

Stochastik II

Prof. Dr. Markus Bibinger

Institut für Mathematik

Lehrstuhl für Mathematik VIII (Angewandte Stochastik)

Patrick Bossert

Assistent

Homepage

www.mathematik.uni-wuerzburg.de/appliedstochastics

Julius-Maximilians-Universität Würzburg
Lehrstuhl für Angewandte Stochastik (Mathematik VIII)
Fakultät für Mathematik und Informatik, Institut für Mathematik

Emil-Fischer-Straße 30, D-97074 Würzburg
Tel. +49 931 31-87610
markus.bibinger@mathematik.uni-wuerzburg.de
patrick.bossert@mathematik.uni-wuerzburg.de
www.mathematik.uni-wuerzburg.de/appliedstochastics

Version: 19. März 2023

Inhaltsverzeichnis

1 Grundlagen der maßtheoretischen Stochastik	5
1.1 Zufallsvariablen, Erwartungswerte und Unabhängigkeit	5
1.2 Charakteristische Funktionen	15
2 Mehrdimensionale Verteilungen	21
2.1 Diskrete gemeinsame und marginale Verteilungen	21
2.2 Zufallsvektoren und absolutstetige Verteilungen	23
2.3 Multinomialverteilung	25
2.4 Multivariate Normalverteilung	27
2.5 Bedingte Verteilungen	31
2.5.1 Diskrete bedingte Verteilungen	31
2.5.2 Bedingte absolutstetige Verteilungen mit Dichten	36
3 Konvergenzarten der Stochastik	39
3.1 Konvergenz von Zufallsvariablen	39
3.2 Konvergenz von Verteilungen	43
3.3 Straffheit und schwache Konvergenz	49
4 Grundlagen der Statistik	55
4.1 Statistische Modelle und Statistiken	55
4.2 Parameterschätzung	57
4.2.1 Die Maximum-Likelihood-Methode	57
4.2.2 Die Momentenmethode	61
4.2.3 Eigenschaften von Schätzern	61
4.3 Statistische Tests	63
4.3.1 Einführendes Beispiel und Begriffe	63
4.3.2 Einseitige Tests	65
4.3.3 Optimalität von Tests	68
4.3.4 Zweiseitige Tests	69
4.4 Konfidenzintervalle	71
4.5 Zweistichprobenprobleme und <i>t</i> -Test	76
5 Grenzwertsätze der Stochastik	83
5.1 Univariate zentrale Grenzwertsätze	83
5.2 Statistische Anwendung: Asymptotik für Momentenschätzer	88
5.3 Multivariater zentraler Grenzwertsatz	89
5.4 Poisson-Konvergenz	91
Literaturverzeichnis	95

Inhaltsverzeichnis

Vorbemerkung:

Das vorliegende Skript ist begleitend zur 4+2 SWS Veranstaltung “Stochastik II” im Sommersemester 2023 an der Universität Würzburg. Es werden Kenntnisse der Vorlesung Stochastik I vorausgesetzt. Hyperlinks und Links sind **blau markiert**, besonders wichtige Begriffe im Text werden **rotbraun hervorgehoben** und in Sätzen und Definitionen durch **breite Schrift**. Das erste Teilkapitel 1.1 rekapituliert prägnant wichtigste Begriffe und Konzepte aus der Stochastik I und fügt einige neue Resultate und Beispiele hinzu. Für Hinweise zu Tippfehlern und Unklarheiten in diesem Skript bin ich dankbar.

1 Grundlagen der maßtheoretischen Stochastik

1.1 Zufallsvariablen, Erwartungswerte und Unabhängigkeit

Wir rekapitulieren prägnant die wichtigsten Grundbegriffe der maßtheoretischen Stochastik aus der Stochastik I. Sei stets Ω eine nichtleere Menge. $\mathcal{P}(\Omega)$ sei die Potenzmenge von Ω , die Menge aller Teilmengen von Ω . Für $A \subseteq \Omega$ bezeichne $A^c = \Omega \setminus A$ das Komplement von A in Ω . Ein Wahrscheinlichkeitsraum ist ein Tripel $(\Omega, \mathcal{A}, \mathbb{P})$, bestehend aus einer Grundmenge Ω , einer σ -Algebra \mathcal{A} über Ω , sowie einem Wahrscheinlichkeitsmaß \mathbb{P} auf dem Messraum (Ω, \mathcal{A}) . Wir erinnern an folgende Begriffe:

Eine Abbildung $\mathbb{P} : \mathcal{A} \rightarrow [0, 1]$ heißt Wahrscheinlichkeitsmaß auf (Ω, \mathcal{A}) , falls gelten

1. $\mathbb{P}(\Omega) = 1$ (Normiertheit);
2. Für (paarweise) disjunkte $A_1, A_2, \dots \in \mathcal{A}$ gilt

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k) \quad (\text{σ-Additivität}). \quad (1.1)$$

Eine Teilmenge $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ heißt eine σ -Algebra (über Ω), falls $\Omega \in \mathcal{A}$,

$$A \in \mathcal{A} \quad \Rightarrow \quad A^c \in \mathcal{A} \quad (1.2)$$

sowie

$$(A_n)_{n \geq 1} \subseteq \mathcal{A} \Rightarrow \bigcup_{n \geq 1} A_n \in \mathcal{A}. \quad (1.3)$$

Eine σ -Algebra enthält die leere Menge $\emptyset = \Omega^c$, und ist \cap -stabil:

$$(A_n)_{n \geq 1} \subseteq \mathcal{A} \Rightarrow \bigcap_{n \geq 1} A_n \in \mathcal{A}. \quad (1.4)$$

Darüber hinaus enthält \mathcal{A} endliche Vereinigungen und Schnitte. *Interpretation:* Ω Grundraum, $\omega \in \Omega$ Ausgang eines Zufallsexperiments, \mathcal{A} Ereignisraum, $A \in \mathcal{A}$ Ereignis, $\mathbb{P}(A)$ Wahrscheinlichkeit des Ereignisses A .

Eigenschaften von Wahrscheinlichkeitsmaßen: Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $A, B, A_1, A_2, \dots \in \mathcal{A}$. Für das Wahrscheinlichkeitsmaß \mathbb{P} auf dem Messraum (Ω, \mathcal{A}) gelten

1. $\mathbb{P}(\emptyset) = 0$ und damit endliche Additivität.
2. **Monotonie:** Ist $A \subseteq B$, so ist $\mathbb{P}(A) \leq \mathbb{P}(B)$.
3. **Subtraktivität:** Ist $A \subseteq B$, so ist $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$.
4. **Stetigkeit von unten:** Falls $A_n \subseteq A_{n+1}$, so gilt

$$\mathbb{P}(A_n) \uparrow \mathbb{P}\left(\bigcup_{n \geq 1} A_n\right), \quad n \rightarrow \infty.$$

5. **Stetigkeit von oben:** Falls $A_n \supseteq A_{n+1}$, $n \geq 1$, so gilt

$$\mathbb{P}(A_n) \downarrow \mathbb{P}\left(\bigcap_{n \geq 1} A_n\right), \quad n \rightarrow \infty.$$

6. **Inklusion-Exklusions-Formel:**

$$\mathbb{P}(A_1 \cup \dots \cup A_n) = \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} \mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}).$$

Seien $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und (Ω', \mathcal{A}') ein Messraum. Eine $(\mathcal{A} - \mathcal{A}')$ -messbare Abbildung $X : \Omega \rightarrow \Omega'$ heißt $((\Omega', \mathcal{A}'))$ -wertige **Zufallsvariable**. Messbarkeit bedeutet, dass $X^{-1}(A') \in \mathcal{A}$ für alle $A' \in \mathcal{A}'$. Wir verwenden die Notation $X^{-1}(A') := \{\omega \in \Omega | X(\omega) \in A'\}$. Ist
 a. $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B})$ (mit \mathcal{B} Borel σ -Algebra von \mathbb{R}), so heißt X **reellwertige Zufallsvariable**,
 b. $(\Omega', \mathcal{A}') = (\mathbb{R}^d, \mathcal{B}^d)$, so heißt X (d -varierter) **Zufallsvektor**. Ist $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, (Ω', \mathcal{A}') ein Messraum und $X : \Omega \rightarrow \Omega'$ eine Zufallsvariable, so heißt das Bildmaß von \mathbb{P} unter X die **Verteilung** von X . Wir schreiben

$$P_X(A') := \mathbb{P}(X^{-1}(A')) := \mathbb{P}(\{\omega \in \Omega | X(\omega) \in A'\}), \quad A' \in \mathcal{A}'.$$

Dies ist ein Wahrscheinlichkeitsmaß auf (Ω', \mathcal{A}') und somit ist $(\Omega', \mathcal{A}', P_X)$ ein Wahrscheinlichkeitsraum. Zur Charakterisierung von Verteilungen reellwertiger Zufallsvariablen haben wir uns vor allem **Verteilungsfunktionen** angeschaut. Wir erweitern dies hier auch auf Zufallsvektoren.

Definition 1.1. 1. Ist μ ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}^d, \mathcal{B}^d)$, so heißt die Funktion $F_\mu : \mathbb{R}^d \rightarrow [0, 1]$ mit

$$F_\mu(x_1, \dots, x_d) = \mu((-\infty, x_1] \times \dots \times (-\infty, x_d]), \quad x_1, \dots, x_d \in \mathbb{R},$$

die **Verteilungsfunktion** von μ .

2. Ist $X = (X_1, \dots, X_d)^T$ ein d -varierter Zufallsvektor mit Verteilung P_X , so heißt $F_{P_X} =: F_X$ **Verteilungsfunktion** von X , also

$$F_X(x_1, \dots, x_d) = P_X((-\infty, x_1] \times \dots \times (-\infty, x_d]) = \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d).$$

Meist beschäftigen wir uns mit Verteilungen, welche durch **Dichten** charakterisiert werden können. Sei μ ein σ -endliches Maß auf dem Messraum (Ω', \mathcal{A}') , und $f : \Omega' \rightarrow [0, \infty]$ sei messbar. Dann ist

$$\mu_f(A) := \int_A f d\mu, \quad A \in \mathcal{A}'$$

ein Maß auf (Ω', \mathcal{A}') . Es heißt Maß mit **Dichte** f bzgl. μ . Es gilt, dass

- i.) μ_f ist ein Wahrscheinlichkeitsmaß $\Leftrightarrow \int f d\mu = 1$.
- ii.) $\mu_f = \mu_g \Leftrightarrow f = g$ μ -fast überall.
- iii.) Ist $A \in \mathcal{A}'$, $\mu(A) = 0$, so ist auch $\mu_f(A) = 0$.

Der **Satz von Radon-Nikodym** aus der Stochastik I garantiert die Existenz einer Dichte, falls Absolutstetigkeit bzgl. μ gilt. Die meisten relevanten Verteilungen lassen sich charakterisieren durch die Fälle

- a.) $\Omega' = \mathbb{R}^d$, $\mu = \lambda^d$, dann heißt die Verteilung **absolutstetig**,
oder
- b.) $\Omega' \subset \mathbb{R}^d$ abzählbar (endlich oder abzählbar unendlich), $\mu = \text{Zählmaß}$ auf Ω' , dann heißt die Verteilung **diskret**.

