{diag}(\overbrace{1, \dots, 1}^{r \text{ Mal}})$, und A_1, A_2, A_3 beliebige $((m-r) \times r)$, $(r \times (n-r))$ bzw. $((m-r) \times (n-r))$ -Matrizen sind.

Insbesondere kann

$$\begin{aligned} A_1 &= 0, \\ A_2 &= 0, \\ A_3 &= \text{diag}(\underbrace{1, \dots, 1}_{s-r \text{ Mal}}, 0, \dots, 0), \\ s &\in \{r, \dots, m\} \end{aligned}$$

gewählt werden, das heißt, $\text{Rang}(A^-) = s \in \{r, \dots, m\}$ für

$$A^- = Q \begin{pmatrix} s & 0 \\ 0 & 0 \end{pmatrix} P.$$

Beweis Zeigen wir, daß für A^- wie in (11.7) gegeben, $AA^-A = A$ gilt. Aus Lemma 11.43 folgt, daß

$$\begin{aligned} A &= P^{-1} \cdot \text{diag}(1, \dots, 1, 0, \dots, 0) \cdot Q^{-1} \quad \text{und somit} \\ AA^-A &= P^{-1} \begin{pmatrix} r & 0 \\ 0 & 0 \end{pmatrix} Q^{-1}Q \cdot \begin{pmatrix} r & A_2 \\ A_1 & A_3 \end{pmatrix} PP^{-1} \begin{pmatrix} r & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} \\ &= P^{-1} \begin{pmatrix} r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} r & A_2 \\ A_1 & A_3 \end{pmatrix} \begin{pmatrix} r & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} = P^{-1} \begin{pmatrix} r & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} \\ &= A. \end{aligned}$$

□

Lemma 11.45 Sei A eine beliebige $(n \times m)$ -Matrix mit $\text{Rang}(A) = r \leq m$, $m \leq n$.

1. Falls $(A^\top A)^-$ eine verallgemeinerte Inverse von $A^\top A$ ist, dann ist $((A^\top A)^-)^\top$ ebenfalls eine verallgemeinerte Inverse von $A^\top A$.

2. Es gilt die Darstellung

$$\begin{aligned} (A^\top A)(A^\top A)^-A^\top &= A^\top \quad \text{bzw.} \\ A(A^\top A)^-(A^\top A) &= A. \end{aligned}$$

Beweis 1. $A^\top A$ ist symmetrisch, also

$$\underbrace{\left(A^\top A(A^\top A)^-A^\top A \right)^\top}_{=A^\top A((A^\top A)^-)^\top A^\top A} = \left(A^\top A \right)^\top = A^\top A.$$

Also ist $((A^\top A)^-)^\top$ eine verallgemeinerte Inverse von $A^\top A$.

2. Es sei $B = (A^\top A)(A^\top A)^{-1}A^\top - A^\top$. Wir zeigen, daß $B = 0$, indem wir zeigen, daß $BB^\top = 0$.

$$\begin{aligned} BB^\top &= \left((A^\top A)(A^\top A)^{-1}A^\top - A^\top \right) \left(A \left((A^\top A)^{-1} \right)^\top A^\top A - A \right) \\ &= A^\top A (A^\top A)^{-1} A^\top A \left((A^\top A)^{-1} \right)^\top A^\top A - \underbrace{A^\top A (A^\top A)^{-1} A^\top A}_{=A^\top A} \\ &\quad - \underbrace{A^\top A \left((A^\top A)^{-1} \right)^\top \cdot A^\top A}_{=A^\top A} + A^\top A = A^\top A - 2A^\top A + A^\top A = 0. \end{aligned}$$

Die Aussage $A(A^\top A)^{-1}A^\top A = A$ erhält man, indem man die Matrizen an beiden Seiten der Gleichung $A^\top A(A^\top A)^{-1}A^\top = A^\top$ transponiert.

□

11.3.2 MKQ-Schätzer für β

Satz 11.46 Es sei X eine $(n \times m)$ -Designmatrix mit $\text{Rang}(X) = r \leq m$ in der linearen Regression $Y = X\beta + \varepsilon$. Die allgemeine Lösung der Normalengleichung

$$(X^\top X)\beta = X^\top Y$$

sieht folgendermaßen aus:

$$\beta = (X^\top X)^{-1} X^\top Y + \left(I_m - (X^\top X)^{-1} X^\top X \right) z, \quad z \in \mathbb{R}^m. \quad (11.8)$$

Beweis 1. Zeigen wir, daß β wie in (11.8) angegeben, eine Lösung der Normalengleichung darstellt.

$$\begin{aligned} X^\top X\beta &= \underbrace{(X^\top X)(X^\top X)^{-1} X^\top Y}_{=X^\top \text{ (Lemma 11.45, 2.)}} + \left(X^\top X - \underbrace{X^\top X(X^\top X)^{-1} X^\top X}_{=X^\top X} \right) z \\ &= X^\top Y \end{aligned}$$

2. Zeigen wir, daß eine beliebige Lösung β' der Normalengleichung die Form (11.8) besitzt. Sei β die Lösung (11.8). Wir bilden die Differenz der Gleichungen

$$\begin{array}{rcl} (X^\top X)\beta' &= X^\top Y \\ - (X^\top X)\beta &= X^\top Y \\ \hline (X^\top X)(\beta' - \beta) &= 0 \end{array}$$

$$\begin{aligned}
\beta' &= (\beta' - \beta) + \beta \\
&= \beta' - \beta + (X^\top X)^{-} X^\top Y + \left(m - (X^\top X)^{-} X^\top X \right) z \\
&= (X^\top X)^{-} X^\top Y + \left(m - (X^\top X)^{-} X^\top X \right) z + (\beta' - \beta) - \underbrace{(X^\top X)^{-} X^\top X (\beta' - \beta)}_{=0} \\
&= (X^\top X)^{-} X^\top Y + \left(m - (X^\top X)^{-} X^\top X \right) \left(\underbrace{z + \beta' - \beta}_{=z_0} \right) \\
&\implies \beta' \text{ besitzt die Darstellung (11.8).}
\end{aligned}$$

□

Bemerkung 11.47 Der Satz 11.46 liefert die Menge aller Extrempunkte der MKQ-Minimierungsaufgabe

$$e(\beta) = \frac{1}{n} |Y - X\beta|^2 \longrightarrow \min_{\beta} .$$

Deshalb soll die Menge aller MKQ-Schätzer von β in (11.8) zusätzliche Anforderungen erfüllen.

Satz 11.48 1. Alle MKQ-Schätzer von β haben die Form

$$\bar{\beta} = (X^\top X)^{-} X^\top Y, \quad \text{wobei}$$

$(X^\top X)^{-}$ eine beliebige verallgemeinerte Inverse von $X^\top X$ ist.

2. $\bar{\beta}$ ist nicht erwartungstreu, denn

$$\mathbb{E} \bar{\beta} = (X^\top X)^{-} X^\top X\beta.$$

3. Es gilt:

$$\text{Cov } \bar{\beta} = \sigma^2 (X^\top X)^{-} (X^\top X) ((X^\top X)^{-})^\top .$$

Beweis 1. Zeigen wir, daß $e(\beta) \geq e(\bar{\beta}) \quad \forall \beta \in \mathbb{R}^m$.

$$\begin{aligned}
n \cdot e(\beta) &= |Y - X\beta|^2 = (Y - X\bar{\beta} + X(\bar{\beta} - \beta))^\top (Y - X\bar{\beta} + X(\bar{\beta} - \beta)) \\
&= (Y - X\bar{\beta})^\top (Y - X\bar{\beta}) + (X(\bar{\beta} - \beta))^\top (X(\bar{\beta} - \beta)) \\
&\quad + 2(\bar{\beta} - \beta)^\top X^\top (Y - X\bar{\beta}) \\
&= n \cdot e(\bar{\beta}) + \underbrace{2 \cdot (\bar{\beta} - \beta)^\top (X^\top Y - (X^\top X\bar{\beta}))}_{=0} + |X(\bar{\beta} - \beta)|^2 \\
&\geq n \cdot e(\bar{\beta}) + 0 = n \cdot e(\bar{\beta}), \quad \text{denn}
\end{aligned}$$

$\bar{\beta}$ hat die Form (11.8) mit $z = 0$ und ist somit eine Lösung der Normalengleichung.

2. Es gilt:

$$\begin{aligned}\mathbb{E} \bar{\beta} &= \mathbb{E} \left((X^\top X)^{-1} X^\top Y \right) = \left(X^\top X \right)^{-1} X^\top \mathbb{E} Y \\ &= (X^\top X)^{-1} X^\top X \beta, \quad \text{weil aus} \\ Y &= X\beta + \varepsilon, \quad \mathbb{E} \varepsilon = 0 \quad \text{die Relation } \mathbb{E} Y = X\beta \text{ folgt.}\end{aligned}$$

Warum ist $\bar{\beta}$ nicht erwartungstreue? Also warum ist $(X^\top X)^{-1} X^\top X \beta \neq \beta$, $\beta \in \mathbb{R}^m$?

Da $\text{Rang}(X) = r < m$, ist $\text{Rang}(X^\top X) < m$ und damit $\text{Rang}((X^\top X)^{-1} X^\top X) < m$. Darum existiert ein $\beta \neq 0$, für das gilt:

$$(X^\top X)^{-1} X^\top X \beta = 0 \neq \beta,$$

also ist $\bar{\beta}$ nicht erwartungstreue. Es gilt sogar, daß alle Lösungen von (11.8) keine erwartungstreuen Schätzer sind. Wenn wir den Erwartungswert an (11.8) anwenden, so erhielten wir im Falle der Erwartungstreue:

$$\begin{aligned}\forall \beta \in \mathbb{R}^m : \quad \beta &= (X^\top X)^{-1} X^\top X \beta + \left(I_m - (X^\top X)^{-1} (X^\top X) \right) z, \quad z \in \mathbb{R}^m. \\ &\implies \left(I_m - (X^\top X)^{-1} (X^\top X) \right) (z - \beta) = 0 \quad \forall z, \beta \in \mathbb{R}^m \\ &\implies (X^\top X)^{-1} (X^\top X) (\beta - z) = \beta - z, \quad \forall z, \beta \in \mathbb{R}^m.\end{aligned}$$

Da diese Gleichung nicht für alle $\beta \in \mathbb{R}^m$ gelten kann (siehe oben), führt die Annahme der Erwartungstreue zum Widerspruch.

3. Es gilt:

$$\begin{aligned}\text{Cov} \left(\bar{\beta}_i, \bar{\beta}_j \right) &= \text{Cov} \left(\left(\underbrace{(X^\top X)^{-1} X^\top Y}_{:= A = (a_{kl})} \right)_i, \left((X^\top X)^{-1} X^\top Y \right)_j \right) \\ &= \text{Cov} \left(\sum_{k=1}^n a_{ik} Y_k, \sum_{l=1}^n a_{jl} Y_l \right) \\ &= \sum_{k,l=1}^n a_{ik} a_{jl} \underbrace{\text{Cov} \left(Y_k, Y_l \right)}_{=\sigma^2 \cdot \delta_{kl}} = \sigma^2 \sum_{k=1}^n a_{ik} a_{jk} = \left(\sigma^2 A A^\top \right)_{i,j} \\ &= \left(\sigma^2 (X^\top X)^{-1} X^\top X \left((X^\top X)^{-1} \right)^\top \right)_{i,j}.\end{aligned}$$

□

11.3.3 Erwartungstreue schätzbare Funktionen

Definition 11.49 Eine Linearkombination $a^\top \beta$ von β_1, \dots, β_m , $a \in \mathbb{R}^m$ heißt (erwartungstreue) schätzbar, falls

$$\exists c \in \mathbb{R}^n : \quad \mathbb{E} \left(c^\top Y \right) = a^\top \beta,$$

das heißt, falls es einen linearen, erwartungstreuen Schätzer $c^\top Y$ für $a^\top \beta$ gibt.