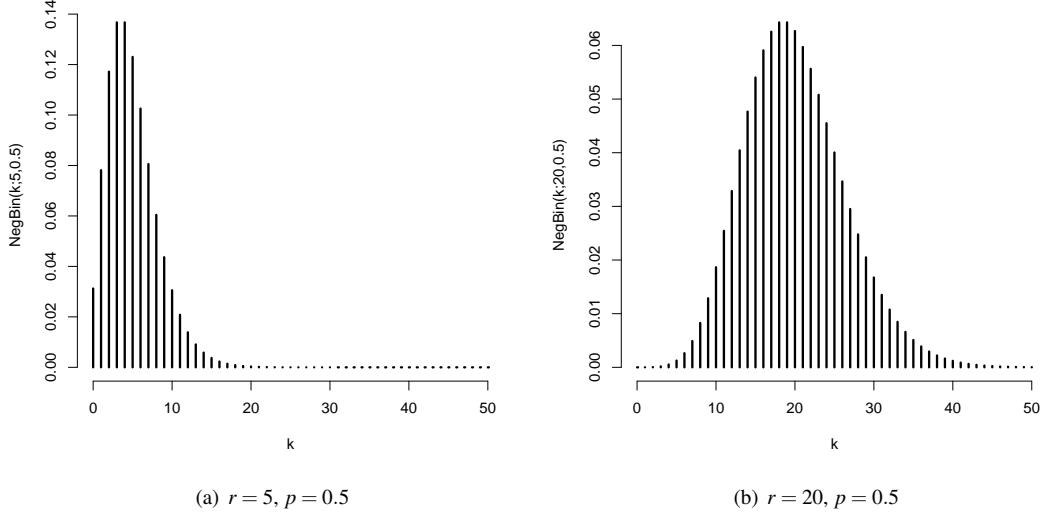
Beispiele 1.2. Eindimensionale Verteilungen

a. Die **Cauchy-Verteilung** ist eine absolutstetige Verteilung mit Lebesgue-Dichte

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

b. Wir führen eine diskrete Verteilung mit Zähldichte (auch ‘‘Wahrscheinlichkeitsfunktion’’) auf \mathbb{N}_0 ein. Gesucht ist die Verteilung der Zufallsvariable X , die die Anzahl der Misserfolge vor dem r -ten Erfolg ($r \geq 1$ fest) bei unendlich vielen Bernoulli- p -Experimenten darstellt. Für ein $X = k$ müssen bis zum $(r+k)$ -ten Experiment k Misserfolge und r Erfolge (davon einer im $(r+k)$ -ten Experiment) auftreten. Dafür gibt es

$$\binom{k+r-1}{k} = \binom{k+r-1}{r-1}$$


 Abbildung 1.1: Zähldichte der negativen Binomialverteilung $\text{Negbin}(r, p)$.

Möglichkeiten, von denen jede die Wahrscheinlichkeit $p^r(1-p)^k$ hat. Somit setzen wir

$$p(k) = \binom{k+r-1}{k} p^r (1-p)^k, \quad k \in \mathbb{N}_0.$$

Ist $r \in \mathbb{N}$ und $p \in (0, 1)$, so ist $p : \mathbb{N}_0 \rightarrow [0, 1]$ mit

$$p(k) = \binom{k+r-1}{k} p^r (1-p)^k$$

eine Wahrscheinlichkeitsfunktion. Die zugehörige Verteilung heißt die **negative Binomialverteilung** mit Parametern r und p . Ist eine Zufallsvariable X verteilt nach $X \sim p$, so schreiben wir $X \sim \text{Negbin}(r, p)$.

\mathbb{P} ist eine Wahrscheinlichkeitsverteilung, also muss gelten

$$\sum_{k=0}^{\infty} \binom{k+r-1}{k} p^r (1-p)^k \stackrel{!}{=} 1. \quad (1.5)$$

Mit der Binomischen Reihe gilt für $|x| < 1$ und $a \in \mathbb{R}$

$$(1+x)^a = \sum_{k=0}^{\infty} \binom{a}{k} x^k, \quad \binom{a}{k} = \frac{a \cdot (a-1) \cdots (a-k+1)}{k!} = \prod_{i=1}^k \frac{a+1-i}{i}. \quad (1.6)$$

Negieren wir in (1.6) x und a , so erhalten wir

$$(1-x)^{-a} = \sum_{k=0}^{\infty} \binom{-a}{k} (-1)^k x^k. \quad (1.7)$$

Dabei ist

$$\begin{aligned} \binom{-a}{k} &= \frac{-a \cdot (-a-1) \cdots (-a-k+1)}{k!} \\ &= (-1)^k \frac{a \cdot (a+1) \cdots (a+k-1)}{k!} = (-1)^k \binom{a+k-1}{k}, \end{aligned}$$

also erhalten wir aus (1.7)

$$(1-x)^{-a} = \sum_{k=0}^{\infty} \binom{a+k-1}{k} x^k.$$

Setzen wir $x = (1-p)$ und $a = r$, so ergibt sich

$$p^{-r} = (1-(1-p))^{-r} = \sum_{k=0}^{\infty} \binom{k+r-1}{k} (1-p)^k,$$

also (1.5). \diamond

Einer der zentralen Begriffe der Stochastik I war der **Erwartungswert**. Heuristisch ist der Erwartungswert der mittlere Wert einer Zufallsvariablen, also das gemäß eines Wahrscheinlichkeitsmaßes gewichtete Mittel. Die allgemeine formale Definition war über das Integral einer messbaren Funktion auf einem Maßraum. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und sei $X : \Omega \rightarrow \overline{\mathbb{R}}$ eine numerische Zufallsvariable¹. Ist X quasiintegrierbar bzgl. \mathbb{P} , d.h. $\min(\int X^+ d\mathbb{P}, \int X^- d\mathbb{P}) < \infty$, so setze

$$\mathbb{E}[X] := \int_{\Omega} X d\mathbb{P}.$$

$\mathbb{E}[X]$ ist der **Erwartungswert** von X . Soll das zugrunde liegende Maß betont werden, schreiben wir $\mathbb{E}_{\mathbb{P}}[X]$. Ist X integrierbar, also $\mathbb{E}[|X|] < \infty$, so hat X einen **endlichen Erwartungswert**.

Erinnerung zur Integralkonstruktion:

Die Konstruktion des allgemeinen Maßintegrals verlief in drei Schritten. Analog überträgt sich dies auf Erwartungswerte.

1. Für nicht-negative Treppenfunktionen $X = \sum_{k=1}^n \alpha_k \mathbf{1}_{A_k}$ mit $\alpha_k \geq 0$, A_1, \dots, A_n eine Zerlegung von Ω , ist

$$\mathbb{E}[X] = \int_{\Omega} X d\mathbb{P} = \sum_{k=1}^n \alpha_k \mathbb{P}(A_k).$$

2. Für nicht-negative numerische Zufallsvariablen $X : \Omega \rightarrow [0, \infty]$ existiert eine monoton wachsende Folge (X_n) von nicht-negativen Treppenfunktionen mit $X_n \nearrow X$ punktweise. Setze

$$\mathbb{E}[X] = \lim_{n \rightarrow \infty} \mathbb{E}[X_n].$$

3. Für allgemeines X , setze $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$. Wir werden die **Konvergenzsätze der Integrations-theorie** aus der Stochastik I häufig benötigen. Seien X_n, X numerische Zufallsvariablen.

1. **Monotone Konvergenz.** Ist $X_n \geq 0$, $X_n \nearrow X$ punktweise, so gilt $\mathbb{E}[X_n] \rightarrow \mathbb{E}[X]$.

2. **Domierte/Majorisierte Konvergenz (Satz von Lebesgue).** Gilt $X_n \rightarrow X$ \mathbb{P} -fast sicher, und existiert eine nicht-negative Zufallsvariable Y mit $|X_n| \leq Y$ für alle n sowie $\mathbb{E}[Y] < \infty$, so sind X_n und X alle integrierbar und es gilt

$$\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - X|] = 0, \quad \text{insbesondere } \mathbb{E}[X] = \lim_{n \rightarrow \infty} \mathbb{E}[X_n].$$

3. **Lemma von Fatou.** Ist $X_n \geq 0$, so gilt

$$\mathbb{E}\left[\liminf_{n \rightarrow \infty} X_n\right] \leq \liminf_{n \rightarrow \infty} (\mathbb{E}[X_n]).$$

Die konkrete Berechnung von Erwartungswerten erfolgt häufig über die Verteilung der Zufallsvariablen. Sei $X : \Omega \rightarrow \Omega'$ eine (Ω', \mathcal{A}') -wertige Zufallsvariable und $h : \Omega' \rightarrow \mathbb{R}$ sei Borel-messbar. Dann ist h eine Zufallsvariable auf $(\Omega', \mathcal{A}', P_X)$, und $h(X)$ eine Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$. Nach der **Transformations-formel** ist $h(X)$ genau dann quasiintegrierbar (über $(\Omega, \mathcal{A}, \mathbb{P})$), falls h quasiintegrierbar über $(\Omega', \mathcal{A}', P_X)$ ist, und dann gilt

$$\mathbb{E}[h(X)] = \int_{\Omega'} h dP_X. \tag{1.8}$$

¹Siehe Kapitel 3.3 in Stochastik I zur erweiterten Zahlengerade $\overline{\mathbb{R}}$ und numerischen Zufallsvariablen.

Für den Erwartungswert einer reellwertigen Zufallsvariablen erhalten wir also $\mathbb{E}[X] = \int_{\mathbb{R}} x dP_X(x)$. Wir nennen $\mathbb{E}[h(X)]$ das h -Moment von X , insbesondere

a. für $h(x) = x^k$: **k -tes Moment**.

b. für $h(x) = (x - \mathbb{E}[X])^k$: **k -tes zentriertes Moment**. Für $k = 2$ erhält man die **Varianz** von X .

Nach (1.8) hängen die Momente, also insbesondere der Erwartungswert, nur von der Verteilung der Zufallsvariable ab.

Wir hatten einige wichtige Ungleichungen mit Erwartungswerten, wie die Markov-Ungleichung, in der Stochastik I behandelt. Eine der wichtigsten lernen wir nachfolgend kennen. Dazu benötigen wir den Begriff der konvexen Funktionen aus der Analysis I. Sei $I \subset \mathbb{R}$ ein Intervall, dann heißt $\phi : I \rightarrow \mathbb{R}$ **konvex**, falls für alle $x, y \in I$, $\lambda \in [0, 1]$ gilt

$$\phi(\lambda x + (1 - \lambda)y) \leq \lambda \phi(x) + (1 - \lambda)\phi(y).$$

ϕ heißt **strikt konvex**, falls für alle $x \neq y \in I$, $\lambda \in (0, 1)$ sogar gilt

$$\phi(\lambda x + (1 - \lambda)y) < \lambda \phi(x) + (1 - \lambda)\phi(y).$$

Satz 1.3 (Jensen-Ungleichung). Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $I \subset \mathbb{R}$ ein Intervall und $X : \Omega \rightarrow I$ eine integrierbare Zufallsvariable. Dann ist $\mathbb{E}[X] \in I$ und ist $\phi : I \rightarrow \mathbb{R}$ konvex, so ist $\phi(X)$ quasiintegrierbar mit

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]. \quad (1.9)$$

Ist ϕ sogar strikt konvex und X nicht fast sicher konstant, so gilt

$$\phi(\mathbb{E}[X]) < \mathbb{E}[\phi(X)]. \quad (1.10)$$

Bevor wir zum Beweis kommen, benötigen wir einige Tatsachen über konvexe Funktionen.

Einschub über konvexe Funktionen.

Satz 1.4. Sei $\phi : I \rightarrow \mathbb{R}$ konvex. Dann ist ϕ stetig in $\text{int}(I)$ und insbesondere Borel-messbar. Ist $x \in \text{int}(I)$, so existiert die rechtseitige Ableitung

$$\phi'_+(x) = \lim_{h \searrow 0} \frac{\phi(x+h) - \phi(x)}{h} \in \mathbb{R},$$

und die linksseitige Ableitung

$$\phi'_-(x) = \lim_{h \searrow 0} \frac{\phi(x) - \phi(x-h)}{h} \in \mathbb{R},$$

und $\phi'_-(x) \leq \phi'_+(x)$. Weiter gilt für alle $c \in [\phi'_-(x), \phi'_+(x)]$

$$\phi(y) \geq \phi(x) + c(y-x), \quad y \in I, \quad (1.11)$$

und ist ϕ strikt konvex, so gilt sogar

$$\phi(y) > \phi(x) + c(y-x), \quad y \in I \setminus \{x\}.$$

Beweis. Für $x, y \in I$, $x < y$, definieren wir den Differenzenquotienten

$$D\phi(x, y) = \frac{\phi(y) - \phi(x)}{y - x}$$

Wir behaupten:

$$\forall x, y, z \in I, x < y < z \quad \text{gilt} \quad D\phi(x, y) \leq D\phi(x, z) \leq D\phi(y, z), \quad (1.12)$$

wobei im strikt konvexen Fall die strikten Ungleichungen gelten. Zum Nachweis der ersten Ungleichung benutze man die Konvexität für

$$y = \left(1 - \frac{y-x}{z-x}\right)x + \frac{y-x}{z-x}z.$$

Umstellen liefert dann die Behauptung. Die zweite ergibt sich ähnlich.

1. ϕ ist stetig in $x \in \text{int}(I)$: Seien $x_1, x_2 \in \text{int}(I)$ mit $x_1 < x < x_2$. Wir zeigen, dass ϕ in $[x_1, x_2]$ Lipschitzstetig ist. Dazu wähle $x_2 < z \in I$, sind dann $y_1, y_2 \in [x_1, x_2]$, $y_1 < y_2$, so folgt mit (1.12)

$$D\phi(y_1, y_2) \leq D\phi(x_2, z),$$

also die Lipschitz-Stetigkeit mit Konstante $D\phi(x_2, z)$.

2. (1.12) impliziert, dass der Differenzenquotient an einer festen Stelle monoton fallend ist, somit existieren linksseitige und rechtsseitige Ableitungen. Für $y > x$ folgt wegen $D\phi(x, y) \geq \phi'_+(x)(y - x)$ dann

$$\phi(y) = \phi(x) + D\phi(x, y)(y - x) \geq \phi(x) + \phi'_+(x)(y - x),$$

und analog für $y < x$ wegen $D\phi(x, y) \leq \phi'_-(x)$

$$\phi(y) = \phi(x) + D\phi(x, y)(y - x) \geq \phi(x) + \phi'_-(x)(y - x).$$

Wegen $\phi'_-(x) \leq \phi'_+(x)$, liefert dies (1.11). Analog ergibt sich die Aussage bei strikter Konvexität. ■

Wir erinnern weiter daran, dass eine auf I stetige, in $\text{int}(I)$ zweimal differenzierbare Funktion ϕ mit $\phi'' \geq 0$ in $\text{int}(I)$ konvex ist. Gilt für alle $x \in \text{int}(I)$ sogar $\phi''(x) > 0$, so ist ϕ strikt konvex.

Beweis von Satz 1.3. Ist X fast sicher konstant, so ist $X = \mathbb{E}[X]$ fast sicher und in (1.9) steht eine Gleichheit.

Sei nun X nicht fast sicher konstant. Dann ist $\mathbb{E}[X] \in \text{int}(I)$. Setze $x = \mathbb{E}[X]$, dann folgt aus (1.11)

$$\phi(X) \geq \phi(x) + \phi'_+(x)(X - x).$$

Da die rechte Seite integrierbar ist, ist die linke Seite $\phi(X)$ quasiintegrierbar. Erwartungswertbildung liefert (1.9). Ist ϕ strikt konvex, so ist wegen $\mathbb{P}(X \neq \mathbb{E}[X]) > 0$ auch

$$\mathbb{P}(\phi(X) > \phi(x) + \phi'_+(x)(X - x)) > 0,$$

daher folgt (1.10). ■

Korollar 1.5. Sei $p > 1$ und $\mathbb{E}[|X|^p] < \infty$, so ist auch $\mathbb{E}[|X|] < \infty$ und es gilt

$$\mathbb{E}[|X|] \leq (\mathbb{E}[|X|^p])^{1/p},$$

wobei sogar die strikte Ungleichung gilt, falls X nicht fast sicher konstant ist.

Beweis. Da $|X| \leq 1 + |X|^p$ ist $\mathbb{E}[|X|] < \infty$. Die Jensen-Ungleichung angewendet auf $|X|$ mit der strikt konvexen Funktion $\phi(t) = t^p$ liefert

$$(\mathbb{E}[|X|])^p = \phi(\mathbb{E}[|X|]) \leq \mathbb{E}[\phi(|X|)] = \mathbb{E}[|X|^p],$$

und damit die Ungleichung. ■

Die folgenden Hölder- und Minkowski-Ungleichungen geben wir für beliebige Maßräume an. Sei $(\Omega, \mathcal{A}, \mu)$ ein Maßraum und $f : \Omega \rightarrow \mathbb{R}$ messbar. Sei $p > 0$. Gilt $\int_{\Omega} |f|^p d\mu < \infty$, so setze

$$\|f\|_p := \left(\int_{\Omega} |f|^p d\mu \right)^{1/p}.$$

$\|\cdot\|_p$ für $p \in [1, \infty)$ ist eine Seminorm, die L_p -Seminorm. Korollar 1.5 entnehmen wir zunächst folgendes Korollar.

Korollar 1.6. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $0 < r < s$ und X eine reellwertige Zufallsvariable mit $\mathbb{E}[|X|^s] < \infty$. Dann ist auch $\mathbb{E}[|X|^r] < \infty$ und es gilt

$$\|X\|_r \leq \|X\|_s,$$

wobei sogar die strikte Ungleichung gilt, falls X nicht fast sicher konstant ist.

Beweis. Wende Korollar 1.5 mit $|X|^r$ und $p = s/r$ an. ■

Satz 1.7. Sei $(\Omega, \mathcal{A}, \mu)$ ein Maßraum und seien $f, g : \Omega \rightarrow \mathbb{R}$ messbar. Dann gilt

i.) **Hölder-Ungleichung:** Ist $p \in (1, \infty)$ und $q := p/(p-1)$, also $1/p + 1/q = 1$, und sind $\int |f|^p d\mu < \infty$, $\int |g|^q d\mu < \infty$, so ist $f \cdot g$ integrierbar und es gilt

$$\|fg\|_1 \leq \|f\|_p \|g\|_q. \quad (1.13)$$

Insbesondere gilt für $p = q = 2$ die **Cauchy-Schwarz-Ungleichung**

$$\|fg\|_1 \leq \|f\|_2 \|g\|_2. \quad (1.14)$$

ii.) **Minkowski-Ungleichung:** Ist $p \in (1, \infty)$, und sind $\int |f|^p d\mu < \infty$, $\int |g|^p d\mu < \infty$, so ist

$$\|f+g\|_p \leq \|f\|_p + \|g\|_p. \quad (1.15) \quad ■$$

Beweis. Zu i.) Der Beweis basiert auf (einem Spezialfall) der Youngschen Ungleichung: Für $x, y \geq 0$ ist

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}.$$

Dies kann aus der Konvexität der Exponentialfunktion gefolgert werden:

$$xy = \exp\left(\frac{1}{p} \log x^p + \frac{1}{q} \log y^q\right) \leq \frac{1}{p} \exp(\log x^p) + \frac{1}{q} \exp(\log y^q) = \frac{x^p}{p} + \frac{y^q}{q}.$$

Für $x = f/\|f\|_p$, $y = g/\|g\|_q$ liefert dies (der Fall $f = 0$ oder $g = 0$ μ -f.ü. ist klar)

$$\frac{|fg|}{\|f\|_p \|g\|_q} \leq \frac{|f|^p}{p \|f\|_p^p} + \frac{|g|^q}{q \|g\|_q^q}.$$

Integration und $1/p + 1/q = 1$ liefern die Behauptung.

Zu ii.) Da $|f+g| \leq 2 \max(|f|, |g|)$, folgt

$$|f+g|^p \leq 2^p \max(|f|^p, |g|^p) \leq 2^p (|f|^p + |g|^p),$$

also aus $\int |f|^p d\mu < \infty$, $\int |g|^p d\mu < \infty$ auch $\int |f+g|^p d\mu < \infty$. Wende nun auf $|f|$ und $|f+g|^{p-1}$ die Hölder-Ungleichung mit p und $q = p/(p-1)$ an und erhalte

$$\int |f| |f+g|^{p-1} \leq \|f\|_p \|f+g\|_p^{p-1}.$$

Analog $\int |g| |f+g|^{p-1} \leq \|g\|_p \|f+g\|_p^{p-1}$. Addition der beiden Ungleichungen liefert

$$\begin{aligned} (\|f\|_p + \|g\|_p) \|f+g\|_p^{p-1} &\geq \int (|f| + |g|) |f+g|^{p-1} d\mu \\ &\geq \int |f+g|^p d\mu = \|f+g\|_p^p, \end{aligned}$$

und damit die Behauptung. ■

Ein wesentlicher Begriff der Stochastik ist die **Unabhängigkeit** von Ereignissen, Zufallsvariablen und Mengensystemen. Wir erinnern, dass endlich viele Ereignisse A_1, \dots, A_n unabhängig heißen, falls für alle $I \subseteq \{1, \dots, n\}$ gilt

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i).$$

Eine beliebe Familie von Ereignissen $(A_i)_{i \in I} \subset \mathcal{A}$ heißt unabhängig, falls jede endliche Teilstamme unabhängig ist.

Mengensysteme $\mathcal{E}_1, \dots, \mathcal{E}_n \subseteq \mathcal{A}$ heißen unabhängig, falls für alle $A_1 \in \mathcal{E}_1, \dots, A_n \in \mathcal{E}_n$ die Ereignisse A_1, \dots, A_n unabhängig sind. Schließlich heißt eine beliebige Familie von Mengensystemen $(\mathcal{E}_i)_{i \in I}$, $\mathcal{E}_i \subseteq \mathcal{A}$ unabhängig, falls wiederum jede endliche Teilstamme unabhängig ist.

Von der Unabhängigkeit von Erzeugern kann unter Voraussetzungen bereits auf die Unabhängigkeit erzeugter σ -Algebren geschlossen werden. Das Resultat aus der Stochastik I dazu war: Sind $\mathcal{E}_1, \dots, \mathcal{E}_n \subseteq \mathcal{A}$ unabhängige Mengensysteme und ist jedes $\mathcal{E}_i \cap$ -stabil, so sind auch $\sigma(\mathcal{E}_1), \dots, \sigma(\mathcal{E}_n)$ unabhängig.