Satz 11.50 Die Funktion $a^\top \beta$, $a \in \mathbb{R}^m$ ist genau dann erwartungstreu schätzbar, wenn eine der folgenden Bedingungen erfüllt ist:

1. $\exists c \in \mathbb{R}^n : a^\top = c^\top X$.
2. a erfüllt die Gleichung

$$a^\top (X^\top X)^- X^\top X = a^\top. \quad (11.9)$$

Beweis 1. „ \Rightarrow “: Falls $a^\top \beta$ schätzbar, dann existiert ein $d \in \mathbb{R}^n$ mit $\mathbb{E}(d^\top Y) = a^\top \beta \quad \forall \beta \in \mathbb{R}^m$. Also

$$\begin{aligned} a^\top \beta &= d^\top \mathbb{E} Y = d^\top X \beta \Rightarrow (a^\top - d^\top X) \beta = 0, \quad \forall \beta \in \mathbb{R}^m \\ &\Rightarrow a^\top = d^\top X, \end{aligned}$$

setze $c = d$, damit ist die erste Richtung bewiesen.

„ \Leftarrow “: $\mathbb{E}(c^\top Y) = c^\top \mathbb{E} Y = c^\top X \beta = a^\top \beta$, also ist $a^\top \beta$ erwartungstreu schätzbar.

2. „ \Rightarrow “: Falls $a^\top \beta$ erwartungstreu schätzbar ist, dann gilt:

$$a^\top (X^\top X)^- X^\top X \stackrel{\text{Punkt 1}}{=} \underbrace{c^\top X \cdot (X^\top X)^- X^\top X}_{=X \text{ (Lemma 11.45)}} = c^\top X \stackrel{\text{Punkt 1}}{=} a^\top.$$

Also ist (11.9) erfüllt.

„ \Leftarrow “: Falls $a^\top (X^\top X)^- X^\top X = a^\top$, dann gilt mit $c = (a^\top (X^\top X)^- X^\top)^T$ nach Punkt 1, daß $a^\top \beta$ schätzbar ist.

□

Bemerkung 11.51 Im Falle der Regression mit $\text{Rang}(X) = m$ ist die Gleichung (11.9) immer erfüllt, denn $(X^\top X)^- = (X^\top X)^{-1}$ und damit ist $a^\top \beta$ schätzbar für alle $a \in \mathbb{R}^m$.

Satz 11.52 (Beispiele schätzbarer Funktionen) Falls $\text{Rang}(X) = r < m$, dann sind folgende Linearkombinationen von β schätzbar:

1. Die Koordinaten $\sum_{j=1}^m x_{ij} \beta_j$, $i = 1, \dots, n$ des Erwartungswertvektors $\mathbb{E} Y = X\beta$.
2. Beliebige Linearkombinationen schätzbarer Funktionen.

Beweis 1. Führe die Bezeichnung $\tilde{x}_i = (x_{i1}, \dots, x_{im})$, $i = 1, \dots, n$ ein.
Dann ist

$$\sum_{j=1}^m x_{ij}\beta_j = \tilde{x}_i^\top \beta \quad \forall i = 1, \dots, n,$$

$$X\beta = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)^\top \beta.$$

$\tilde{x}_i\beta$ ist schätzbar, falls \tilde{x}_i die Gleichung (11.9) erfüllt, die für alle $i = 1, \dots, n$ folgendermaßen in Matrixform dargestellt werden kann:

$$X(X^\top X)^{-1}X^\top X = X,$$

was nach Lemma 11.45 Gültigkeit besitzt.

2. Für $a_1, \dots, a_k \in \mathbb{R}^m$ seien $a_1^\top \beta, \dots, a_k^\top \beta$ schätzbare Funktionen. Für alle $\lambda = (\lambda_1, \dots, \lambda_k)^\top \in \mathbb{R}^k$ zeigen wir, daß $\sum_{i=1}^k \lambda_i \cdot a_i^\top \beta = \lambda^\top A\beta$ schätzbar ist, wobei $A = (a_1, \dots, a_k)^\top$. Zu zeigen bleibt: $b = (\lambda^\top A)^\top$ erfüllt (11.9), also

$$\lambda^\top A(X^\top X)^{-1}X^\top X = \lambda^\top A.$$

Diese Gleichung stimmt, weil $a_i^\top (X^\top X)^{-1}X^\top X = a_i^\top$, $i = 1, \dots, k$. Nach Satz 11.50, 2.) ist $\lambda^\top A\beta$ schätzbar.

□

Satz 11.53 (Gauß-Markov) Es sei $a^\top \beta$ eine schätzbare Funktion, $a \in \mathbb{R}^m$ im linearen Regressionsmodell $Y = X\beta + \varepsilon$ mit $\text{Rang}(X) \leq m$.

1. Der beste lineare erwartungstreue Schätzer (engl. BLUE - best linear unbiased estimator) von $a^\top \beta$ ist durch $a^\top \bar{\beta}$ gegeben, wobei

$$\bar{\beta} = (X^\top X)^{-1}X^\top Y$$

ein MKQ-Schätzer für β ist.

2. $\text{Var } (a^\top \bar{\beta}) = \sigma^2 a^\top (X^\top X)^{-1} a$.

Beweis Die Linearität von $a^\top \bar{\beta} = a^\top (X^\top X)^{-1}X^\top Y$ als Funktion von Y ist klar. Zeigen wir die Erwartungstreue:

$$\begin{aligned} \mathbb{E}(a^\top \bar{\beta}) &= a^\top \mathbb{E}\bar{\beta} = a^\top (X^\top X)^{-1}X^\top X\beta \\ &= \underbrace{c^\top}_{=X \text{ (Lemma 11.45)}} \underbrace{(X^\top X)^{-1}X^\top X}_{=a^\top} \beta = \underbrace{c^\top}_{=a^\top} X\beta = a^\top \beta \quad \forall \beta \in \mathbb{R}^m. \end{aligned}$$

Berechnen wir $\text{Var}(a^\top \bar{\beta})$ (also beweisen wir Punkt 2), und zeigen, daß sie minimal ist.

$$\begin{aligned}
\text{Var}(a^\top \bar{\beta}) &= \text{Var}\left(\sum_{i=1}^m a_i \bar{\beta}_i\right) = \sum_{i,j=1}^m a_i a_j \cdot \text{Cov}(\bar{\beta}_i, \bar{\beta}_j) \\
&= a^\top \text{Cov}(\bar{\beta}) a \stackrel{(\text{Satz 11.48})}{=} a^\top \sigma^2 ((X^\top X)^{-1} X^\top X (X^\top X)^{-1})^\top a \\
&= \sigma^2 \cdot a^\top \underbrace{((X^\top X)^{-1})^\top}_{=(X^\top X)^{-1}} X^\top X \underbrace{((X^\top X)^{-1})^\top}_{=(X^\top X)^{-1}} a \\
&\stackrel{\text{Lemma 11.45, 1.)}}{=} \sigma^2 a^\top (X^\top X)^{-1} X^\top X (X^\top X)^{-1} a \\
&\stackrel{\text{Satz 11.50, 1.)}}{=} \sigma^2 \cdot c^\top \underbrace{X \cdot ((X^\top X)^{-1} X^\top X)}_{=X} (X^\top X)^{-1} X^\top c \\
&= \sigma^2 \underbrace{c^\top X}_{=a^\top} (X^\top X)^{-1} \underbrace{X^\top c}_{=a} = \sigma^2 a^\top (X^\top X)^{-1} a.
\end{aligned}$$

Jetzt zeigen wir, daß für einen beliebigen linearen, erwartungstreuen Schätzer $b^\top Y$ von $a^\top \beta$ gilt: $\text{Var}(b^\top Y) \geq \text{Var}(a^\top \bar{\beta})$. Weil $b^\top Y$ erwartungstreue ist, gilt: $\mathbb{E}(b^\top Y) = a^\top \beta$. Nach Satz 11.50 gilt: $a^\top = b^\top X$. Betrachten wir die Varianz von

$$\begin{aligned}
0 \leq \text{Var}(b^\top Y - a^\top \bar{\beta}) &= \text{Var}(b^\top Y) - 2\text{Cov}(b^\top Y, a^\top \bar{\beta}) + \text{Var}(a^\top \bar{\beta}) \\
&= \text{Var}(b^\top Y) - 2\sigma^2 a^\top (X^\top X)^{-1} a + \sigma^2 a^\top (X^\top X)^{-1} a = \text{Var}(b^\top Y) - \text{Var}(a^\top \bar{\beta})
\end{aligned}$$

mit

$$\begin{aligned}
\text{Cov}(b^\top Y, a^\top \bar{\beta}) &= \text{Cov}(b^\top Y, a^\top (X^\top X)^{-1} X^\top Y) = \sigma^2 a^\top (X^\top X)^{-1} \underbrace{X^\top b}_{=a} \\
&= \sigma^2 a^\top (X^\top X)^{-1} a.
\end{aligned}$$

Damit ist $\text{Var}(b^\top Y) \geq \text{Var}(a^\top \bar{\beta})$ und $a^\top \bar{\beta}$ ist ein bester, linearer, erwartungstreuer Schätzer für $a^\top \beta$. \square

Bemerkung 11.54 1. Falls $\text{Rang}(X) = m$, dann ist $a^\top \hat{\beta}$ der beste lineare, erwartungstreue Schätzer für $a^\top \beta$, $a \in \mathbb{R}^m$.

2. Wie im folgenden Satz gezeigt wird, hängt der Schätzer $a^\top \bar{\beta} = a^\top (X^\top X)^{-1} X^\top Y$ nicht von der Wahl der verallgemeinerten Inversen ab.

Satz 11.55 Der beste lineare, erwartungstreue Schätzer $a^\top \bar{\beta}$ für $a^\top \beta$ ist eindeutig bestimmt.

Beweis

$$a^\top \bar{\beta} = a^\top (X^\top X)^{-1} X^\top Y \stackrel{\text{Satz 11.50, 1.)}}{=} c^\top X (X^\top X)^{-1} X^\top Y.$$

Wir zeigen, daß $X(X^\top X)^{-}X^\top$ nicht von der Wahl von $(X^\top X)^{-}$ abhängt. Zeigen wir, daß für beliebige verallgemeinerte Inverse A_1 und A_2 von $(X^\top X)$ gilt: $XA_1X^\top = XA_2X^\top$. Nach Lemma 11.45, 2.) gilt:

$$XA_1X^\top X = X = XA_2X^\top X.$$

Multiplizieren wir alle Teile der Gleichung mit A_1X^\top von rechts:

$$XA_1 \underbrace{X^\top X A_1 X^\top}_{=X^\top} = XA_1X^\top = XA_2 \underbrace{X^\top X A_1 X^\top}_{=X^\top}$$

Also ist $XA_1X^\top = XA_2X^\top$. □

11.3.4 Normalverteilte Störgrößen

Sei $Y = X\beta + \varepsilon$ ein lineares Regressionsmodell mit $\text{Rang}(X) = r < m$ und $\varepsilon \sim N(0, \sigma^2)$. Genauso wie in Abschnitt 11.2.3 können Maximum-Likelihood-Schätzer $\tilde{\beta}$ und $\tilde{\sigma}^2$ für β und σ^2 hergeleitet werden. Und genauso wie im Satz 11.32 kann gezeigt werden, daß

$$\begin{aligned}\tilde{\beta} &= \bar{\beta} = (X^\top X)^{-}X^\top Y \quad \text{und} \\ \tilde{\sigma}^2 &= \frac{1}{n} |Y - X\bar{\beta}|^2.\end{aligned}$$

Jetzt werden die Verteilungseigenschaften von $\bar{\beta}$ und $\tilde{\sigma}^2$ untersucht. Wir beginnen mit der Erwartungstreue von $\tilde{\sigma}^2$. Wir zeigen, daß $\tilde{\sigma}^2$ nicht erwartungstreue ist, dafür ist aber der korrigierte Schätzer

$$\bar{\sigma}^2 = \frac{1}{n-r} |Y - X\beta|^2 = \frac{n}{n-r} \tilde{\sigma}^2$$

erwartungstreue.

Satz 11.56 Der Schätzer $\bar{\sigma}^2$ ist erwartungstreue für σ^2 .

Der Beweis des Satzes 11.56 folgt dem Beweis des Satzes 11.31, in dem $\hat{\beta} = (X^\top X)^{-1}X^\top Y$ und $\hat{\sigma}^2 = \frac{1}{n-m} |Y - X\beta|^2$ im Fall $\text{Rang}(X) = m$ betrachtet wurden. Somit ist die Aussage des Satzes 11.31 ein Spezialfall des Satzes 11.56. Führen wir die Matrix $D = -X(X^\top X)^{-}X^\top$ ein.

Lemma 11.57 Für D gelten folgende Eigenschaften:

1. $D^\top = D$ (Symmetrie),
2. $D^2 = D$ (Idempotenz),
3. $DX = 0$,
4. $\text{Spur}(D) = n - r$.