Seien $(\Omega_i, \mathcal{A}_i)$ Messräume und $X_i : \Omega \rightarrow \Omega_i$ Zufallsvariablen, $i = 1, \dots, n$. Dann heißen X_1, \dots, X_n **unabhängig**, falls für alle $A_i \in \mathcal{A}_i$ die Ereignisse $\{X_i \in A_i\}$, $i = 1, \dots, n$, unabhängig sind. Eine beliebige Familie von Zufallsvariablen heißt unabhängig, falls jede endliche Teilstamme unabhängig ist. Für Zufallsvariablen $X_i : \Omega \rightarrow \Omega_i$, $i = 1, \dots, n$, sind nach Resultaten aus der Stochastik I äquivalent:

1. X_1, \dots, X_n sind unabhängig.
2. Die σ -Algebren $\sigma(X_i) := X_i^{-1} \mathcal{A}_i = \{X_i^{-1} A_i, A_i \in \mathcal{A}_i\}$ sind unabhängig.
3. Für alle $A_i \in \mathcal{A}_i$ gilt

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdot \dots \cdot \mathbb{P}(X_n \in A_n).$$

4. Für reellwertige Zufallsvariablen gilt

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

5a. Sind $X_i : \Omega \rightarrow S_i$ diskret mit S_i abzählbar, so gilt

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdot \dots \cdot \mathbb{P}(X_n = x_n) \quad \forall x_i \in S_i.$$

5b. Sei $\mathbf{X} = (X_1, \dots, X_n)^T$ ein Zufallsvektor mit der Dichte $f : \mathbb{R}^n \rightarrow \mathbb{R}$, sowie f_i die Lebesgue-Dichten von X_i für alle $i = 1, \dots, n$. Dann gilt für $\mathbf{x} \in \mathbb{R}^n$ Lebesgue-fast überall

$$f(\mathbf{x}) = f_1(x_1) \cdot \dots \cdot f_n(x_n).$$

6. Die Verteilung von (X_1, \dots, X_n) auf $\Omega_1 \times \dots \times \Omega_n$ ist das **Produktmaß**

$$P_{(X_1, \dots, X_n)} = P_{X_1} \otimes \dots \otimes P_{X_n}.$$

Unter Unabhängigkeit hatten wir die folgenden Resultate zu Erwartungswerten in der Stochastik I. Sind

$X : \Omega \rightarrow \Omega_1, Y : \Omega \rightarrow \Omega_2$ unabhängige Zufallsvariablen und $f : \Omega_1 \rightarrow \mathbb{R}, g : \Omega_2 \rightarrow \mathbb{R}$ messbar, und gilt $f(X), g(Y) \geq 0$ oder $f(X)$ und $g(Y)$ integrierbar, so ist

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]. \quad (1.16)$$

Sind X_1, \dots, X_n unabhängig und gilt $X_i \geq 0, i = 1, \dots, n$, oder X_i integrierbar, $i = 1, \dots, n$, so gilt

$$\mathbb{E}\left[\prod_{i=1}^n X_i\right] = \prod_{i=1}^n \mathbb{E}[X_i].$$

Als Anwendung zeigen wir nun eine spezielle Version der **Hoeffding-Ungleichung**, einer wichtigen **Exponentialungleichung** unter den **Konzentrationsungleichungen**.

Satz 1.8. Seien X_1, \dots, X_n unabhängige Zufallsvariablen mit $\mathbb{E}[X_i] = 0, |X_i| \leq M$ f.s., $i = 1, \dots, n$ für ein $M > 0$. Dann gelten

$$\mathbb{P}\left(\sum_{j=1}^n X_j \geq t\right) \leq \exp\left(-\frac{t^2}{2nM^2}\right), \quad t \geq 0,$$

sowie

$$\mathbb{P}\left(\left|\sum_{j=1}^n X_j\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2nM^2}\right), \quad t \geq 0.$$

■

Die Hoeffding-Ungleichung kann umgeschrieben werden zu

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{j=1}^n X_j\right| \geq t\right) \leq 2 \exp\left(-\frac{nt^2}{2M^2}\right), \quad t > 0. \quad (1.17)$$

Lemma 1.9. Sei Y eine Zufallsvariable mit $|Y| \leq M$ sowie $\mathbb{E}[Y] = 0$. Dann gilt

$$\mathbb{E}[\exp(sY)] \leq \cosh(sM) \leq \exp(s^2 M^2 / 2), \quad s \geq 0.$$

Beweis. Sei $Z = (Y + M)/(2M)$, dann ist $Z \in [0, 1]$ mit $\mathbb{E}[Z] = 1/2$ und

$$Y = ZM + (1 - Z)(-M).$$

Wegen der Konvexität von $x \mapsto \exp(sx)$ folgt

$$\exp(sY) \leq Z \exp(sM) + (1 - Z) \exp(-sM).$$

Anwendung des Erwartungswertes liefert die erste Ungleichung. Für die zweite beachte dass

$$\cosh x = \frac{1}{2}(e^x + e^{-x}) = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} \leq \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n n!} = e^{x^2/2}.$$

■

Wir benötigen folgende Version der Markov-Ungleichung.

Lemma 1.10. Sei $Z : \Omega \rightarrow \mathbb{R}$ Zufallsvariable, $\phi : \mathbb{R} \rightarrow [0, \infty)$ eine monoton wachsende Funktion. Dann gilt für $t \in \mathbb{R}$ mit $\phi(t) > 0$

$$\mathbb{P}(Z \geq t) \leq \frac{1}{\phi(t)} \mathbb{E}[\phi(Z) \mathbf{1}_{\{Z \geq t\}}] \leq \frac{1}{\phi(t)} \mathbb{E}[\phi(Z)].$$

Beweis. Da $\phi(z) \geq 0$ für alle $z \in \mathbb{R}$, folgt

$$\phi(Z) \mathbf{1}_{\{Z \geq t\}} \leq \phi(Z),$$

und Erwartungswertbildung liefert die zweite Ungleichung. Für die erste beachte

$$\phi(t) \mathbf{1}_{\{Z \geq t\}} \leq \phi(Z) \mathbf{1}_{\{Z \geq t\}},$$

da ϕ monoton wachsend. Anwendung des Erwartungswertes liefert die Behauptung. ■

Beweis von Satz 1.8. Für $s \geq 0$ gilt mit der Markov-Ungleichung aus Lemma 1.10, dass

$$\mathbb{P}\left(\sum_{j=1}^n X_j \geq t\right) \leq e^{-st} \mathbb{E}\left[\exp\left(s \sum_{j=1}^n X_j\right)\right] = e^{-st} \prod_{j=1}^n \mathbb{E}\left[\exp(sX_j)\right],$$

wobei wir im zweiten Schritt die Faktorisierung der Erwartungswerte unter Unabhängigkeit angewendet haben und genutzt haben, dass

$$\exp(sX_1), \dots, \exp(sX_n)$$

unabhängig sind. Mit Lemma 1.9 können wir weiter abschätzen

$$\mathbb{P}\left(\sum_{j=1}^n X_j \geq t\right) \leq \exp(ns^2 M^2 / 2 - st) \quad \forall s \geq 0.$$

Für $s = t/(nM^2)$ ergibt sich die erste Aussage des Satzes. Für die zweite beachte, dass für $t > 0$

$$\mathbb{P}\left(\left|\sum_{j=1}^n X_j\right| \geq t\right) = \mathbb{P}\left(\sum_{j=1}^n X_j \geq t\right) + \mathbb{P}\left(\sum_{j=1}^n (-X_j) \geq t\right),$$

und wende auf beide Terme die einseitige Ungleichung an (auch $-X_1, \dots, -X_n$ erfüllen die Voraussetzung für die einseitige Ungleichung). ■

Beispiel 1.11. Seien X_1, \dots, X_n unabhängig und $\text{Ber}(p)$ -verteilt, $p \in (0, 1)$. Aus der Tschebyschev-Ungleichung erhalten wir

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{j=1}^n X_j - p\right| \geq t\right) \leq \frac{1}{4nt^2}, \quad t > 0,$$

da $\text{Var}(X_i) = p(1-p) \leq 1/4$, $p \in (0, 1)$. Aus der symmetrischen Version der Hoeffding-Ungleichung, angewendet auf $X_i - p$ mit $M = 1$ folgt

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{j=1}^n X_j - p\right| \geq t\right) \leq 2 \exp\left(-\frac{nt^2}{2}\right), \quad t > 0.$$

Ist speziell $p = 1/2$, also $X_i \sim \text{Ber}(1/2)$, so ist $|X_i - 1/2| = 1/2$ und es kann $M = 1/2$ gewählt werden, so dass sich die bessere Abschätzung

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{j=1}^n X_j - \frac{1}{2}\right| \geq t\right) \leq 2 \exp\left(-2nt^2\right), \quad t > 0$$

ergibt. Aus einer allgemeineren Version der Hoeffding-Ungleichung kann man diese verbesserte Abschätzung für jeden Parameter p erhalten.

1.2 Charakteristische Funktionen

Die Verteilungsfunktion eines Wahrscheinlichkeitsmaßes μ auf $(\mathbb{R}^d, \mathcal{B}^d)$ charakterisiert μ eindeutig. Wir lernen hier eine weitere Funktionenklasse kennen, sogenannte **charakteristische Funktionen** bzw. **Fouriertransformierte**, welche Wahrscheinlichkeitsmaße eindeutig bestimmen. Diese haben zum einen analytische Vorteile gegenüber Verteilungsfunktionen. Darauf hinaus erlauben sie Dimensionsreduktion vieler Aussagen von $d > 1$ auf $d = 1$, und sind von besonderer Bedeutung bei der Beschreibung von Verteilungskonvergenz und für den zentralen Grenzwertsatz.

Integration komplexwertiger Funktionen

Sei $(\Omega, \mathcal{A}, \mu)$ ein Maßraum. Eine messbare Funktion $f : \Omega \rightarrow \mathbb{C}$ heißt integrierbar, falls der Realteil $\Re f$ und der Imaginärteil $\Im f$ integrierbar sind (äquivalent, falls $|f|$ integrierbar ist). Man setzt dann

$$\int_{\Omega} f d\mu := \int_{\Omega} \Re f d\mu + i \int_{\Omega} \Im f d\mu \in \mathbb{C}.$$

Eigenschaften:

1. Komplex linear: Für alle integrierbaren $f, g : \Omega \rightarrow \mathbb{C}$, $a, b \in \mathbb{C}$ gilt $\int (af + bg) = a \int f + b \int g$.
2. $\Re \int f = \int (\Re f)$, $\Im \int f = \int (\Im f)$, $\int \bar{f} = \overline{\int f}$.
3. $|\int f| \leq \int |f|$.

Zu 3.: Wähle $\theta \in [0, 2\pi)$, so dass $|\int f| = e^{i\theta} \int f$. Dann gilt

$$\begin{aligned} \left| \int f \right| &= e^{i\theta} \int f = \Re \left(\int (e^{i\theta} f) \right) \\ &= \int \Re (e^{i\theta} f) \leq \int |f|. \end{aligned}$$

Definition 1.12. Sei μ ein W-Maß auf $(\mathbb{R}^d, \mathcal{B}^d)$. Dann heißt die Abbildung $\varphi_{\mu} : \mathbb{R}^d \rightarrow \mathbb{C}$ mit

$$\varphi_{\mu}(t) = \int_{\mathbb{R}^d} e^{i \langle t, x \rangle} d\mu(x), \quad (1.18)$$

wobei $\langle t, x \rangle = t^T x$ das Skalarprodukt auf \mathbb{R}^d ist, die **charakteristische Funktion** (oder **Fouriertransformation**) von μ .

Dabei existiert das Integral in (1.18) für alle $t \in \mathbb{R}^d$, da $|e^{i \langle t, x \rangle}| = 1$. Hat μ die Dichte f bzgl. des Lebesgue-Maßes λ^d , so folgt

$$\varphi_{\mu}(t) = \int_{\mathbb{R}^d} e^{i \langle t, x \rangle} f(x) d\lambda^d(x) =: \mathcal{F}(f)(t).$$

Obiges Integral ist allgemein für Lebesgue-integrierbares f definiert, und heißt **Fouriertransformation von f** .

Wir bemerken, dass die Fouriertransformation $\mathcal{F}(f)(t)$ eine lineare Abbildung auf dem Raum L_1 der integrierbaren Funktionen ist.

Ist $X : \Omega \rightarrow \mathbb{R}^d$ ein Zufallsvektor mit Verteilung P_X , so heißt

$$\varphi_X(t) := \varphi_{P_X}(t) = \int_{\mathbb{R}^d} e^{i \langle t, x \rangle} dP_X(x) = \mathbb{E}[e^{i \langle t, X \rangle}] \quad (1.19)$$

die **charakteristische Funktion von X** . Das letzte Gleichheitszeichen ergibt sich aus der Transformationsformel (1.8).