Beweis 1. Es gilt:

$$\begin{aligned} D^\top &= \left(-X(X^\top X)^{-1}X^\top \right)^\top = -X \left((X^\top X)^{-1} \right)^\top X^\top \\ &= -X(X^\top X)^{-1}X^\top = D, \end{aligned}$$

weil $\left((X^\top X)^{-1} \right)^\top$ auch eine verallgemeinerte Inverse von $X^\top X$ ist (vergleiche Lemma 11.45, 1.)).

2. Es gilt:

$$\begin{aligned} D^2 &= \left(-X(X^\top X)^{-1}X^\top \right)^2 = -2X(X^\top X)^{-1}X^\top + \underbrace{X(X^\top X)^{-1}X^\top X}_{=X(\text{Lemma 11.45, 2.})} (X^\top X)^{-1}X^\top \\ &= -X(X^\top X)^{-1}X^\top = D. \end{aligned}$$

$$3. DX = X - \underbrace{X(X^\top X)^{-1}X^\top X}_{=X(\text{Lemma 11.45, 2.})} = X - X = 0.$$

4. Es gilt:

$$\text{Spur}(D) = \text{Spur}(I) - \text{Spur}\left(X(X^\top X)^{-1}X^\top\right) = n - \text{Spur}\left(X(X^\top X)^{-1}X^\top\right).$$

Verwenden wir die Eigenschaft der symmetrischen idempotenten Matrizen A aus der linearen Algebra, daß $\text{Spur}(A) = \text{Rang}(A)$. Da $X(X^\top X)^{-1}X^\top$ symmetrisch und idempotent ist, genügt es zu zeigen, daß $\text{Rang}(X(X^\top X)^{-1}X^\top) = r$. Nach Lemma 11.45 2.) gilt:

$$\begin{aligned} \text{Rang}(X) &= r = \text{Rang}(X(X^\top X)^{-1}X^\top) \\ &\leq \min \left\{ \text{Rang}(X(X^\top X)^{-1}X^\top), \underbrace{\text{Rang}(X)}_{=r} \right\} \\ &\leq \text{Rang}\left(X(X^\top X)^{-1}X^\top\right) \leq \text{Rang}(X) = r \\ &\implies \text{Rang}\left(X(X^\top X)^{-1}X^\top\right) = r \\ &\implies \text{Spur}\left(X(X^\top X)^{-1}X^\top\right) = r. \end{aligned}$$

□

Beweis des Satzes 11.56 Mit Hilfe des Lemmas 11.57 bekommt man

$$\begin{aligned}\bar{\sigma}^2 &= \frac{1}{n-r} |Y - X\bar{\beta}|^2 = \frac{1}{n-r} |Y - X(X^\top X)^{-1}X^\top Y|^2 = \frac{1}{n-r} |DY|^2 \\ &= \frac{1}{n-r} \left| \underbrace{DX}_{=0} \beta + D\varepsilon \right|^2 = \frac{1}{n-r} |D\varepsilon|^2 = \frac{1}{n-r} \varepsilon^\top \underbrace{D^\top D}_{=D^2=D} \varepsilon = \frac{1}{n-r} \varepsilon^\top D\varepsilon.\end{aligned}$$

Deshalb gilt:

$$\begin{aligned}\mathbb{E} \bar{\sigma}^2 &= \frac{1}{n-r} \mathbb{E} (\varepsilon^\top D\varepsilon) = \frac{1}{n-r} \mathbb{E} \text{Spur}(\varepsilon^\top D\varepsilon) = \frac{1}{n-r} \text{Spur}(D \cdot \mathbb{E} (\underbrace{\varepsilon\varepsilon^\top}_{\sigma^2})) \\ &= \frac{\sigma^2}{n-r} \cdot \text{Spur}(D) = \sigma^2 \text{ nach Lemma 11.57, 4.), weil } \mathbb{E} \varepsilon\varepsilon^\top = \sigma^2 \\ &\text{wegen } \varepsilon \sim N(0, \sigma^2).\end{aligned}$$

□

Satz 11.58 Es gelten folgende Verteilungseigenschaften:

1. $\bar{\beta} \sim N\left((X^\top X)^{-1}X^\top X\beta, \sigma^2(X^\top X)^{-1}(X^\top X)((X^\top X)^{-1})^\top\right)$,
2. $\frac{(n-r)\bar{\sigma}^2}{\sigma^2} \sim \chi_{n-r}^2$,
3. $\bar{\beta}$ und $\bar{\sigma}^2$ sind unabhängig.

Beweis 1. Es gilt:

$$\bar{\beta} = (X^\top X)^{-1}X^\top Y = (X^\top X)^{-1}X^\top(X\beta + \varepsilon) = \underbrace{(X^\top X)^{-1}X^\top X\beta}_{=\mu} + \underbrace{(X^\top X)^{-1}X^\top \varepsilon}_{=A}$$

und mit der Definition von $N(\cdot, \cdot)$ bekommt man

$$\begin{aligned}\bar{\beta} &\sim N\left(\mu, \sigma^2 AA^\top\right) = N\left((X^\top X)^{-1}X^\top X\beta, \sigma^2(X^\top X)^{-1}X^\top X((X^\top X)^{-1})^\top\right) \\ &\text{mit } AA^\top = (X^\top X)^{-1}X^\top X((X^\top X)^{-1})^\top\end{aligned}$$

2. Es gilt $\bar{\sigma}^2 = \frac{1}{n-r} \varepsilon^\top D\varepsilon$ aus dem Beweis des Satzes 11.56. Deshalb

$$\frac{(n-r)\bar{\sigma}^2}{\sigma^2} = \underbrace{\left(\frac{\varepsilon}{\sigma}\right)^\top}_{\sim N(0,)} D \left(\frac{\varepsilon}{\sigma}\right) \stackrel{\text{(Satz 11.24)}}{\sim} \chi_{n-r}^2.$$

3. Betrachten wir $A\varepsilon$ und $\varepsilon^\top D\varepsilon$. Es genügt zu zeigen, daß sie unabhängig sind, um die Unabhängigkeit von $\bar{\beta}$ und $\bar{\sigma}^2$ zu beweisen, weil $\bar{\beta} = \mu + A\varepsilon$, $\bar{\sigma}^2 = \frac{1}{n-r} \varepsilon^\top D\varepsilon$. Es gilt: $A \cdot \sigma^2 \cdot D = 0$. Nach Satz 11.25 sind dann $A\varepsilon$ und $\varepsilon^\top D\varepsilon$ unabhängig.

□

11.3.5 Hypothesentests

Betrachten wir die Hypothesen $H_0 : H\beta = d$ vs. $H_1 : H\beta \neq d$, wobei H eine $(s \times m)$ -Matrix ($s \leq m$) mit $\text{Rang}(H) = s$ ist, und $d \in \mathbb{R}^s$.

Im Satz 11.39 haben wir im Fall $\text{Rang}(X) = r = m$ folgende Testgröße dafür betrachtet:

$$T = \frac{(H\hat{\beta} - d)^\top (H(X^\top X)^{-1}H^\top)^{-1}(H\hat{\beta} - d)}{s\hat{\sigma}^2} \stackrel{(H_0)}{\sim} F_{s,n-m}.$$

Im allgemeinen Fall betrachten wir

$$T = \frac{(H\bar{\beta} - d)^\top (H(X^\top X)^{-1}H^\top)^{-1}(H\bar{\beta} - d)}{s\bar{\sigma}^2}. \quad (11.10)$$

Wir wollen zeigen, daß $T \stackrel{(H_0)}{\sim} F_{s,n-r}$. Dann wird H_0 verworfen, falls $T > F_{s,n-r,1-\alpha}$. Dies ist ein Test zum Niveau $\alpha \in (0, 1)$.

Definition 11.59 Die Hypothese $H_0 : H\beta = d$ heißt *testbar*, falls alle Koordinaten des Vektors $H\beta$ schätzbare Funktionen sind.

Satz 11.50 gibt Bedingungen an H an, unter denen $H_0 : H\beta = d$ testbar ist. Diese werden im folgendem Lemma formuliert:

Lemma 11.60 Die Hypothese $H_0 : H\beta = d$ ist testbar genau dann, wenn

1. $\exists (s \times n)$ -Matrix $C : H = CX$, oder
2. $H(X^\top X)^{-1}X^\top X = H$.

Wir zeigen, daß die Testgröße T in (11.10) wohldefiniert ist, das heißtt, die $(s \times s)$ -Matrix $H(X^\top X)^{-1}H^\top$ positiv definit und damit invertierbar ist. Aus Folgerung 11.44 haben wir $X^\top X = P^{-1} \begin{pmatrix} r & 0 \\ 0 & 0 \end{pmatrix} P^{-1}$ für eine $(m \times m)$ -Matrix P , die invertierbar und symmetrisch ist. Deshalb gilt

$$(X^\top X)^{-1} = P \cdot \begin{pmatrix} r & 0 \\ 0 & m-r \end{pmatrix} P^{-1} = P \cdot P^{-1} = I,$$

das heißtt, daß es eine eindeutige verallgemeinerte Inverse von $X^\top X$ mit dieser Darstellung gibt. Daraus folgt, daß die $(s \times s)$ -Matrix $HPPH^\top = (PH^\top)^\top \cdot PH^\top$ positiv definit ist, weil $\text{Rang}(PH^\top) = s$. Sei nun $(X^\top X)^{-1}$ eine beliebige verallgemeinerte Inverse von $X^\top X$. Dann ist mit Lemma 11.60

$$H(X^\top X)^{-1}H^\top = CX(X^\top X)^{-1}X^\top C^\top = CXPPX^\top C^\top = HPPH^\top,$$

denn $X(X^\top X)^{-1}X^\top$ ist invariant bezüglich der Wahl von $(X^\top X)^{-1}$, laut Beweis des Satzes 11.55. Also ist $H(X^\top X)^{-1}H^\top$ positiv definit für eine beliebige verallgemeinerte Inverse $(X^\top X)^{-1}$ und die Testgröße T somit wohldefiniert.

Satz 11.61 Falls $H_0 : H\beta = d$ testbar ist, dann gilt $T \xrightarrow{(H_0)} F_{s,n-r}$.

Beweis ähnlich, wie in Satz 11.39 gilt

$$H\bar{\beta} - d = H(X^\top X)^{-1}X^\top(X\beta + \varepsilon) - d = \underbrace{H(X^\top X)^{-1}X^\top X\beta - d}_{=\mu} + \underbrace{H(X^\top X)^{-1}X^\top\varepsilon}_{=B\varepsilon}.$$

Zeigen wir, daß $\mu \xrightarrow{(H_0)} 0$.

$$\mu \stackrel{\text{(Lemma 11.60)}}{=} C \cdot \underbrace{X(X^\top X)^{-1}X^\top X}_{=X \text{ (Lemma 11.45, 2.)}} \cdot \beta - d = CX\beta - d = H\beta - d \xrightarrow{(H_0)} 0.$$

Nach Satz 11.58 sind $(H\bar{\beta} - d)^\top (H(X^\top X)^{-1}H^\top)^{-1}(H\bar{\beta} - d)$ und $s \cdot \bar{\sigma}^2$ unabhängig, $\frac{(n-r)\bar{\sigma}^2}{\sigma^2} \sim \chi_{n-r}^2$. Also bleibt nur noch zu zeigen, daß

$$\left(\underbrace{H\bar{\beta} - d}_{=\varepsilon^\top B^\top} \right)^\top (H(X^\top X)^{-1}H^\top)^{-1} \left(\underbrace{H\bar{\beta} - d}_{=B\varepsilon} \right) \xrightarrow{(H_0)} \chi_s^2.$$

Es gilt

$$\begin{aligned} & \varepsilon^\top B^\top (H(X^\top X)^{-1}H^\top)^{-1} B\varepsilon \\ &= \varepsilon^\top \underbrace{X \left((X^\top X)^{-1} \right)^\top H^\top (H(X^\top X)^{-1}H^\top)^{-1} H(X^\top X)^{-1} X^\top \varepsilon}_A \end{aligned}$$

Man kann leicht zeigen, daß A symmetrisch, idempotent und $\text{Rang}(A) = s$ ist. Zeigen wir zum Beispiel die Idempotenz:

$$\begin{aligned} A^2 &= X \left((X^\top X)^{-1} \right)^\top H^\top (H(X^\top X)^{-1}H^\top)^{-1} \underbrace{H(X^\top X)^{-1}X^\top X}_{H \text{ (Lemma 11.60, 2.)}} \left((X^\top X)^{-1} \right)^T H^\top \\ &\quad \cdot (H(X^\top X)^{-1}H^\top)^{-1} H(X^\top X)^{-1} X^\top \\ &= X \left((X^\top X)^{-1} \right)^\top H^\top (H(X^\top X)^{-1}H^\top)^{-1} H(X^\top X)^{-1} X^\top = A, \end{aligned}$$

weil $\left((X^\top X)^{-1} \right)^\top$ auch eine verallgemeinerte Inverse von $X^\top X$ ist (nach Lemma 11.45). Somit hängt auch $H(X^\top X)^{-1}H^\top = CX(X^\top X)^{-1}X^\top C^\top$ nicht von der Wahl von $(X^\top X)^{-1}$ ab, vgl. den Beweis des Satzes 11.55. Nach Satz 11.24 ist $\frac{\varepsilon^\top}{\sigma} A \frac{\varepsilon}{\sigma} \sim \chi_s^2$, wegen $\varepsilon \sim N(0, \sigma^2)$ und somit $T \xrightarrow{H_0} F_{s,n-r}$. \square

11.3.6 Konfidenzbereiche

ähnlich wie in Abschnitt 11.2.5 werden wir Konfidenzbereiche für unterschiedliche Funktionen vom Parametervektor β angeben. Aus dem Satz 11.61 ergibt sich unmittelbar folgender Konfidenzbereich zum Niveau $1 - \alpha \in (0, 1)$:

Folgerung 11.62 Sei $Y = X\beta + \varepsilon$ ein multivariates Regressionsmodell mit $\text{Rang}(X) = r < m$, H eine $(s \times m)$ -Matrix mit $\text{Rang}(H) = s$, $s \in \{1, \dots, m\}$ und $H_0 : H\beta = d$ testbar $\forall d \in \mathbb{R}^s$. Dann ist

$$\left\{ d \in \mathbb{R}^s : \frac{(H\bar{\beta} - d)^\top (H(X^\top X)^{-1} H^\top)^{-1} (H\bar{\beta} - d)}{s \cdot \bar{\sigma}^2} \leq F_{s, n-r, 1-\alpha} \right\}$$

ein Konfidenzbereich für $H\beta$ zum Niveau $1 - \alpha$.

Folgerung 11.63 Sei $h^\top \beta$ eine schätzbare lineare Funktion von β , $h \in \mathbb{R}^m$. Dann ist

$$\left(h^\top \bar{\beta} - t_{n-r, 1-\alpha/2} \cdot \bar{\sigma} \sqrt{h^\top (X^\top X)^{-1} h}, h^\top \bar{\beta} + t_{n-r, 1-\alpha/2} \cdot \bar{\sigma} \sqrt{h^\top (X^\top X)^{-1} h} \right)$$

ein Konfidenzintervall für $h^\top \beta$ zum Niveau $1 - \alpha$.

Beweis Setzen wir $s = 1$ und $H = h^\top$. Aus Satz 11.61 folgt

$$\begin{aligned} T &= \frac{(h^\top \bar{\beta} - d)^\top (h^\top (X^\top X)^{-1} h)^{-1} (h^\top \bar{\beta} - d)}{\bar{\sigma}^2} = \frac{(h^\top \bar{\beta} - d)(h^\top \bar{\beta} - d)}{\bar{\sigma}^2 (h^\top (X^\top X)^{-1} h)} \\ &= \frac{(h^\top \bar{\beta} - d)^2}{\bar{\sigma}^2 (h^\top (X^\top X)^{-1} h)} \sim F_{1, n-r} \end{aligned}$$

unter der Voraussetzung $h^\top \beta = d$, weil $h^\top (X^\top X)^{-1} h$ eindimensional (eine Zahl) ist. Deshalb gilt

$$\sqrt{T} = \frac{h^\top \beta - h^\top \bar{\beta}}{\bar{\sigma} \sqrt{h^\top (X^\top X)^{-1} h}} \sim t_{n-r}$$

und somit

$$P(-t_{n-r, 1-\alpha/2} \leq \sqrt{T} \leq t_{n-r, 1-\alpha/2}) = 1 - \alpha.$$

Daraus folgt das obige Konfidenzintervall. \square

Man kann sogar eine stärkere Version von 11.63 beweisen, die für alle h aus einem linearen Unterraum gilt:

Satz 11.64 (Konfidenzband von Scheffé) Sei $H = (h_1, \dots, h_s)^\top$, $h_1, \dots, h_s \in \mathbb{R}^m$, $1 \leq s \leq m$ und $H\beta = d$ testbar $\forall d \in \mathbb{R}^s$. Sei $\text{Rang}(H) = s$ und $\mathcal{L} = \langle h_1, \dots, h_s \rangle$ der lineare Unterraum, der von den Vektoren h_1, \dots, h_s aufgespannt wird. Dann gilt:

$$P \left(\max_{h \in \mathcal{L}} \left\{ \frac{(h^\top \beta - h^\top \bar{\beta})^2}{\bar{\sigma}^2 h^\top (X^\top X)^{-1} h} \right\} \leq s F_{s, n-r, 1-\alpha} \right) = 1 - \alpha$$

Somit ist

$$\left[h^\top \bar{\beta} - \sqrt{s F_{s, n-r, 1-\alpha}} \cdot \bar{\sigma} \sqrt{h^\top (X^\top X)^{-1} h}, h^\top \bar{\beta} + \sqrt{s F_{s, n-r, 1-\alpha}} \cdot \bar{\sigma} \sqrt{h^\top (X^\top X)^{-1} h} \right]$$

ein (gleichmäßiges bzgl. $h \in \mathcal{L}$) Konfidenzintervall für $h^\top \beta$.

Beweis Aus dem Satz 11.61 folgt $\forall \alpha \in (0, 1)$:

$$P \left(\underbrace{\left(H\bar{\beta} - H\beta \right)^\top \left(H(X^\top X)^{-1} H^\top \right)^{-1} \left(H\bar{\beta} - H\beta \right)}_{T_1} \leq s \cdot \bar{\sigma}^2 F_{s, n-r, 1-\alpha} \right) = 1 - \alpha.$$

Falls wir zeigen können, daß

$$T_1 = \max_{x \in \mathbb{R}^s, x \neq 0} \left\{ \frac{(x^\top (H\bar{\beta} - H\beta))^2}{x^\top (H(X^\top X)^{-1} H^\top) x} \right\}, \quad (11.11)$$

dann ist der Satz bewiesen, denn

$$\begin{aligned} 1 - \alpha &= P \left(T_1 \leq \underbrace{s \bar{\sigma}^2 F_{s, n-r, 1-\alpha}}_t \right) = P \left(\max_{x \in \mathbb{R}^s, x \neq 0} \left\{ \frac{(x^\top (H\bar{\beta} - H\beta))^2}{x^\top (H(X^\top X)^{-1} H^\top) x} \right\} \leq t \right) \\ &= P \left(\max_{x \in \mathbb{R}^s, x \neq 0} \left\{ \frac{((H^\top x)^\top \bar{\beta} - (H^\top x)^\top \beta)^2}{(H^\top x)^\top (X^\top X)^{-1} (H^\top x)} \right\} \leq t \right) \quad \text{und weil } H^\top x = h \in \mathcal{L} \\ &= P \left(\max_{h \in \mathcal{L}} \left\{ \frac{(h^\top \bar{\beta} - h^\top \beta)^2}{h^\top (X^\top X)^{-1} h} \right\} \leq s \bar{\sigma}^2 F_{s, n-r, 1-\alpha} \right). \end{aligned}$$

Also, zeigen wir die Gültigkeit von (11.11). Es genügt zu zeigen, daß T_1 die obere Schranke von

$$\frac{(x^\top (H\bar{\beta} - H\beta))^2}{x^\top (H(X^\top X)^{-1} H^\top) x}$$

darstellt, die auch angenommen wird. Da $H(X^\top X)^{-1}H^\top$ positiv definit ist und invertierbar, existiert eine invertierbare $(s \times s)$ -Matrix B mit der Eigenschaft $BB^\top = H(X^\top X)^{-1}H^\top$. Dann gilt

$$\begin{aligned} \left(x^\top (H\bar{\beta} - H\beta) \right)^2 &= \left(\underbrace{x^\top B}_{(B^\top x)^\top} \cdot B^{-1}(H\bar{\beta} - H\beta) \right)^2 \\ &\leq |B^\top x|^2 \cdot |B^{-1}(H\bar{\beta} - H\beta)|^2 \quad (\text{wegen der Ungleichung von Cauchy-Schwarz}) \\ &= x^\top BB^\top x \left(H\bar{\beta} - H\beta \right)^\top \cdot \underbrace{(B^{-1})^\top B^{-1}}_{=(B^\top)^{-1}B^{-1}=(BB^\top)^{-1}} (H\bar{\beta} - H\beta) \\ &= x^\top H(X^\top X)^{-1}H^\top x \cdot \left(H\bar{\beta} - H\beta \right)^\top \left(H(X^\top X)^{-1}H^\top \right)^{-1} (H\bar{\beta} - H\beta). \end{aligned}$$

Somit gilt

$$\frac{\left(x^\top (H\bar{\beta} - H\beta) \right)^2}{x^\top (H(X^\top X)^{-1}H^\top) x} \leq \left(H\bar{\beta} - H\beta \right)^\top \left(H(X^\top X)^{-1}H^\top \right)^{-1} \left(H\bar{\beta} - H\beta \right) = T_1.$$

Man kann leicht prüfen, daß diese Schranke für $x = \left(H(X^\top X)^{-1}H^\top \right)^{-1} (H\bar{\beta} - H\beta)$ angenommen wird. \square

11.3.7 Einführung in die Varianzanalyse

In diesem Abschnitt geben wir ein Beispiel für die Verwendung linearer Modelle mit Design-Matrix, die keinen vollen Rang besitzt. Dabei handelt es sich um die Aussage der *Variabilität der Erwartungswerte* in der Stichprobe $Y = (Y_1, \dots, Y_n)^\top$, die auf englisch *analysis of variance*, kurz *ANOVA*, heißt. Später werden wir auch denselben Begriff *Varianzanalyse* dafür verwenden.

Betrachten wir zunächst die *einfaktorielle Varianzanalyse*, bei der man davon ausgeht, daß die Stichprobe (Y_1, \dots, Y_n) in k homogene Teilklassen $(Y_{ij}, j = 1, \dots, n_i), i = 1, \dots, k$ zerlegbar ist, mit den Eigenschaften:

1. $\mathbb{E}(Y_{ij}) = \mu_i = \mu + \alpha_i, \quad j = 1, \dots, n_i, \quad i = 1, \dots, k.$
2. $n_i > 1, \quad i = 1, \dots, k, \quad \sum_{i=1}^k n_i = n, \quad \sum_{i=1}^k n_i \alpha_i = 0.$

Dabei ist μ ein Faktor, der allen Klassen gemeinsam ist, und α_i verkörpert die *klassenspezifischen Differenzen* zwischen den Erwartungswerten μ_1, \dots, μ_k . Die Nummer $i = 1, \dots, k$ der Klassen wird als *Stufe eines Einflussfaktors* (zum Beispiel die Dosis eines Medikaments in einer klinischen Studie) und $\alpha_i, i = 1, \dots, k$ als *Effekt* der i -ten Stufe gedeutet. Die Nebenbedingung $\sum_{i=1}^k n_i \alpha_i = 0$ bewirkt, daß die Umrechnung $(\mu_1, \dots, \mu_k) \longleftrightarrow (\mu, \alpha_1, \dots, \alpha_k)$

eindeutig wird und daß $\mu = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} \mathbb{E} Y_{ij}$. Es wird vorausgesetzt, daß μ_i mit unkorrelierten Meßfehlern ε_{ij} gemessen werden kann, das heißt

$$Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \dots, k, j = 1, \dots, n_i \quad (11.12)$$

$$\mathbb{E} \varepsilon_{ij} = 0, \quad \text{Var } \varepsilon_{ij} = \sigma^2, \quad \varepsilon_{ij} \text{ unkorreliert}, \quad i = 1, \dots, k, j = 1, \dots, n_i. \quad (11.13)$$

Es soll die *klassische ANOVA-Hypothese* getestet werden, daß *keine* Variabilität in den Erwartungswerten μ_i auffindbar ist:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k,$$

was bedeutet, daß

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k.$$

Aus der Nebenbedingung

$$\sum_{i=1}^k n_i \alpha_i = 0.$$

folgt: $\alpha_i = 0$

Die Problemstellung (11.12) kann in der Form der multivariaten linearen Regression folgendermaßen umgeschrieben werden:

$$Y = X\beta + \varepsilon, \text{ wobei } Y = (Y_{11}, \dots, Y_{1n_1}, Y_{21}, \dots, Y_{2n_2}, \dots, Y_{k1}, \dots, Y_{kn_k})^\top,$$

$$\beta = (\mu, \alpha_1, \dots, \alpha_k)^\top,$$

$$\varepsilon = (\varepsilon_{11}, \dots, \varepsilon_{1n_1}, \dots, \varepsilon_{k1}, \dots, \varepsilon_{kn_k})^\top,$$

$$X = \begin{pmatrix} 1 & 1 & 0 & \dots & \dots & 0 \\ 1 & 1 & 0 & \dots & \dots & 0 \\ \vdots & & & & & \\ 1 & 1 & 0 & \dots & \dots & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & & \\ 1 & 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & & \\ 1 & 0 & \dots & \dots & 0 & 1 \\ \vdots & & & & & \\ 1 & 0 & \dots & \dots & 0 & 1 \end{pmatrix} \left. \right\} \begin{array}{l} n_1 \\ n_2 \\ \vdots \\ n_k \end{array}$$

Die $(n \times (k+1))$ -Matrix X hat den Rang $k < m = k+1$, somit ist die Theorie von Abschnitt 11.3 auf dieses Modell komplett anwendbar.