Rechenregeln

Sei $X : \Omega \rightarrow \mathbb{R}^d$ ein Zufallsvektor mit charakteristischer Funktion φ_X . Dann gilt

1. $\varphi_X(0) = 1, |\varphi_X(t)| \leq 1$.

2. $\varphi_X(-t) = \overline{\varphi_X(t)}$.

3. Für $A \in \mathbb{R}^{m \times d}$ und $b \in \mathbb{R}^m$ ist die charakteristische Funktion von $AX + b$ gegeben durch

$$\varphi_{AX+b}(s) = \mathbb{E}[e^{i\langle AX+b, s \rangle}] = e^{i\langle b, s \rangle} \mathbb{E}[e^{iX^T A^T s}] = e^{i\langle b, s \rangle} \varphi_X(A^T s). \quad (1.20)$$

4. Sind μ_1, \dots, μ_k W-Maße auf $(\mathbb{R}^d, \mathcal{B}^d)$ und $\lambda_1, \dots, \lambda_k \geq 0$, $\sum_j \lambda_j = 1$, so ist

$$\varphi_{\lambda_1 \mu_1 + \dots + \lambda_k \mu_k} = \lambda_1 \varphi_{\mu_1} + \dots + \lambda_k \varphi_{\mu_k}. \quad (1.21)$$

Beispiele 1.13. 1. Dirac-Maß: Zu δ_x dem Dirac-Maß bei x ist

$$\varphi_{\delta_x}(t) = e^{i\langle t, x \rangle}.$$

2. Poissonverteilung: Sei $X \sim \text{Poi}(\lambda)$, also $\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$, $k = 0, 1, 2, \dots$, so ist

$$\begin{aligned} \varphi_X(t) &= \mathbb{E}[e^{itX}] = \sum_{k \geq 0} e^{itk} e^{-\lambda} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k \geq 0} \frac{(e^{it}\lambda)^k}{k!} = e^{-\lambda} \exp(e^{it}\lambda) = \exp(\lambda(e^{it} - 1)). \end{aligned}$$

3. Uniforme Verteilung: Ist $X \sim U(-a, a)$, dann ist

$$\begin{aligned} \varphi_X(t) &= \frac{1}{2a} \int_{-a}^a e^{itx} dx = \frac{1}{2a} \frac{e^{itx}}{it} \Big|_{x=-a}^{x=a} \\ &= \frac{1}{at} \frac{e^{ita} - e^{-ita}}{2i} = \frac{\sin(at)}{at}. \end{aligned}$$

4. Standardnormalverteilung: Ist $X \sim \mathcal{N}(0, 1)$, so ist

$$\varphi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{itx} e^{-x^2/2} dx = e^{-t^2/2}.$$

Für die letzte Gleichung zeigt man, dass $\varphi'_X(t) = -t\varphi_X(t)$. Da $\varphi_X(0) = 1$, hat die lineare Differentialgleichung die angegebene eindeutige Lösung.

Ist $Y = \sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$, so folgt mit Rechenregel (1.20), dass $\varphi_Y(t) = e^{i\mu t} e^{-\sigma^2 t^2/2}$.

5. Exponential- und Laplace-Verteilung: Ist $X \sim \text{Exp}(1)$, also mit Dichte $f(x) = e^{-x} \mathbf{1}_{(0,\infty)}(x)$, so ist

$$\varphi_X(t) = \int_0^\infty e^{itx-x} dx = \frac{e^{x(it-1)}}{(it-1)} \Big|_{x=0}^{x=\infty} = \frac{1}{1-it}.$$

Die Laplace-Verteilung hat Dichte $e^{-|x|}/2 = (f(x) + f(-x))/2$ für obiges f , somit nach (1.21)

$$\varphi_X(t) = \frac{1}{2} \left(\frac{1}{1-it} + \frac{1}{1+it} \right) = \frac{1}{1+t^2}. \quad (1.22)$$

Analytische Eigenschaften

Satz 1.14. Sei μ ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}^d, \mathcal{B}^d)$ mit charakteristischer Funktion φ_μ . Dann gilt

1. φ_μ ist gleichmäßig stetig.

2. φ_μ ist positiv semidefinit, d.h. für alle $t_1, \dots, t_n \in \mathbb{R}^d$, $n \geq 1$, $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ gilt

$$\sum_{j,k=1}^n \lambda_j \bar{\lambda}_k \varphi_\mu(t_j - t_k) \geq 0.$$

3. (Riemann-Lebesgue-Lemma.) Hat das Wahrscheinlichkeitsmaß μ eine Lebesgue-Dichte, so gilt $\lim_{|t| \rightarrow \infty} \varphi_\mu(t) = 0$. ■

Beweis. Zu 1.: Für $h \in \mathbb{R}^d$ gilt

$$\begin{aligned} \sup_{t \in \mathbb{R}^d} |\varphi(t+h) - \varphi(t)| &= \sup_{t \in \mathbb{R}^d} \left| \int e^{i \langle t, x \rangle} (e^{i \langle h, x \rangle} - 1) d\mu(x) \right| \\ &\leq \sup_{t \in \mathbb{R}^d} \int |e^{i \langle t, x \rangle}| |e^{i \langle h, x \rangle} - 1| d\mu(x) \\ &= \int |e^{i \langle h, x \rangle} - 1| d\mu(x) \rightarrow 0, \quad h \rightarrow 0, \end{aligned}$$

nach dem Satz von der majorisierten Konvergenz, da $|e^{i \langle h, x \rangle} - 1| \leq 2$ und $|e^{i \langle h, x \rangle} - 1| \rightarrow 0$, $h \rightarrow 0$.

Zu 2.:

$$\sum_{j,k=1}^n \lambda_j \bar{\lambda}_k \varphi_\mu(t_j - t_k) = \int \sum_{j,k=1}^n \lambda_j \bar{\lambda}_k e^{i \langle t_j - t_k, x \rangle} d\mu(x) = \int \left| \sum_{j=1}^n \lambda_j e^{i \langle t_j, x \rangle} \right|^2 d\mu(x) \geq 0.$$

Zu 3.: Man zeigt die Aussage zunächst für spezielle Dichten, z.B. Treppenfunktionen $\sum_{k=1}^m \alpha_k \mathbf{1}_{(a_k, b_k]}$, $a_k, b_k \in \mathbb{R}^d$, durch direkte Rechnung, und dann für allgemeine Dichten durch L_1 -Approximation. ■

Man kann zeigen (Satz von Bochner, siehe z. Bsp. Satz 15.29 in Klenke (2008)), dass die Punkte 1. und 2. charakteristische Funktionen von Wahrscheinlichkeitsmaßen auf \mathbb{R}^d charakterisieren in dem Sinne: Jede Funktion $\varphi : \mathbb{R}^d \rightarrow \mathbb{C}$ mit den beiden Eigenschaften und $\varphi(0) = 1$ ist charakteristische Funktion eines Wahrscheinlichkeitsmaßes auf \mathbb{R}^d .

Charakteristische Funktionen und Unabhängigkeit

Satz 1.15. Sei $X = (X_1, \dots, X_n)^T$ ein Zufallsvektor. Dann sind äquivalent:

1. X_1, \dots, X_n sind unabhängig.
2. Für die charakteristischen Funktionen gilt

$$\varphi_X(t) = \varphi_{X_1}(t_1) \cdot \dots \cdot \varphi_{X_n}(t_n) \quad \forall t = (t_1, \dots, t_n)^T \in \mathbb{R}^n.$$

Beweis. 1. \Rightarrow 2.:

$$\begin{aligned} \varphi_X(t) &= \mathbb{E}[e^{i(t_1 X_1 + \dots + t_n X_n)}] = \mathbb{E}\left[\prod_{j=1}^n e^{it_j X_j}\right] \\ &= \prod_{j=1}^n \mathbb{E}[e^{it_j X_j}] = \prod_{j=1}^n \varphi_{X_j}(t_j). \end{aligned}$$

2. \Rightarrow 1.: Die obige Rechnung zeigt, dass die charakteristische Funktion der Produktverteilung das Produkt der charakteristischen Funktionen der einzelnen Verteilungen ist. Wegen des Eindeutigkeitssatzes, siehe Satz 1.18 unten, ist dadurch die Produktverteilung eindeutig bestimmt. ■

Satz 1.16. Sind X, Y unabhängige Zufallsvektoren, so ist für alle $t \in \mathbb{R}^d$

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t).$$

Somit ist die charakteristische Funktion der Faltung das Produkt der charakteristischen Funktionen.

■

Beweis. Es gilt

$$\mathbb{E}[e^{i(X+Y)^T t}] = \mathbb{E}[e^{iX^T t} e^{iY^T t}] = \mathbb{E}[e^{iX^T t}] \mathbb{E}[e^{iY^T t}].$$

■

Beispiel 1.17. Sind $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, so gilt

$$\varphi_{X+Y}(t) = e^{i\mu_1 t} e^{-\sigma_1^2 t^2/2} e^{i\mu_2 t} e^{-\sigma_2^2 t^2/2} = e^{i(\mu_1 + \mu_2)t} e^{-(\sigma_1^2 + \sigma_2^2)t^2/2},$$

also $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Der Eindeutigkeitssatz

Satz 1.18 (Eindeutigkeitssatz). Sind μ, ν Wahrscheinlichkeitsmaße auf \mathbb{R}^d , so stimmen ihre charakteristischen Funktionen genau dann überein, also $\varphi_\mu = \varphi_\nu$, wenn $\mu = \nu$. ■

Beweis. Ist $\mu = \nu$, so folgt offenbar $\varphi_\mu = \varphi_\nu$. Es bleibt die Rückrichtung zu zeigen.

Sei $f_\sigma(z) = (2\pi\sigma^2)^{-n/2} \exp(-(2\sigma^2)^{-1} \sum_{j=1}^n z_j^2)$ die Lebesgue-Dichte von $Z = (Z_1, \dots, Z_n)^\top$ mit $Z_j \sim \mathcal{N}(0, \sigma^2)$, $1 \leq j \leq n$, sowie

$$\varphi_Z(u) = \exp\left(\frac{-\sum_{j=1}^n u_j^2 \sigma^2}{2}\right)$$

die charakteristische Funktion von Z (hierfür benutzen wir die Hinrichtung von Satz 1.15 und die charakteristische Funktion der Normalverteilung). Es gilt also

$$\int_{\mathbb{R}^n} f_\sigma(z) e^{i\langle u, z \rangle} dz = \prod_{j=1}^n \exp(-u_j^2 \sigma^2/2) = \varphi_Z(u).$$

Weiterhin folgt

$$\begin{aligned} f_\sigma(v-w) &= (2\pi\sigma^2)^{-n/2} \varphi_Z\left(\frac{v-w}{\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} \int_{\mathbb{R}^n} f_\sigma(z) e^{i\langle z, \sigma^{-2}(v-w) \rangle} dz. \end{aligned}$$

Ist nun $X \sim \mu$ und $Y \sim \nu$ mit $\varphi_X = \varphi_Y$, dann gilt

$$\begin{aligned} \int f_\sigma(v-w) \mu(dv) &= \int (2\pi\sigma^2)^{-n/2} \left(\int f_\sigma(z) e^{i\langle z, \sigma^{-2}(v-w) \rangle} dz \right) \mu(dv) \\ &= \int (2\pi\sigma^2)^{-n/2} f_\sigma(z) \varphi_X(z/\sigma^2) e^{-i\langle \sigma^{-2}z, w \rangle} dz \end{aligned}$$

und analog für ν , womit sich aus $\varphi_X = \varphi_Y$ ergibt, dass

$$\int g(x) \mu(dx) = \int g(x) \nu(dx), \quad (1.23)$$

für alle Funktionen g der Form $v \mapsto f_\sigma(v - w)$. Nach dem **Satz von Stone-Weierstrass** aus der Analysis (siehe z. Bsp. Klenke (2008) S. 302) liegt der Raum solcher Funktionen dicht in C_0 bezüglich gleichmäßiger Konvergenz. C_0 bezeichnet den Raum stetiger Funktionen mit $\lim_{\|z\| \rightarrow \infty} |f(z)| = 0$. Damit überträgt sich (1.23) auf alle Funktionen $g \in C_0$. Weiter können Indikatorfunktionen von offenen Mengen monoton durch C_0 -Funktionen approximiert werden, so dass $\mu(A) = v(A)$, für alle offenen Teilmengen A des \mathbb{R}^n , mit monotoner Konvergenz folgt. Nach dem Maßeindeutigkeitssatz (Übereinstimmung auf \sqcap -stabilem Erzeuger) erhalten wir $\mu = v$. ■

Aus der Inversionsformel der Fouriertransformation leitet sich die folgende Inversionsformel für charakteristische Funktionen ab.