Übungsaufgabe 11.65 Zeigen Sie, dass die ANOVA-Hypothese

$$H_0 : \alpha_i = 0, \quad \forall i = 1, \dots, k$$

nicht testbar ist!

Um eine äquivalente testbare Hypothese aufzustellen, benutzt man

$$H_0 : \alpha_1 - \alpha_2 = 0, \dots, \alpha_1 - \alpha_k = 0 \quad \text{bzw.} \quad H_0 : H\beta = 0$$

für die $(k-1) \times (k+1)$ -Matrix

$$H = \begin{pmatrix} 0 & 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & 0 & -1 & \dots & 0 \\ \vdots & & & & & \\ 0 & 1 & 0 & \dots & -1 & 0 \\ 0 & 1 & 0 & \dots & 0 & -1 \end{pmatrix}.$$

(Zeigen Sie es!)

Bei der *zweifaktoriellen Varianzanalyse* wird die Stichprobe (Y_1, \dots, Y_n) in Abhängigkeit von 2 Faktoren in $k_1 \cdot k_2$ homogene Gruppen aufgeteilt:

$$Y_{i_1 i_2 j}, \quad j = 1, \dots, n_{i_1 i_2}$$

für $i_1 = 1, \dots, k_1$, $i_2 = 1, \dots, k_2$, sodaß

$$\sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} n_{i_1 i_2} = n.$$

Hier wird angenommen, daß

$$\mathbb{E} Y_{i_1 i_2 j} = \mu_{i_1 i_2} = \mu + \alpha_{i_1} + \beta_{i_2} + \gamma_{i_1 i_2}, \quad i_1 = 1, \dots, k_1, i_2 = 1, \dots, k_2,$$

somit stellt man folgendes lineares Modell auf:

$$Y_{i_1 i_2 j} = \mu_{i_1 i_2} + \varepsilon_{i_1 i_2 j} = \mu + \alpha_{i_1} + \beta_{i_2} + \gamma_{i_1 i_2} + \varepsilon_{i_1 i_2 j}, \\ j = 1, \dots, n_{i_1 i_2}, i_1 = 1, \dots, k_1, i_2 = 1, \dots, k_2.$$

Übungsaufgabe 11.66 Schreiben Sie die Design-Matrix X für diesen Fall explizit auf! Zeigen Sie, daß sie wieder keinen vollen Rang besitzt.

Literaturverzeichnis

- [1] P. A. R. Ade, N. Aghanim, Y. Akrami, P. K. Aluri, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, and et al. Planck 2015 results. XVI. Isotropy and statistics of the CMB. arXiv Preprint 1506.07135, 2015. <https://arxiv.org/abs/1506.07135>.
- [2] P. A. R. Ade, N. Aghanim, Y. Akrami, P. K. Aluri, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, and et al. Planck 2015 results. *Astronomy & Astrophysics*, 594:A16, Sep 2016.
- [3] A. R. Admati, P. M. DeMarzo, M. Hellwig, and P. Pfleiderer. Fallacies, irrelevant facts, and myths in the discussion of capital regulation: Why bank equity is not socially expensive. Preprints of the Max Planck Institute for Research on Collective Goods 2013/23, Bonn, 2013.
- [4] N. I. Akhiezer. *The classical moment problem and some related questions in analysis*, volume 82 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, [2021] ©2021.
- [5] G. A. Anastassiou. *Probabilistic inequalities*, volume 7 of *Series on Concrete and Applicable Mathematics*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2010.
- [6] S. Asmussen and P. W. Glynn. *Stochastic simulation: algorithms and analysis*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2007.
- [7] S. A. Aĭvazyan, V. M. Buchstaber, I. S. Enyukov, and L. D. Meshalkin. *Applied statistics*. “Finansy i Statistika”, Moscow, 1989. Classification and reduction of dimensionality.
- [8] H. Dehling, B. Haupt. *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Springer, Berlin, 2003.
- [9] N. Balakrishnan and C.-D. Lai. *Continuous bivariate distributions*. Springer, Dordrecht, second edition, 2009.

- [10] P. S. Bandyopadhyay and M. R. Forster, editors. *Philosophy of statistics*, volume 7 of *Handbook of the Philosophy of Science*. Elsevier/North-Holland, Amsterdam, 2011.
- [11] N. S. Barnett, P. Cerone, and S. S. Dragomir. *Inequalities for distributions on a finite interval*. Advances in Mathematical Inequalities Series. Nova Science Publishers, Inc., New York, 2008.
- [12] R. G. Bartle. *The Elements of Integration and Lebesgue Measure*. Wiley, 1995.
- [13] U. Bäsel and A. Duma. Buffon's problem with a cluster of needles and a lattice of rectangles. *Gen. Math.*, 18(4):127–138, 2010.
- [14] H. Bauer. *Wahrscheinlichkeitstheorie*. de Gruyter, Berlin, 1991.
- [15] H. Bauer. *Maß- und Integrationstheorie*. de Gruyter, Berlin, 2 edition, 1992.
- [16] T. Bedürftig and R. Murawski. *Philosophie der Mathematik*. De Gruyter, Berlin, revised edition, 2015.
- [17] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen. Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.*, 6(1):17–39, 1997.
- [18] P. Bickel. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume 1. Prentice Hall, 2 edition, 2001.
- [19] Borovkov. *Mathematical Statistics*. Gordon & Breach, 1998.
- [20] A. A. Borovkov. *Wahrscheinlichkeitstheorie: eine Einführung*. Birkhäuser, Basel, 1976.
- [21] A. A. Borovkov. *Probability theory*. Universitext. Springer, London, 2013. Translated from the 2009 Russian fifth edition by O. B. Borovkova and P. S. Ruzankin, Edited by K. A. Borovkov.
- [22] N. Breslow. Lessons in biostatistics. In X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, and J.-L. Wang, editors, *Past, Present, and Future of Statistical Science*, pages 335–347. Chapman & Hall, 2014.
- [23] K. Burdzy. *The search for certainty*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2009. On the clash of science and philosophy of probability.
- [24] T. Carleman. *Les fonctions quasi-analytiques*. Collection Borel. Gauthier-Villars, Paris, 1926.

- [25] B. P. Carlin and T. A. Louis. *Bayesian methods for data analysis*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, third edition, 2009.
- [26] U. Cherubini, E. Luciano, and W. Vecchiato. *Copula methods in finance*. Wiley Finance Series. John Wiley & Sons, Ltd., Chichester, 2004.
- [27] Umberto Cherubini, Fabio Gobbi, and Sabrina Mulinacci. *Convolution copula econometrics*. SpringerBriefs in Statistics. Springer, Cham, 2016.
- [28] D. V. Chudnovsky and G. V. Chudnovsky. Approximations and complex multiplication according to Ramanujan. In *Ramanujan revisited (Urbana-Champaign, Ill., 1987)*, pages 375–472. Academic Press, Boston, MA, 1988.
- [29] D. Colander, H. Föllmer, A. Haas, M. G. Goldberg, K. Juselius, A. Kirman, T. Lux, and B. Sloth. The financial crisis and the systemic failure of academic economics. *VOPROSY ECONOMIKI*, 6, 2010.
- [30] David Colander, Hans Föllmer, Armin Haas, Michael Goldberg Goldberg, Katarina Juselius, Alan Kirman, Thomas Lux, and Birgitte Sloth. The financial crisis and the systemic failure of academic economics. Discussion Papers 09-03, University of Copenhagen. Department of Economics, 2009.
- [31] J. A. Costa and A. O. Hero, III. Determining intrinsic dimension and entropy of high-dimensional shape spaces. In *Statistics and analysis of shapes*, Model. Simul. Sci. Eng. Technol., pages 231–252. Birkhäuser Boston, Boston, MA, 2006.
- [32] C. Czado. *Analyzing dependent data with vine copulas*, volume 222 of *Lecture Notes in Statistics*. Springer, Cham, 2019. A practical guide with R.
- [33] Y. Davydov, I. Molchanov, and S. Zuyev. Stability for random measures, point processes and discrete semigroups. *Bernoulli*, 17(3):1015–1043, 2011.
- [34] C. D. Daykin, T. Pentikäinen, and M. Pesonen. *Practical risk theory for actuaries*, volume 53 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London, 1994.
- [35] P. Diaconis and B. Skyrms. *Ten great ideas about chance*. Princeton University Press, Princeton, NJ, 2018.

- [36] D. Drouet Mari and S.I Kotz. *Correlation and dependence*. Imperial College Press, London; distributed by World Scientific Publishing Co., Inc., River Edge, NJ, 2001.
- [37] D. Duffie. Prone to fail: The pre-crisis financial system. *Journal of Economic Perspectives*, 33(1):81–106, February 2019.
- [38] A. Duma and M. Stoka. Geometrical probabilities for convex test bodies. II. *Ann. I.S.U.P.*, 42(2-3):39–45, 1998.
- [39] A. Duma and M. Stoka. Geometrical probabilities for convex test bodies. *Beiträge Algebra Geom.*, 40(1):15–25, 1999.
- [40] A. Duma and M. Stoka. Geometrical probabilities for convex test bodies. III. *Rend. Circ. Mat. Palermo (2) Suppl.*, (65, part I):99–108, 2000. III International Conference in “Stochastic Geometry, Convex Bodies and Empirical Measures”, Part I (Mazara del Vallo, 1999).
- [41] A. Duma and M. Stoka. Problems of Buffon type for non convex lattices. *Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur.*, 139:23–33, 2005.
- [42] A. Duma and M. Stoka. Geometric probabilities for hexagonal lattices with hexagonal obstacles. *Rend. Circ. Mat. Palermo (2) Suppl.*, 2(80):153–159, 2008.
- [43] F. Durante and C. Sempi. *Principles of copula theory*. CRC Press, Boca Raton, FL, 2016.
- [44] A. Eagle, editor. *Philosophy of probability*. Routledge Contemporary Readings in Philosophy. Routledge/Taylor & Francis Group, London, 2011. Contemporary readings.
- [45] J. Elstrodt. *Maß- und Integrationstheorie*. Springer, Berlin, 8 edition, 2018.
- [46] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events*, volume 33 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997. For insurance and finance.
- [47] K. Falconer. *Fractal geometry*. John Wiley & Sons, Ltd., Chichester, third edition, 2014. Mathematical foundations and applications.
- [48] H. Federer. *Geometric measure theory*. Springer, Berlin, repr. of the 1969 edition, 1996.
- [49] W. Feller. *An introduction to probability theory and its applications. Vol I/II*. J. Wiley & Sons, New York, 1970/71.