Satz 1.19 (Inversion). Ist μ ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}^d, \mathcal{B}^d)$ mit charakteristischer Funktion φ_μ und gilt

$$\int_{\mathbb{R}^d} |\varphi_\mu(t)| dt < \infty,$$

so hat μ eine stetige Lebesgue-Dichte bzgl. λ^d , die durch

$$f(y) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\langle y, t \rangle} \varphi_\mu(t) dt$$

gegeben ist. Es gilt: $f \in L_2(\lambda^d)$ genau dann wenn $\varphi \in L_2(\lambda^d)$ und $\|f\|_2 = \|\varphi\|_2$ (Plancherel). ■

Unterschiedliche Beweise finden sich in Werner (2011), Kapitel V.2, in Feller (1968), Kapitel XV.3, sowie in Durret (2010), Theorem 3.3.5.

Beispiel 1.20. (Cauchy-Verteilung).

Nach (1.22) ist die charakteristische Funktion der Laplace-Verteilung, also der Dichte $f(x) = e^{-|x|}/2$, gegeben durch $\varphi(t) = (\mathcal{F}f)(t) = 1/(1+t^2)$.

Die Dichte g der Cauchy-Verteilung ist $g(x) = \varphi(x)/\pi$, und daher ergibt sich aus obiger Inversionsformel für die charakteristische Funktion der Cauchy-Verteilung

$$\mathcal{F}(g)(t) = \frac{1}{\pi} \mathcal{F}(\varphi)(t) = \frac{1}{\pi} \mathcal{F}(\mathcal{F}f)(t) = 2f(-t) = e^{-|t|}.$$

Sind X_1, \dots, X_n unabhängig und Cauchy-verteilt und ist $Z = (X_1 + \dots + X_n)/n$, so gilt

$$\varphi_Z(t) = e^{-|t|/n} \cdot \dots \cdot e^{-|t|/n} = e^{-|t|},$$

also ist auch Z Cauchy-verteilt.

Dimensionsreduktion

Satz 1.21. Sei $X : \Omega \rightarrow \mathbb{R}^d$ ein Zufallsvektor mit Verteilung P_X . Dann ist P_X eindeutig durch die Verteilungen $P_{a^T X}$ aller Linearkombinationen $a^T X$ für $a \in \mathbb{R}^d$ bestimmt. ■

Beweis. Es ist

$$\varphi_X(t) = \mathbb{E}[e^{it^T X}] = \varphi_{t^T X}(1),$$

somit folgt die Behauptung aus dem Eindeutigkeitssatz in Satz 1.18. ■

Korollar 1.22. Ein Wahrscheinlichkeitsmaß μ auf $(\mathbb{R}^d, \mathcal{B}^d)$ ist durch Auswertung auf allen Halbräumen $H_{\mathbf{a}, t} = \{x \in \mathbb{R}^d : \langle x, \mathbf{a} \rangle \leq t\}$, $\mathbf{a} \in \mathbb{R}^d, t \in \mathbb{R}$, also die Werte $\mu(H_{\mathbf{a}, t})$, eindeutig bestimmt.

Beweis. Ist X Zufallsvektor mit Verteilung μ und $\mathbf{a} \in \mathbb{R}^d$, so gilt für $t \in \mathbb{R}$

$$F_{\mathbf{a}^T X}(t) = \mathbb{P}(\mathbf{a}^T X \leq t) = \mu(H_{\mathbf{a},t}).$$

Die Verteilungsfunktion $F_{\mathbf{a}^T X}$ bestimmt eindeutig die Verteilung von $\mathbf{a}^T X$, und die Verteilungen aller $\mathbf{a}^T X$ nach obigem Satz die Verteilung von X selbst. ■

Charakteristische Funktion und Momente

Satz 1.23. Es sei $X \sim F$ eine reellwertige Zufallsvariable mit Verteilungsfunktion F und mit

$$\mathbb{E}[|X|^n] = \int_{-\infty}^{\infty} |x|^n dF(x) < \infty, n \geq 1.$$

Dann existieren n stetige Ableitungen der charakteristischen Funktion $\varphi = \mathbb{E}[\exp(itX)]$ und

$$\varphi^{(k)}(t) = i^k \int_{-\infty}^{\infty} x^k e^{itx} dF(x), k = 1, \dots, n.$$

Insbesondere gilt $\mathbb{E}[X^k] = i^{-k} \varphi^{(k)}(0)$. ■

Beweis. Betrachte die Differenzenquotienten

$$\frac{\varphi(t+h) - \varphi(t)}{h} = \int_{-\infty}^{\infty} e^{ithx} (e^{ihx} - 1) h^{-1} dF(x).$$

Da für $h \rightarrow 0$ gilt, dass $h^{-1}(e^{ihx} - 1) \rightarrow ix$ und $h^{-1}|e^{ihx} - 1| \leq |x|$, wobei $\int_{-\infty}^{\infty} |x| dF(x) < \infty$ nach Voraussetzung, folgt mit majorisierter Konvergenz

$$\varphi'(t) = i \int_{-\infty}^{\infty} e^{itx} x \mu(dx),$$

wenn μ das zu F gehörende Maß (Lebesgue-Stieltjes-Maß) ist. Dies ist obige Identität für $k = 1$ und für $k > 1$ führt man eine Induktion. ■

Eine mehrdimensionale Formulierung des Satzes findet man als Theorem 13.2 bei [Jacod and Protter \(2000\)](#).

2 Mehrdimensionale Verteilungen

2.1 Diskrete gemeinsame und marginale Verteilungen

Beispiel 2.1. Beim zweifachen Wurf eines fairen Würfels (uniforme Verteilung auf dem Ergebnisraum $\Omega = \{1, \dots, 6\}^2$) definieren wir die Zufallsvariablen

$$\begin{aligned} X : \Omega &\rightarrow \{1, \dots, 6\}, \quad X((\omega_1, \omega_2)) = \omega_1 & (\text{Augenzahl des ersten Würfels}), \\ Y : \Omega &\rightarrow \{1, \dots, 6\}, \quad Y((\omega_1, \omega_2)) = \max(\omega_1, \omega_2) & (\text{größere Zahl}). \end{aligned}$$

Jetzt besteht die gemeinsame Verteilung von X, Y aus den Wahrscheinlichkeiten

$$\mathbb{P}(\{X = i\} \cap \{Y = j\}), \quad i, j = 1, \dots, 6$$

mit den Werten aus [Tabelle 2.1](#).

Tabelle 2.1: Gemeinsame Verteilung des Maximums und der ersten Augenzahl beim zweifachen Würfeln

$X \setminus Y$	1	2	3	4	5	6	Σ
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
2	0	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
3	0	0	$\frac{3}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
4	0	0	0	$\frac{4}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
5	0	0	0	0	$\frac{5}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
6	0	0	0	0	0	$\frac{6}{36}$	$\frac{1}{6}$
Σ	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	1

Definition 2.2. Sind $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum und $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$ diskrete Zufallsvariablen, so heißt die \mathbb{R}^n -wertige Zufallsvariable

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} : \Omega \rightarrow \mathbb{R}^n, \quad \mathbf{X}(\omega) = \begin{pmatrix} X_1(\omega) \\ \vdots \\ X_n(\omega) \end{pmatrix}$$

ein (\mathbb{R}^n -wertiger), diskreter **Zufallsvektor**.

Definition 2.3. Sind $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$ Zufallsvariablen und $\mathbf{X} = (X_1, \dots, X_n)^T$, so heißt

$$\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{X_1, \dots, X_n} : \mathcal{P}(\mathbf{X}(\Omega)) \rightarrow [0, 1], \quad \mathbb{P}_{\mathbf{X}}(A) = \mathbb{P}(\mathbf{X} \in A), \quad A \subseteq \mathbf{X}(\Omega) \subseteq \mathbb{R}^n$$

die **gemeinsame Verteilung** von X_1, \dots, X_n .

Bemerkung. 1. Die gemeinsame Verteilung von diskreten Zufallsvariablen wird bereits bestimmt durch die zugehörige **Wahrscheinlichkeitsfunktion** (auch **Zähldichte**)

$$p_{\mathbf{X}} : \mathbf{X}(\Omega) \rightarrow [0, 1], \quad p_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x}) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n), \quad \mathbf{x} \in \mathbf{X}(\Omega) \subseteq \mathbb{R}^n.$$

2. Es ist $\mathbf{X}(\Omega) \subseteq X_1(\Omega) \times \cdots \times X_n(\Omega)$. Für

$$x_i \in X_i(\Omega) \text{ für ein } i \in \{1, \dots, n\}, \quad \mathbf{x} = (x_1, \dots, x_n)^T \notin \mathbf{X}(\Omega)$$

setzen wir $p_{\mathbf{X}}(\mathbf{x}) = 0$.

Definition 2.4. Sind $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$ Zufallsvariablen und ist $1 \leq i_1 < \dots < i_k \leq n$, so heißt die gemeinsame Verteilung von X_{i_1}, \dots, X_{i_k} die **marginale Verteilung** von X_1, \dots, X_n .