- [50] A. T. Fomenko. *Methods of statistical analysis of narrative texts with applications to chronology.* (in Russian). Moscow State University, 1990.
- [51] A. T. Fomenko. *Methods of mathematical analysis of historical texts. Applications to chronology.* (in Russian). Nauka, Moscow, 2nd edition, 1996.
- [52] M. Fréchet. Sur l'intégrale d'une fonctionnelle étendue à un ensemble abstrait. *Bull. Soc. Math. France*, 43:248–265, 1915.
- [53] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis.* Texts in Statistical Science Series. CRC Press, Boca Raton, FL, third edition, 2014.
- [54] H. O. Georgii. *Stochastik.* de Gruyter, Berlin, 2002.
- [55] R. D. Gill. The Monty Hall problem is not a probability puzzle. *Stat. Neerl.*, 65(1):58–71, 2011.
- [56] B. V. Gnedenko. *Einführung in die Wahrscheinlichkeitstheorie.* Akademie, Berlin, 1991.
- [57] B. V. Gnedenko. Development of probability theory. In *Outlines of the history of mathematics (Russian)*, pages 247–338. Izdat. Moskov. Univ., Moscow, 1997.
- [58] S. Gnedin. The Mondee Gills game. *Math. Intelligencer*, 34(1):34–41, 2012.
- [59] C. Graham and D. Talay. *Stochastic simulation and Monte Carlo methods*, volume 68 of *Stochastic Modelling and Applied Probability*. Springer, Heidelberg, 2013. Mathematical foundations of stochastic simulation.
- [60] A. Hald. *A history of probability and statistics and their applications before 1750.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1990.
- [61] A. Hald. *A history of mathematical statistics from 1750 to 1930.* Wiley Series in Probability and Statistics: Texts and References Section. John Wiley & Sons, Inc., New York, 1998.
- [62] A. Hald. *A history of parametric statistical inference from Bernoulli to Fisher, 1713–1935.* Sources and Studies in the History of Mathematics and Physical Sciences. Springer, New York, 2007.

- [63] R. Haller and F. Barth. *Berühmte Aufgaben der Stochastik*. De Gruyter Studium. De Gruyter, Berlin, 2017. Von den Anfängen bis heute. [From the beginnings until today], Second, revised and extended edition.
- [64] J. Y. Halpern. *Reasoning about uncertainty*. MIT Press, Cambridge, MA, second edition, 2017.
- [65] N. Henze. *Stochastik: Eine Einführung mit Grundzügen der Maßtheorie*. Springer, Berlin, 2019.
- [66] C. Hesse. *Angewandte Wahrscheinlichkeitstheorie*. Vieweg, Braunschweig, 2003.
- [67] H. Heuser. *Lehrbuch der Analysis. Teil 1*. Vieweg + Teubner, Wiesbaden, revised edition, 2003.
- [68] C. C. Heyde and E. Seneta, editors. *Statisticians of the centuries*. Springer-Verlag, New York, 2001.
- [69] M. Hofert, I. Kojadinovic, M. Mächler, and J. Yan. *Elements of copula modeling with R*. Use R! Springer, Cham, 2018.
- [70] M. L. Huber. *Perfect simulation*, volume 148 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2016.
- [71] D. Huff. *How to Lie with Statistics*. Norton, New York, 1954.
- [72] T. P. Hutchinson and C. D. Lai. *Continuous bivariate distributions, emphasising applications*. Rumsby Scientific Publishing, Adelaide, 1990.
- [73] T. P. Hutchinson and C. D. Lai. *The engineering statistician's guide to continuous bivariate distributions*. Rumsby Scientific Publishing, Adelaide, 1991.
- [74] R. Ibragimov and A. Prokhorov. *Heavy tails and copulas*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2017. Topics in dependence modelling in economics and finance, With a foreword by Joe Harry.
- [75] Jacod, J. and Protter, P. *Probability essentials*. Springer, Berlin, 2003.
- [76] P. Jaworski, F. Durante, W. Härdle, and T. Rychlik, editors. *Copula theory and its applications*, volume 198 of *Lecture Notes in Statistics—Proceedings*. Springer, Heidelberg, 2010.
- [77] H. Joe. *Dependence modeling with copulas*, volume 134 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015.

- [78] N. L. Johnson, A. W. Kemp, and S. Kotz. *Univariate discrete distributions*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2005.
- [79] N. L. Johnson and S. Kotz. *Distributions in statistics. Continuous univariate distributions. 2*. Houghton Mifflin Co., Boston, Mass., 1970.
- [80] N. L. Johnson and S. Kotz. *Distributions in statistics: continuous multivariate distributions*. John Wiley & Sons, Inc., New York-London-Sydney, 1972. Wiley Series in Probability and Mathematical Statistics.
- [81] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions. Vol. 1*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, second edition, 1994. A Wiley-Interscience Publication.
- [82] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions. Vol. 2*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, second edition, 1995. A Wiley-Interscience Publication.
- [83] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete multivariate distributions*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1997. A Wiley-Interscience Publication.
- [84] N. L. Johnson and S.l Kotz. *Distributions in statistics. Continuous univariate distributions. 1*. Houghton Mifflin Co., Boston, Mass., 1970.
- [85] O. Kallenberg. *Foundations of modern probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, Cham, third edition, 2021.
- [86] A. F. Karr. *Probability*. Springer, New York, 1993.
- [87] D. A. Klain and G.-C. Rota. *Introduction to geometric probability*. Lezioni Lincee. [Lincei Lectures]. Cambridge University Press, Cambridge, 1997.
- [88] A. Klenke. *Wahrscheinlichkeitstheorie*. Universitext. Springer, Berlin, 2. edition, 2008.
- [89] S. Kocherlakota and K. Kocherlakota. *Bivariate discrete distributions*, volume 132 of *Statistics: Textbooks and Monographs*. Marcel Dekker, Inc., New York, 1992.
- [90] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933.

- [91] A. N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea Publishing Company, New York, N. Y., 1950.
- [92] A.N. Kolmogorov and S.V. Fomin. *Reelle Funktionen und Funktionalanalysis*. Hochschulbücher für Mathematik. VEB Deutscher Verlag der Wissenschaften, Berlin, 1975.
- [93] A. E. Kossovksy. *Benford's law*. World Scientific Publishing, 2015.
- [94] S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continuous multivariate distributions. Vol. 1*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, second edition, 2000. Models and applications.
- [95] S. Kotz and S. Nadarajah. *Extreme value distributions*. Imperial College Press, London, 2000. Theory and applications.
- [96] L. F. Kozachenko and N. N. Leonenko. A statistical estimate for the entropy of a random vector. *Problemy Perekachi Informatsii*, 23(2):9–16, 1987.
- [97] T. J. Kozubowski and K. Podgórski. A generalized Sibuya distribution. *Ann. Inst. Statist. Math.*, 70(4):855–887, 2018.
- [98] W. Krämer. *So lügt man mit Statistik*. Piper Verlag, München, 2. edition, 2001.
- [99] W. Krämer. *Denkste!: Trugschlüsse aus der Welt der Zahlen und des Zufalls*. Piper Verlag, München, 2011.
- [100] M. Krein. On a problem of extrapolation of A. N. Kolmogoroff. *C. R. (Doklady) Acad. Sci. URSS (N. S.)*, 46:306–309, 1945.
- [101] U. Krengel. *Einführung in die Wahrscheinlichkeitstheorie*. Vieweg, Braunschweig, 2002.
- [102] D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo methods*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2011.
- [103] N Kusolitsch. *Maß- und Wahrscheinlichkeitstheorie*. Springer, Berlin, 2014.
- [104] N. Leonenko, L. Pronzato, and V. Savani. A class of Rényi information estimators for multidimensional densities. *Ann. Statist.*, 36(5):2153–2182, 2008.
- [105] N. Leonenko, L. Pronzato, and V. Savani. Estimation of entropies and divergences via nearest neighbors. *Tatra Mt. Math. Publ.*, 39:265–273, 2008.

- [106] Zh. Lin and Zh. Bai. *Probability inequalities*. Science Press Beijing, Beijing; Springer, Heidelberg, 2010.
- [107] J. E. Littlewood. *A mathematician's miscellany*. Methuen, London, 1957.
- [108] Jan-Frederik Mai and Matthias Scherer. *Simulating copulas*, volume 4 of *Series in Quantitative Finance*. Imperial College Press, London, 2012. Stochastic models, sampling algorithms, and applications.
- [109] Jan-Frederik Mai and Matthias Scherer. *Simulating copulas*, volume 6 of *Series in Quantitative Finance*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2017. Stochastic models, sampling algorithms, and applications, Second edition of [MR2906392], With contributions by Claudia Czado, Elke Korn, Ralf Korn and Jacob Stöber.
- [110] N. Metropolis and S. Ulam. The Monte Carlo method. *J. Amer. Statist. Assoc.*, 44:335–341, 1949.
- [111] L. Milla. A detailed proof of the Chudnovsky formula with means of basic complex analysis – Ein ausführlicher Beweis der Chudnovsky-Formel mit elementarer Funktionentheorie. *arXiv e-prints*, page arXiv:1809.00533, September 2018.
- [112] R. M. Mnatsakanov, H. Albrecher, and S. Loisel. Approximations of Copulas via Transformed Moments. *Methodol. Comput. Appl. Probab.*, 24(4):3175–3193, 2022.
- [113] I. Molchanov. *Theory of random sets*, volume 87 of *Probability Theory and Stochastic Modelling*. Springer-Verlag, London, 2017. Second edition.
- [114] P. Müller, F. A. Quintana, A. Jara, and T. Hanson. *Bayesian nonparametric data analysis*. Springer Series in Statistics. Springer, Cham, 2015.
- [115] J. A. Murphy. An analysis of the financial crisis of 2008: Causes and solutions. Ssrn preprint, 2008.
- [116] R. B. Nelsen. *An introduction to copulas*. Springer Series in Statistics. Springer, New York, second edition, 2006.
- [117] B. L. Nelson. *Foundations and methods of stochastic simulation*, volume 187 of *International Series in Operations Research & Management Science*. Springer, New York, 2013. A first course.
- [118] M. Nilsson and W. B. Kleijn. On the estimation of differential entropy from data located on embedded manifolds. *IEEE Trans. Inform. Theory*, 53(7):2330–2341, 2007.

- [119] G. V. Nosovskiy and A. T. Fomenko. *Introduction into new chronology (Which century is it now?)*. (in Russian). Kraft, Moscow, 2001.
- [120] G.V. Nosovskiy. *Age counting from Christ and calendar disputes. New chronology of Fomenko - Nosovskiy* (in Russian). Astrel, AST, Moscow, 2009.
- [121] R. Oloff. *Wahrscheinlichkeitsrechnung und Maßtheorie*. Springer, Berlin, 2017.
- [122] H. H. Panjer. Recursive evaluation of a family of compound distributions. *Astin Bull.*, 12(1):22–26, 1981.
- [123] L. Pastur and M. Shcherbina. *Eigenvalue distribution of large random matrices*, volume 171 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2011.
- [124] H.-O. Peitgen, H. Jürgens, and D. Saupe. *Chaos and fractals*. Springer-Verlag, New York, second edition, 2004. New frontiers of science.
- [125] M. D. Penrose and J. E. Yukich. Limit theory for point processes in manifolds. *Ann. Appl. Probab.*, 23(6):2161–2211, 2013.
- [126] V. V. Petrov. *Limit theorems of probability theory*, volume 4 of *Oxford Studies in Probability*. The Clarendon Press, Oxford University Press, New York, 1995. Sequences of independent random variables, Oxford Science Publications.
- [127] V. V. Petrov. Strengthenings of the Lyapunov, Hölder, and Minkowski inequalities. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 420(Veroyatnost i Statistika. 20):149–156, 2013.
- [128] I. Pinelis, V. H. de la Peña, R. Ibragimov, A. Osękowski, and I. Shevtsova. *Inequalities and extremal problems in probability and statistics*. Academic Press, London, 2017. Selected topics, Edited by the author.
- [129] O. Pons. *Inequalities in analysis and probability*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2017. 2nd Ed.
- [130] Monty Hall problem. Wikipedia, https://en.wikipedia.org/wiki/Monty_Hall_problem, Nov. 2018.
- [131] Yu. V. Prokhorov. Asymptotic behavior of the binomial distribution. *Uspehi Matem. Nauk (N.S.)*, 8(3(55)):135–142, 1953.
- [132] F. Pukelsheim. The three sigma rule. *Amer. Statist.*, 48(2):88–91, 1994.