Bemerkung. Ist $A_{i_j} \subseteq X_{i_j}(\Omega)$, $j = 1, \dots, k$, und $A_j = X_j(\Omega)$, für $j \in \{1, \dots, n\} \setminus \{i_1, \dots, i_k\}$, dann gilt

$$\mathbb{P}_{X_{i_1}, \dots, X_{i_k}}(A_{i_1} \times \cdots \times A_{i_k}) = \mathbb{P}_{X_1, \dots, X_n}(A_1 \times \cdots \times A_n).$$

Insbesondere gilt für $x_{i_l} \in X_{i_l}(\Omega)$, mit $\{j_1, \dots, j_{n-k}\} := \{1, \dots, n\} \setminus \{i_1, \dots, i_k\}$, dass

$$p_{X_{i_1}, \dots, X_{i_k}}(x_{i_1}, \dots, x_{i_k}) = \sum_{x_{j_1} \in X_{j_1}(\Omega)} \cdots \sum_{x_{j_{n-k}} \in X_{j_{n-k}}(\Omega)} p_{\mathbf{X}}(x_1, \dots, x_n).$$

Bemerkung. Sind $X, Y : \Omega \rightarrow \mathbb{R}$ Zufallsvariablen mit endlichen Bildbereichen, x_1, \dots, x_n eine Abzählung von $X(\Omega)$ und y_1, \dots, y_m eine Abzählung von $Y(\Omega)$, dann lässt sich die gemeinsame Verteilung von X, Y mit einer 2D-Kontingenztafel nach dem Schema in Tabelle 2.2 grafisch darstellen, siehe als Beispiel Tabelle 2.1. Dabei sei

Tabelle 2.2: Aufbau einer 2D-Kontingenztafel

		Y	y_1	\cdots	y_m	Σ
X			p_{x_1, y_1}	\cdots	p_{x_1, y_m}	
x_1						$p_x(x_1)$
\vdots			\vdots	\ddots	\vdots	\vdots
x_n			p_{x_n, y_1}	\cdots	p_{x_n, y_m}	$p_x(x_n)$
Σ			$p_y(y_1)$	\cdots	$p_y(y_m)$	1

$$p_{x_i, y_j} = \mathbb{P}_{X, Y}(x_i, y_j) = \mathbb{P}(X = x_i, Y = y_j), \quad p_x(x_i) = \mathbb{P}(X = x_i), \quad p_y(y_j) = \mathbb{P}(Y = y_j).$$

Bemerkung. Seien $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$ diskrete Zufallsvariablen und $h : \mathbb{R}^n \rightarrow \mathbb{R}^k$ eine Funktion. Definieren wir nun die Zufallsvektoren

$$\mathbf{X} : \Omega \rightarrow \mathbb{R}^n, \quad \mathbf{X} = (X_1, \dots, X_n)^T, \quad \mathbf{Y} : \Omega \rightarrow \mathbb{R}^k, \quad \mathbf{Y} = h(\mathbf{X}),$$

so ist für $\mathbf{y} = (y_1, \dots, y_k)^T \in \mathbf{Y}(\Omega)$

$$\begin{aligned} \{\mathbf{Y} = \mathbf{y}\} &= \{\omega \in \Omega \mid h(\mathbf{X}(\omega)) = \mathbf{y}\} = \{\omega \in \Omega \mid \mathbf{X}(\omega) = \mathbf{x} \text{ und } h(\mathbf{x}) = \mathbf{y}\} \\ &= \bigcup_{\substack{\mathbf{x} \in \mathbf{X}(\Omega) \\ \text{mit } h(\mathbf{x}) = \mathbf{y}}} \{\mathbf{x} = \mathbf{x}\}, \\ p_{\mathbf{Y}}(\mathbf{y}) &= \mathbb{P}(\mathbf{Y} = \mathbf{y}) = \mathbb{P}\left(\bigcup_{\substack{\mathbf{x} \in \mathbf{X}(\Omega) \\ \text{mit } h(\mathbf{x}) = \mathbf{y}}} \{\mathbf{x} = \mathbf{x}\}\right) = \sum_{\substack{\mathbf{x} \in \mathbf{X}(\Omega) \\ \text{mit } h(\mathbf{x}) = \mathbf{y}}} p_{\mathbf{X}}(\mathbf{x}). \end{aligned}$$

Ist $h : \mathbb{R}^n \rightarrow \mathbb{R}$ und $Y = h(\mathbf{X})$ mit $\mathbb{E}[|Y|] < \infty$, dann ist

$$\mathbb{E}[Y] = \sum_{\mathbf{x} \in \mathbf{X}(\Omega)} h(\mathbf{x}) \cdot p_{\mathbf{X}}(\mathbf{x}).$$

Beispiel 2.5. Sind X, Y die Zufallsvariablen, die im zweifachen Würfelwurf die Einzelergebnisse darstellen, so gilt

$$\mathbb{E}[X \cdot Y] = \sum_{i=1}^6 \sum_{j=1}^6 i \cdot j \cdot \mathbb{P}(X = i, Y = j) = \frac{1}{36} \cdot \sum_{i=1}^6 i \cdot \sum_{j=1}^6 j = \frac{49}{4}.$$

2.2 Zufallsvektoren und absolutstetige Verteilungen

Wir bezeichnen mit \mathcal{B}_n die **Borel- σ -Algebra** über \mathbb{R}^n , also die kleinste σ -Algebra $\mathcal{B}_n \subset \mathcal{P}(\mathbb{R}^n)$, die alle Quader

$$(\mathbf{a}, \mathbf{b}] := \{\mathbf{x} \in \mathbb{R}^n \mid a_i < x_i \leq b_i \text{ für alle } i = 1, \dots, n\}, \quad \mathbf{a}, \mathbf{b} \in \mathbb{R}^n$$

enthält.

Definition 2.6. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Dann heißt $\mathbf{X} : \Omega \rightarrow \mathbb{R}^n$ ein **Zufallsvektor**, falls

$$\forall B \in \mathcal{B}_n : \{\mathbf{X} \in B\} = \{\omega \in \Omega \mid \mathbf{X}(\omega) \in B\} \in \mathcal{A}. \quad (2.1)$$

Sei $\mathbf{X} : \Omega \rightarrow \mathbb{R}^n$ eine Abbildung mit $\mathbf{X} = (X_1, \dots, X_n)^T$. Nach Resultaten zur Messbarkeit aus der Stochastik I sind äquivalent:

1. \mathbf{X} ist ein Zufallsvektor, erfüllt also (2.1).
2. X_1, \dots, X_n sind Zufallsvariablen, also gilt

$$\forall B \in \mathcal{B} = \mathcal{B}_1 : \{X_i \in B\} \in \mathcal{A}, \quad i = 1, \dots, n.$$

Definition 2.7. Ist $\mathbf{X} : \Omega \rightarrow \mathbb{R}^n$ ein Zufallsvektor, so heißt die **Verteilung von \mathbf{X}**

$$\mathbb{P}_{\mathbf{X}} : \mathcal{B}_n \rightarrow [0, 1], \quad B \mapsto \mathbb{P}_{\mathbf{X}}(B) = \mathbb{P}(\mathbf{X} \in B)$$

die **gemeinsame Verteilung** von X_1, \dots, X_n . Ein Zufallsvektor \mathbf{X} heißt **absolutstetig verteilt mit Dichte** $f : \mathbb{R}^n \rightarrow [0, \infty)$, falls

$$\mathbb{P}_{\mathbf{X}}(B) = \mathbb{P}(\mathbf{X} \in B) = \int_B f(\mathbf{x}) d\mathbf{x}, \quad B \in \mathcal{B}_n. \quad (2.2)$$

Definition 2.8. Für $d \in \mathbb{N}$ heißt die durch die Dichtefunktion

$$f(\mathbf{x}) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{\mathbf{x}^T \mathbf{x}}{2}\right), \quad \mathbf{x} \in \mathbb{R}^d, \quad \mathbf{x}^T \mathbf{x} = (x_1, \dots, x_d)^T \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} = \sum_{i=1}^d x_i^2$$

charakterisierte **Verteilung d -variate Standard-Normalverteilung**.

Bemerkung. Es gilt tatsächlich $\int_{\mathbb{R}^d} f(\mathbf{x}) d\mathbf{x} = 1$, da sich das Integral über das Produkt faktorweise auswerten lässt.

Definition 2.9. Ist $\mathbf{X} = (X_1, \dots, X_n)^\top : \Omega \rightarrow \mathbb{R}^n$ ein Zufallsvektor und $1 \leq i_1 < \dots < i_k \leq n$, so heißt die gemeinsame Verteilung von X_{i_1}, \dots, X_{i_k} deren **marginale Verteilung**.

Ist der Zufallsvektor \mathbf{X} absolutstetig mit Dichte f , so ist auch $(X_{i_1}, \dots, X_{i_k})^\top$ absolutstetig mit der Dichte

$$h(x_{i_1}, \dots, x_{i_k}) = \underbrace{\int_{\mathbb{R}} \cdots \int_{\mathbb{R}}}_{n-k} f(x_1, \dots, x_n) dx_{j_1} \cdots dx_{j_{n-k}}, \quad \{j_1, \dots, j_{n-k}\} = \{1, \dots, n\} \setminus \{i_1, \dots, i_k\}.$$

Beispiel 2.10. Sei $(X, Y)^\top$ uniform verteilt auf der Kreisscheibe

$$B = \{(x, y)^\top \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}.$$

Dann hat X die Dichte

$$h(x) = \frac{1}{\pi} \int_{\mathbb{R}} 1_{\{x^2 + y^2 \leq 1\}} dy = \begin{cases} \frac{2}{\pi} \sqrt{1 - x^2}, & \text{wenn } x^2 \leq 1 \\ 0, & \text{wenn } x^2 > 1 \end{cases}.$$

Bevor wir multivariate Verallgemeinerungen einführen, erinnern wir an den Begriff der **Kovarianz** aus der Stochastik I. Sind $X, Y : \Omega \rightarrow \mathbb{R}$ integrierbare Zufallsvariablen derart, dass auch XY integrierbar, so heißt

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

die Kovarianz von X und Y . Für $X, Y \in L_2(\Omega, \mathbb{P})$, d.h. $\mathbb{E}[X^2] < \infty$ und $\mathbb{E}[Y^2] < \infty$, existiert $\text{Cov}(X, Y)$. Sind reellwertige Zufallsvariablen $X, Y : \Omega \rightarrow \mathbb{R}$, mit $\mathbb{E}[X^2] < \infty$ und $\mathbb{E}[Y^2] < \infty$, absolutstetig mit gemeinsamer Dichte $f_{X,Y}$, so ist

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \iint_{\mathbb{R}^2} (x - \mathbb{E}[X])(y - \mathbb{E}[Y]) f_{X,Y}(x, y) dx dy.$$

Wir wiederholen die wichtigsten Rechenregeln für Kovarianzen. Es ist $\text{Cov}(X, Y) = \text{Cov}(Y, X)$, und $\text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y)$, $a, b, c, d \in \mathbb{R}$. Sind $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$, so sind X, Y und XY integrierbar und es gelten

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y).$$

Sind allgemeiner $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$ mit $\mathbb{E}[X_i^2] < \infty$, so gilt die **Varianz-Kovarianz-Formel**:

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j).$$

Definition 2.11. Ist $\mathbf{X} = (X_1, \dots, X_n)^\top : \Omega \rightarrow \mathbb{R}^n$ ein Zufallsvektor, so heißen

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix} \in \mathbb{R}^n, \quad \text{Cov}(\mathbf{X}) = (\text{Cov}(X_i, X_j))_{i,j=1,\dots,n} \in \mathbb{R}^{n \times n}$$

Erwartungswertvektor und Kovarianzmatrix von \mathbf{X} .

Es gilt die **Linearität**

$$\mathbb{E}[AX + b] = A\mathbb{E}[X] + b,$$

mit einer Matrix A und einem Vektor b . Für $\mathbf{A} \in \mathbb{R}^{m \times n}$ gilt $\text{Cov}(\mathbf{AX}) = \mathbf{A} \text{Cov}(\mathbf{X}) \mathbf{A}^T$. Für einen (Spalten)vektor $\mathbf{a} \in \mathbb{R}^n$ ist also insbesondere $\text{Cov}(\mathbf{a}^T \mathbf{X}) = \mathbf{a}^T \text{Cov}(\mathbf{X}) \mathbf{a}$. Eine **Kovarianzmatrix** $\text{Cov}(X)$ ist stets **symmetrisch und positiv semidefinit**: Für $a \in \mathbb{R}^d$ ist

$$a^T \text{Cov}(X) a = \text{Var}(a^T X) \geq 0,$$

und $\text{Cov}(X)$ ist genau dann degeneriert, falls ein $a \in \mathbb{R}^d$ existiert, so dass $a^T X = \text{const}$ fast sicher, äquivalent zu $\text{Var}(a^T X) = 0$.

Definition 2.12. Sind $\mathbf{X} = (X_1, \dots, X_n)^T : \Omega \rightarrow \mathbb{R}^n$ und $\mathbf{Y} = (Y_1, \dots, Y_m)^T : \Omega \rightarrow \mathbb{R}^m$ Zufallsvektoren, so heißt

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = (\text{Cov}(X_i, Y_j))_{\substack{i=1, \dots, n \\ j=1, \dots, m}} \in \mathbb{R}^{n \times m}$$

die **Kovarianzmatrix** von \mathbf{X} und \mathbf{Y} .