- [133] S. Ramanujan. Modular equations and approximations to π [Quart. J. Math. **45** (1914), 350–372]. In *Collected papers of Srinivasa Ramanujan*, pages 23–39. AMS Chelsea Publ., Providence, RI, 2000.
- [134] C. R. Rao. *Statistics and truth*. World Scientific Publishing Co., Inc., River Edge, NJ, second edition, 1997. Putting chance to work, With a foreword by A. P. Mitra.
- [135] A. Rényi. A few fundamental problems of information theory. *Magyar Tud. Akad. Mat. Fiz. Oszt. Közl.*, 10:251–282, 1960.
- [136] A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 547–561. Univ. California Press, Berkeley, Calif., 1961.
- [137] C. A. Rogers. *Hausdorff measures*. Cambridge University Press, 1970.
- [138] T. Rolski, H. Schmidli, V. Schmidt, and J. Teugels. *Stochastic processes for insurance and finance*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 1999.
- [139] S. M. Ross. *Simulation*. Elsevier/Academic Press, Amsterdam, 2013. Fifth edition.
- [140] R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo method*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2017. Third edition [of MR0624270].
- [141] L. Rüschendorf. *Wahrscheinlichkeitstheorie*. Springer, Berlin, 2018.
- [142] K. A. Rybnikov. *History of Mathematics*. MSU, Moscow, 1994. In Russian.
- [143] L. Sachs. *Angewandte Statistik*. Springer, 2004.
- [144] F. Salmon. Recipe for disaster: the formula that killed Wall Street. Online paper 23.02., <https://www.wired.com/2009/02/wp-quant/>, 2009.
- [145] R. L. Schilling. *Maß und Integral*. De Gruyter, 2015.
- [146] R. L. Schilling. *Wahrscheinlichkeit*. De Gruyter, 2017.
- [147] R. L. Schilling and F. Kühn. *Counterexamples in Measure and Integration*. Cambridge University Press, 2021.
- [148] K. D. Schmidt. *Maß und Wahrscheinlichkeit*. Springer, Heidelberg, revised edition, 2011.
- [149] S. Selvin. On the Monty Hall problem (letter to the editor). *American Statistician*, 29:134, 1975.

- [150] S. Selvin. A problem in probability (letter to the editor). *American Statistician*, 29:67, 1975.
- [151] A. Shemyakin and A. Kniazev. *Introduction to Bayesian estimation and copula models of dependence*. John Wiley & Sons, Inc., Hoboken, NJ, 2017.
- [152] A. N. Shiryaev. *Probability*. Springer, New York, 1996.
- [153] M. Sibuya. Generalized hypergeometric, digamma and trigamma distributions. *Ann. Inst. Statist. Math.*, 31(3):373–390, 1979.
- [154] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.
- [155] K.-S. Song. Rényi information, loglikelihood and an intrinsic distribution measure. *J. Statist. Plann. Inference*, 93(1-2):51–69, 2001.
- [156] J. M. Steele. Darrell Huff and fifty years of *How to lie with statistics*. *Stat. Sci.*, 20(3):205–209, 2005.
- [157] W. J. Stewart. *Probability, Markov chains, queues, and simulation*. Princeton University Press, Princeton, NJ, 2009.
- [158] J. M. Stoyanov. *Counterexamples in probability*. Wiley & Sons, 1987.
- [159] B. Sundt and W. S. Jewell. Further results on recursive evaluation of compound distributions. *Astin Bull.*, 12(1):27–39, 1981.
- [160] N. T. Thomopoulos. *Essentials of Monte Carlo simulation*. Springer, New York, 2013.
- [161] H. Tijms. *Understanding probability. Chance rules in everyday life*. Cambridge University Press, 2004.
- [162] I. S. Tyurin. Refinement of the upper bounds of the constants in Lyapunov's theorem. *Russian Mathematical Surveys*, 65:586–588, 2010.
- [163] M. Úbeda Flores, E. de Amo Artero, F. Durante, and J. Fernández Sánchez, editors. *Copulas and dependence models with applications*. Springer, Cham, 2017. Contributions in honor of Roger B. Nelsen.
- [164] R. v. Mises. Fundamentalsätze der Wahrscheinlichkeitsrechnung. *Math. Z.*, 4(1-2):1–97, 1919.
- [165] R. v. Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Math. Z.*, 5(1-2):52–99, 1919.

- [166] Cosmic variance. Wikipedia, last accessed 1.09.2020. https://en.wikipedia.org/wiki/Cosmic_variance.
- [167] M. vos Savant. Ask marylin column. *Parade Magazine*, 9.09.:16, 1990.
- [168] M. vos Savant. Marylin vos savant's reply (letter to the editor). *American Statistician*, 45:347, 1991.
- [169] P. Gänßler, W. Stute. *Wahrscheinlichkeitstheorie*. Springer, Berlin, 1977.
- [170] J. Wells. Of fat cats and fat tails: From the financial crisis to the ‘new’ probabilistic marxism. In *Contradictions: Finance, Greed, and Labor Unequally Paid*, volume 28, pages 197–228. Emerald Publishing Ltd, 2013.
- [171] H. Wolthuis. *Life Insurance Mathematics (The Markovian Model)*. Caire Education Series 2, Brussels, 1994.
- [172] S. J. Yakowitz. *Computational probability and simulation*. Addison-Wesley Publishing Co., Reading, Mass.-London-Amsterdam, 1977. Applied Mathematics and Computation, No. 12.
- [173] W. H. Young. On Integration with Respect to a Function of Bounded Variation. *Proc. London Math. Soc. (2)*, 13:109–150, 1914.
- [174] V. M. Zolotarev. *One-dimensional stable distributions*, volume 65 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1986.

Index

- a-posteriori-Wahrscheinlichkeiten, 23
- a-priori-Wahrscheinlichkeit, 23
- a-posteriori-Verteilung, 150
- a-priori-Verteilung, 150
- Abgeschnittene Normalverteilung
 - Simulation
 - Akzeptanz- und Verwerfungsme-thode, 88
- Ablehnungsbereich, 175
- absolut stetige Verteilung, 34
- absolute Häufigkeit, 96
- Abweichung, mittlere quadratische, 104
- Algebra, 8
 - σ -Algebra, 9
 - Borelsche σ -Algebra, 10
 - Eigenschaften, 9
 - erzeugende Algebra, 10
 - minimale σ -Algebra, 10
- analysis of variance, *siehe* Varianzanaly-se
- Annahmebereich, 175
- ANOVA, *siehe* Varianzanalyse
- Approximation
 - Binomiale, 33
 - Poissonsche, 33
- Approximationssatz, 33
- arithmetisches Mittel, *siehe* Mittel, 100
- asymptotisch erwartungstreu, 132
- asymptotisch normalverteilt, 133
- Ausgangsvariable, 117
- Axiome von Kolmogorow, 10
- Bandbreite, 111
- Bayes-Schätzer, 150
- Bayesche Formel, 22, 23, 150
- bedingte Wahrscheinlichkeit, 20
- Bernoulli-Verteilung, 31
 - Simulation, 78
- Bernoulli-Verteilung
 - asymptotisches Konfidenzintervall, 170
- Berry
 - Satz von Berry-Esséen, 73
- besserer Schätzer, 133
- best linear unbiased estimator (BLUE), 230
- bester erwartungstreuer Schätzer, 133
- bester linearer erwartungstreuer Schät-zer, 230
- Bestimmtheitsmaß, 121, 238
- Bias, 131
- bilineare Form, 219
- bimodal, 97
- Binomiale Approximation, 33
- Binomialverteilung, 28, 31, 195
 - Simulation, 89
- Bonferroni-Ungleichung, 241
- Bootstrap
 - Konfidenzintervall, 156
 - Schätzer, 155
- Bootstrap-Schätzer
 - Monte-Carlo-Methoden, 156
- Borel-Mengen, 10
- Bose-Einstein-Statistik, 17
- Box-Plot, 102
 - modifizierter, 103
- Bravais-Pearson-Koeffizient, 115
- Bravais-Pearson-Korrelationskoeffizient, 113
- Cauchy-Verteilung, 38
- χ^2 -Verteilung, 128
- Cramér-Rao, Ungleichung von, 157
- Daten-Stichproben, 92
- Datenbereinigung, 93
- Datenerhebung, 93

Design-Matrix, 213, 227
 Dichte, 34, 43
 Dichteschätzung, 111
 disjunkt, 8
 diskrete Verteilung, *siehe* Verteilung
 Diskrete Verteilungen
 Simulation
 Akzeptanz- und Verwerfungsme-
 thode, 89
 Inversionsmethode, 88
 gleichmäßiger Abstand D_n , 141
 3σ -Regel, 37

 Effekt, 259
 Eindeutigkeitssatz
 für charakteristische Funktionen, 215
 für momenterzeugende Funktionen, 223
 Einfache lineare Regression, 117
 Einflussfaktor, 117
 einparametrische Exponentialklasse, 194
 empirische(r)
 Kovarianz, 113
 Median, 102
 Standardabweichung, 104
 Varianz, 104
 Variationskoeffizient, 104, 105
 Verteilungsfunktion, 98
 endlicher Wahrscheinlichkeitsraum, 13
 Entropie, 63
 differentiell, 63
 Rényi-, 63
 Rényi-,differentiell, 63
 Shannon-, 63
 Entscheidungsregel, 175
 Ereignis, 7
 Erlangverteilung, 128
 erwartungstreu, 131
 Erwartungstreue, 105
 Erwartungswert, *siehe* Momente
 Esséen
 Satz von Berry-Esséen, 73
 Exklusionsprinzip von Pauli, 17
 Explorative Datenanalyse, 93
 Exponentialverteilung, 38
 Exzess, 62

 Faltungsformel, 51

 Faltungsstabilität der multivariaten Normalverteilung, 218
 Faltungsstabilität
 Normalverteilung, 51
 Fehler 1. Art, 143
 Fehler 1. und 2. Art, 176
 Fermi-Dirac-Statistik, 17
 Fisher
 Fisher-Information, 148
 Fisher-Snedecor-Verteilung, F-
 Verteilung, 131
 Wölbungsmaß von Fisher, 106
 flachgipflig, *siehe* Exzess
 Fréchet-Verteilung, 40

 Gütfunktion, 176
 Gamma-Verteilung
 Simulation, 83
 Gammaverteilung
 Faltungsstabilität, 128
 Momenterzeugende und charakte-
 ristische Funktion, 127
 Satz von Gauß-Markov, 250
 Gaußsche-Verteilung, *siehe* Normalver-
 teilung
 Geburtstagsproblem, 14
 gemischte Momente, *siehe* Momente, 220
 Geometrische Verteilung, 32
 Simulation, 90
 geometrische Wahrscheinlichkeit, 19
 geometrisches Mittel, *siehe* Mittel, 100,
 101
 Gesamtstreuung, 122
 getrimmtes Mittel, 100
 gewichtetes Mittel, *siehe* Mittel
 Gleichverteilung, 32, 44
 Simulation, 76
 Gleichverteilung auf $[a, b]$, 37
 Gliwenko-Cantelli, Satz von, 141
 Grenzwertsätze, 67
 Gesetz der großen Zahlen, 67–70
 Anwendungen, 69–70
 schwaches Gesetz der großen Zah-
 len, 68
 zentraler Grenzwertsatz, 70
 Grenzwertsatz von Lindeberg, 73
 klassischer, 71–72
 Konvergenzgeschwindigkeit, 73
 Grundgesamtheit, 7, 92

Häufigkeit
 absolute, 96
 relative, 96
 harmonisches Mittel, *siehe* Mittel, 101
 Hypothesentest, *siehe* Tests
 Histogramm, 96
 eindimensionales Histogramm, 96
 Hoeffding-Ungleichung, 168
 hypergeometrische Verteilung, 18, 32
 Hypothese, 175
 Alternative, 175
 Haupthypothese, 175
 testbare, 255
 höhere Momente, *siehe* Momente
 identifizierbar, 125
 Indikator-Funktion, 26, 27
 Invarianzeigenschaften, 116
 Irrtumswahrscheinlichkeit, 163
 Jackknife-Schätzer für die/den
 Erwartungswert, 153
 Varianz, 154
 Verzerrung (Bias), 154
 k-tes Moment, *siehe* Momente
 Karl Popper, 175
 kausale und stochastische Unabhängigkeit, 22
 Kchintschin, schwaches Gesetz der großen Zahlen von, *siehe* Grenzwertsätze
 Kerndichteschätzer
 eindimensionaler Kerndichteschätzer, 111
 klassenspezifische Differenzen, 259
 Klassenstärke, 206
 klassische ANOVA-Hypothese, 260
 klassisches Wahrscheinlichkeitsmaß, 13
 Kolmogorow
 Gesetz der großen Zahlen, 69
 Kolmogorow, Satz von, 142
 Kolmogorow-Abstand D_n , 141
 Kolmogorow-Verteilung, 142
 Konfidenzintervall, 136, 163
 asymptotisches, 164, 169
 für die Bernoulli-Verteilung, 170
 für die Poissonverteilung, 170
 Bootstrap, 156
 Lange, 164
 minimales, 164
 Konfidenzniveau, 163
 konsistenter Schätzer, 132
 Konvergenz
 fast sicher, 68
 in Wahrscheinlichkeit, 68
 mit Wahrscheinlichkeit 1, 68
 stochastisch, 68
 Korrelationskoeffizient, 60, 113
 Spearmans, 115
 Kovarianz, *siehe* Momente
 Kovarianz, empirische, 113
 kritischer Bereich, *siehe* Ablehnungsreich
 Kurtosis, 106
 Lagemaß, 100
 laplacesches Wahrscheinlichkeitsmaß, 13
 Likelihood-Funktion, 145
 Lindeberg
 Grenzwertsatz von, *siehe* Grenzwertsätze
 Satz von, 73
 lineare Abhängigkeit, Maß für, *siehe* Korrelationskoeffizient
 lineare Form, 219
 lineare Regression, 213
 einfache, 229
 multiple, 229
 ohne vollen Rang, 244
 multivariate mit vollem Rang, 227
 Lineare Transformation von $N(\mu, K)$, 218
 linksschief, *siehe* Schiefe, 97
 linkssteil, *siehe* Schiefe, 97
 Markow
 schwaches Gesetz der großen Zahlen, *siehe* Grenzwertsätze
 Ungleichung von, *siehe* Ungleichungen
 maximale Streuung, 104
 Maximum-Likelihood-Schätzer, 145, 146
 schwache Konsistenz, 147
 Maxwell–Boltzmann–Statistik, 17
 Median, 100, 103
 empirischer, 102
 messbare Zerlegung, 22
 Messraum, 9
 Methode der kleinsten Quadrate, 227

- Mischungen von Verteilungen, 40
- Mittel
- arithmetisches, 53, 100
 - geometrisches, 53, 100, 101
 - getrimmtes, 100
 - gewichtetes, 53
 - harmonisches, 53, 100, 101
- Mittelwert, 37, 100, 103
- mittlere quadratische Abw. des EW, *siehe* Varianz
- mittlere quadratische Abweichung, 104
- mittlerer quadratischer Fehler, 132
- MKQ-Schätzer, 228
- Modalität, 97
- Modellierung von Daten, 93
- Modellvalidierung, 93
- modifizierter Box-Plot, 103
- Modus, 100, 103
- Momente, 53
- Erwartungswert, 54
 - absolut stetiger ZV, 55
 - Additivität, 54
 - diskreter ZV, 56
 - Monotonie, 55
 - Normalverteilung, 56
 - Poisson–Verteilung, 56
 - gemischte, 61
 - zentrales gemischtes Moment, 61
 - höhere, 61
 - k-tes Moment, 61
 - k-tes zentriertes Moment, 61
- Kovarianz, 57
- Varianz, 57
- Addition, 58
 - Normalverteilung, 59
 - Poisson–Verteilung, 59
- Momentenmethode, 143
- Momentenschätzer, 143
- Mondscheibe – Messung des Diameters, 66
- Monte–Carlo, Methode zur numerischen Integration, 69
- Monte–Carlo–Simulation, 75
- multimodal, 97
- Multinomialverteilung, 206
- Multiplikationssatz, 21
- multivariate Verteilungsfunktion, 41
- Neyman-Pearson
- Fundamentallemma, 190
- Optimalitätssatz, 189
- nicht-zentrale $\chi^2_{n,\mu}$ -Verteilung, 223
- Normalengleichung, 228
- Normalverteilung, *siehe* Verteilung
- abgeschnitten, 87
 - Erwartungswert, *siehe* Momente
 - Konfidenzintervall
 - für eine Stichprobe, 165
 - für zwei Stichproben, 172
 - multivariate, 214
 - Signifikanztests, 184
 - Simulation
 - Iterativer Algorithmus, 84
 - Polarmethode, 87
 - Varianz, *siehe* Momente
- Ordnungsstatistik, 95, 100, 101
- p*-Wert, 180
- paarweise disjunkt, 8
- Panjer–Rekursion, 89, 91
- Parameterraum, 125
- Parametervektor, 125
- Pareto–Verteilung, 39
- Simulation, 79
- Pauli, Exklusionsprinzip von, 17
- Pearson-Teststatistik, 207
- π , Berechnung von, 70
- Plug-in-Methode, 143
- Poisson–Verteilung, 32
- Erwartungswert, *siehe* Momente
 - Simulation, 90
 - Varianz, *siehe* Momente
- Poissonsche Approximation, 33
- Poissonverteilung, 186
- asymptotisches Konfidenzintervall, 170
 - Neyman-Pearson-Test, 192
- Polynomiale Regression, 118
- Polynomiale Verteilung, 44
- Potenzmenge, 8
- Problem von Galilei, 14
- Prüfverteilung, 51
- Pseudozufallszahlen, 76
- Generator, 76
 - Keim, 76
- Punktschätzer, 125
- quadratische Form, 219

Kovarianz, 220
 Quantil, 100, 101
 Quantilplot, 107
 Quartil, 100, 102
 Randomisierungsbereich, 175
 Rangkorrelationskoeffizient, 115
 Rayleigh–Verteilung, 86
 Realisierung, 95, 96
 rechtsschief, *siehe* Schiefe, 97
 rechtssteil, *siehe* Schiefe, 97
 Regressand, 117
 Regression, 117
 einfache lineare, 117
 polynomiale, 118
 Regressionsgerade, 119
 Regressionsgerade, Eigenschaften von, 121
 Regressionskoeffizient, 119
 Regressionskonstante, 119
 Regressionsvarianz, 119
 Regressor, 117
 relative Häufigkeit, 96
 Resampling-Methode, 152
 Residualplot, 124
 Residuen, 119
 Residuum, 238
 Reststreuung, 238
 Routing–Problem, 23
 Säulendiagramm, 97
 Satz
 χ^2 -Verteilung, Spezialfall, 129
 Dichte der t -Verteilung, 130
 Eigenschaften der empirischen Modelle, 133
 Gliwenko-Cantelli, 141
 Invarianzeigenschaften, 116
 Kolmogorow, 142
 Momenterzeugende und charakteristische Funktion der Gamma-verteilung, 127
 Schwache Konsistenz von ML-Schätzern, 147
 Ungleichung von Cramér-Rao, 157
 Schätzer, 131
 besserer, 133
 bester erwartungstreuer, 133
 konsistenter, 132
 Vergleich von, 133
 Schiefe, 61, 100, 106
 linksschief, 62
 linkssteil, 61
 rechtsschief, 61
 rechtssteil, 62
 Schließende Datenanalyse, 93
 Siebformel, 12
 σ –Algebra, *siehe* Algebra
 σ –Subadditivität, 12
 Simulation von Zufallsvariablen, 75
 Box–Muller–Transformation, 85
 Inversionsmethode, 78
 Polarmethode, 85
 Transformationsmethode, 78
 Simulationsmethoden
 ad hoc Methoden, 90
 Akzeptanz– und Verwerfungsme-thode, 79
 Markov–Ketten–Monte–Carlo–Methoden, 89
 Spannweite, 104
 Spearmans Korrelationskoeffizient, 115
 Spitzigkeit, *siehe* Exzess
 Störgrößen, 227
 Stabdiagramm, 97
 Standardabweichung, 37, 57, 105
 Standardnormalverteilung, *siehe* Vertei-lung
 Statistische Merkmale, 93
 steilgipflig, *siehe* Exzess
 Stichproben, 94
 Stichprobenfunktion, 95
 Stichprobenmittel, 95, 100
 Stichprobenraum, 7
 Stichprobenvarianz, 95, 104
 (stochastisch) unabhängig, 12
 stochastische Unabhängigkeit, 12
 Streudiagramm, 124
 Streuung, 37
 Stufe eines Einflussfaktors, 259
 Subadditivität des Wahrscheinlichkeits-maßes P , 12
 sum of squared residuals, 122
 sum of squares explained, 122
 sum of squares total, 122
 Symmetriekoeffizient, 61, 106
 symmetrisch, 97
 symmetrische Differenz, 8
 t -Verteilung, 130

Test
 Anpassungstest, 206
 asymptotischer, 178, 185
 auf Zusammenhang, 237
 besserer, 187
 bester, 188
 χ^2 -Anpassungstest, 206
 für Regressionsparameter, 237
 Kolmogorov-Smirnov, 206
 Macht, 176
 Monte-Carlo-Test, 178
 Neyman-Pearson-Test, 188
 Ablehnungsbereich, 188
 einseitiger, 194
 modifizierter, 202
 Parameter der Poissonverteilung, 192
 Umfang, 188
 NP-Test, *siehe* Neyman-Pearson-Test
 Parameter der Normalverteilung, 184
 parametrischer, 177
 einseitiger, 177
 linksseitiger, 177
 rechtsseitiger, 177
 zweiseitiger, 177
 parametrischer Signifikanztest, 184
 power, *siehe* Macht
 randomisierter, 175, 186
 Schärfe, 176
 Stärke, 176
 Umfang, 187
 unverfälschter, 182
 Wald-Test, 185
 Teststatistik, 165
 totale Wahrscheinlichkeit, *siehe* Bayesische Formel, 23
 Transformationsregel, 104
 Transformationssatz, 48
 linear, 49
 Tschebyschew, Ungleichung von, *siehe* Ungleichungen
 Unabhängigkeit, 46
 Charakterisierung, 46
 Ungleichungen
 Markow, 65
 Tschebyschew, 66
 unimodal, 97, 147
 unkorreliert, 57
 falls unabhängig, 57
 Untergraph, 80
 unvereinbare Ereignisse, 8
 unverzerrt, 131
 Urnenmodell, 15
 Variabilität der Erwartungswerte, 259
 Varianz, *siehe* Momente
 Varianz, empirische, 104
 Varianzanalyse, 259
 einfaktorielle, 259
 zweifaktorielle, 261
 verallgemeinerte Inverse Matrix, 244
 Verfahren von Cramér-Wold, 216
 Verlustfunktion, 150
 Verteilung, 27
 absolut stetig, 34
 Cauchy-, 38
 Eigenschaften, 35
 Exponential-, 38
 Fréchet-, 40
 Gleich-, 37
 Normal-, 37, 51
 Pareto-, 39
 Standardnormal-, 37
 diskret, 30
 Beispiele, 31
 Bernoulli-, 31
 Binomial-, 31
 geometrische, 32
 Gleich-, 32
 hypergeometrische, 32
 Poisson-, 32
 quadrierte ZV, 49
 von Zufallsvektoren, 41
 absolut stetig, 43
 diskret, 42
 Gleichverteilung, 44
 multivariate Gleichverteilung, 47
 multivariate Normalverteilung, 44, 47
 Polynomiale Verteilung, 44
 Verteilungsfunktion, 41
 Verteilung mit monotonem Dichtekoeffizienten, 194
 Verteilungsfunktion, 27
 Asymptotik, 29

- Eigenschaften, 29
Monotonie, 29
rechtsseitige Stetigkeit, 29
Verteilungsfunktion, empirische, 98
Vertrauensintervall, 136
Verzerrung, 131
Visualisierung, 93
- Wölbung, 100
Wölbungsmaß von Fisher, 106
Wahrscheinlichkeit, 10
Wahrscheinlichkeitsfunktion, 31, 42
Wahrscheinlichkeitsmaß, 10
Wahrscheinlichkeitsraum, 10
- Wölbung, *siehe* Exzess
zentrales gemischtes Moment, *siehe* Momente
zentriertes Moment, *siehe* Momente
Zielgröße, 117
Zufallsstichprobe, 95
Zufallsvariable, 26
 Momente, 53
 Produkt, 52
 Quotient, 52
 Summe von, 51
Zufallsvektor, 26
Zähldichte, 31, 42