Sind \mathbf{X} und \mathbf{Y} unabhängig, dann sind für $i = 1, \dots, n$, und $j = 1, \dots, m$, auch deren Komponenten X_i und Y_j unabhängig, also gilt $\text{Cov}(X_i, Y_j) = 0$. Daraus folgt $\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{0}$. Sind \mathbf{X} und \mathbf{Y} Zufallsvektoren, so gilt für den Zufallsvektor $(\mathbf{X}^T, \mathbf{Y}^T)^T$

$$\text{Cov}(\mathbf{Z}) = \begin{pmatrix} \text{Cov}(\mathbf{X}) & \text{Cov}(\mathbf{X}, \mathbf{Y}) \\ \text{Cov}(\mathbf{Y}, \mathbf{X}) & \text{Cov}(\mathbf{Y}) \end{pmatrix}, \quad \text{Cov}(\mathbf{X}, \mathbf{Y}) = \text{Cov}(\mathbf{Y}, \mathbf{X})^T.$$

2.3 Multinomialverteilung

Modell: Verteile n Personen auf s Gruppen. Der Ergebnisraum ist hier $\Omega = \{1, \dots, s\}^n$, wobei in jedem $\omega = (\omega_1, \dots, \omega_n) \in \Omega$ jeweils ω_i die Gruppenzugehörigkeit von Person $i \in \{1, \dots, n\}$ ist. Dann können wir die Ereignisse

$$E_{i_1, \dots, i_s} = \{\omega \in \Omega \mid \forall j \in \{1, \dots, s\} : \text{card}\{k \in \{1, \dots, n\} \mid \omega_k = j\} = i_j\}$$

für $i_1, \dots, i_s \in \{0, \dots, n\}$ mit $i_1 + \dots + i_s = n$ definieren, die darstellen, dass in den einzelnen Gruppen $j \in \{1, \dots, s\}$ jeweils i_j Personen sind. Verteilen wir nun für feste Mitgliederzahlen die Gruppen auf die Personen, so erhalten wir

$$\begin{aligned} \text{card } E_{i_1, \dots, i_s} &= \binom{n}{i_1} \cdot \binom{n-i_1}{i_2} \cdot \dots \cdot \binom{n-i_1-\dots-i_{s-1}}{i_s} \\ &= \frac{n!}{i_1!(n-i_1)!} \cdot \frac{(n-i_1)!}{i_2!(n-i_1-i_2)!} \cdot \dots \cdot \frac{(n-i_1-\dots-i_{s-1})!}{i_s!0!} = \frac{n!}{i_1! \cdot \dots \cdot i_s!}. \end{aligned}$$

Ist $n \in \mathbb{N}$ und sind $i_1, \dots, i_s \in \mathbb{N}_0$ mit $i_1 + \dots + i_s = n$, so heißt

$$\binom{n}{i_1, \dots, i_s} := \frac{n!}{i_1! \cdot \dots \cdot i_s!}$$

Multinomialkoeffizient von n über i_1, \dots, i_s .

Bemerkung. Der Binomialkoeffizient ist ein Multinomialkoeffizient mit $s = 2$, mit der Notation $\binom{n}{k} = \binom{n}{k, n-k}$.

Angenommen, der Ausgang $j \in \{1, \dots, s\}$ habe die Wahrscheinlichkeit $p(j) = p_j$, also $p_1 + \dots + p_s = 1$. Auf Ω definieren wir dann eine Produktverteilung, also

$$p^{\otimes n}(\omega) = \prod_{j=1}^s p_j^{\sum_{k=1}^n 1_{\{\omega_k=j\}}}.$$

Ist nun $\omega \in E_{i_1, \dots, i_s}$, so ist $p^{\otimes n}(\omega) = p_1^{i_1} \cdots p_s^{i_s}$, also

$$\mathbb{P}^{\otimes n}(E_{i_1, \dots, i_s}) = \binom{n}{i_1, \dots, i_s} \cdot p_1^{i_1} \cdots p_s^{i_s}.$$

Wir definieren nun die Ereignisse

$$A_k^{(j)} = \{\omega = (\omega_1, \dots, \omega_n) \in \Omega \mid \omega_k = j\},$$

die beschreiben, dass Person $k \in \{1, \dots, n\}$ in Gruppe $j \in \{1, \dots, s\}$ ist, und die Zufallsvariablen

$$X_j : \Omega \rightarrow \{0, \dots, n\}, \quad X_j(\omega) = \sum_{k=1}^n 1_{A_k^{(j)}}(\omega),$$

die jeweils die Anzahl der Personen in der Gruppe $j \in \{1, \dots, s\}$ beschreiben. Dann gilt

$$\{X_1 = i_1, \dots, X_s = i_s\} = E_{i_1, \dots, i_s}$$

und die gemeinsame Verteilung ist definiert durch

$$\mathbb{P}(X_1 = i_1, \dots, X_s = i_s) = \binom{n}{i_1, \dots, i_s} \cdot p_1^{i_1} \cdots p_s^{i_s}.$$

Definition 2.13. Seien $n, s \in \mathbb{N}$ und $p_1, \dots, p_s \in [0, 1]$ mit $\sum_{i=1}^s p_i = 1$. Dann ist ein Zufallsvektor $\mathbf{X} = (X_1, \dots, X_s)^T$ **multinomialverteilt** mit Parametern n und p_1, \dots, p_s , falls

$$\mathbb{P}(X_1 = i_1, \dots, X_s = i_s) = \binom{n}{i_1, \dots, i_s} \cdot p_1^{i_1} \cdots p_s^{i_s} \quad \text{für } i_1, \dots, i_s \in \mathbb{N}_0 \text{ mit } i_1 + \cdots + i_s = n.$$

Man schreibt dann auch kurz $\mathbf{X} \sim \text{Mult}(n; p_1, \dots, p_s)$.

Bemerkung. 1. $\{X_j = i_j\}$ beschreibt das Ereignis, in n unabhängigen Experimenten mit jeweils s Ausgängen der Wahrscheinlichkeiten p_1, \dots, p_s den Ausgang j genau i_j -mal zu erhalten.

2. Es ist $X_s = n - X_1 - \cdots - X_{s-1}$, daher betrachtet man manchmal nur $(X_1, \dots, X_{s-1})^T$.

3. Die Binomialverteilung tritt als Spezialfall der Multinomialverteilung für $s = 2$ auf, da

$$X \sim \text{Bin}(n, p) \Leftrightarrow (X, n - X)^T \sim \text{Mult}(n; p, 1 - p).$$

Ist $(X_1, \dots, X_s)^T \sim \text{Mult}(n; p_1, \dots, p_s)$, so folgt:

1. $X_j \sim \text{Bin}(n, p_j)$ für $j = 1, \dots, s$,

2. Ist T_1, \dots, T_l eine disjunkte Zerlegung von $\{1, \dots, s\}$ und

$$Y_r = \sum_{j \in T_r} X_j, \quad q_r = \sum_{j \in T_r} p_j, \quad r = 1, \dots, l,$$

dann folgt $(Y_1, \dots, Y_l)^T \sim \text{Mult}(n; q_1, \dots, q_l)$.

Bemerkung. 1. ist ein Spezialfall von 2. für $T_1 = \{j\}$ und $T_2 = \{1, \dots, s\} \setminus \{j\}$. 2. bleibt hier ohne Beweis, ist aber nach Anschauung klar, da man die Ausgänge aus jedem T_i zu je einem Ausgang zusammenfassen kann.

Korrelationen im Multinomialmodell: Sei $(X_1, \dots, X_s)^T \sim \text{Mult}(n; p_1, \dots, p_s)$ und $1 \leq i < j \leq s$. Wegen $X_i \sim \text{Bin}(n, p_i)$ und $X_j \sim \text{Bin}(n, p_j)$ gilt dann $\mathbb{E}[X_i] = np_i$ und $\mathbb{E}[X_j] = np_j$. Wir nutzen nun die Notation

des vorangegangenen Abschnitts, und setzen damit

$$X_i = \sum_{k=1}^n 1_{A_k^{(i)}}, \quad X_j = \sum_{k=1}^n 1_{A_k^{(j)}}.$$

Dann gilt

$$\mathbb{E}[X_i \cdot X_j] = \sum_{k=1}^n \sum_{l=1}^n \mathbb{E}[1_{A_k^{(i)}} 1_{A_l^{(j)}}] = \sum_{k=1}^n \sum_{l=1}^n \mathbb{E}[1_{A_k^{(i)} \cap A_l^{(j)}}] = \sum_{k=1}^n \sum_{l=1}^n \mathbb{P}(A_k^{(i)} \cap A_l^{(j)}).$$

Für $k = l$ ist $A_k^{(i)} \cap A_l^{(j)} = \emptyset$, für $k \neq l$ sind die Ereignisse unabhängig, also gilt

$$\mathbb{P}(A_k^{(i)} \cap A_l^{(j)}) = \mathbb{P}(A_k^{(i)}) \cdot \mathbb{P}(A_l^{(j)}) = p_i p_j$$

und somit

$$\mathbb{E}[X_i \cdot X_j] = \sum_{k=1}^n \sum_{l=1}^n \mathbb{P}(A_k^{(i)} \cap A_l^{(j)}) = \sum_{\substack{k=1 \\ k \neq l}}^n \sum_{l=1}^n p_i p_j = n(n-1) \cdot p_i p_j,$$

also

$$\text{Cov}(X_i, X_j) = \mathbb{E}[X_i \cdot X_j] - \mathbb{E}[X_i] \cdot \mathbb{E}[X_j] = n(n-1) \cdot p_i p_j - np_i \cdot np_j = -np_i p_j.$$

Wir erinnern an die Definition des **Korrelationskoeffizienten** zwischen Zufallsvariablen X und Y :

Sind $\mathbb{E}[X^2] < \infty$ und $\mathbb{E}[Y^2] < \infty$, sowie $\text{Var}(X) > 0$ und $\text{Var}(Y) > 0$, so ist die Korrelation zwischen X und Y , $\rho(X, Y)$, definiert durch

$$\rho(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma(X) \cdot \sigma(Y)}.$$

Für $(X_1, \dots, X_s)^T \sim \text{Mult}(n; p_1, \dots, p_s)$ und $1 \leq i < j \leq s$, gilt damit

$$\rho(X_i, X_j) = \frac{-np_i p_j}{n\sqrt{p_i(1-p_i) \cdot p_j(1-p_j)}} = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}}.$$

2.4 Multivariate Normalverteilung

Ausgangspunkt ist die **d -variate Standardnormalverteilung**. Seien X_1, \dots, X_d u.i.v. Zufallsvariablen mit Verteilung $\mathcal{N}(0, 1)$, also mit der Lebesgue-Dichte $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Die Verteilung des Vektors $\mathbf{X} = (X_1, \dots, X_d)^T$ ist dann die **multivariate oder d -variate Standardnormalverteilung**, $\mathcal{N}(0, I_d)$. Diese hat als Dichte das Produkt der Marginaldichten, also

$$f(\mathbf{x}) = \prod_{k=1}^d \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} = \frac{1}{(2\pi)^{d/2}} e^{-\sum_{k=1}^d x_k^2/2} = \frac{1}{(2\pi)^{d/2}} e^{-\mathbf{x}^T \mathbf{x}/2}, \quad \mathbf{x} = (x_1, \dots, x_d)^T,$$

vergleiche Definition 2.8. Als charakteristische Funktion ergibt das Produkt der charakteristischen Funktionen

$$\varphi_{\mathbf{X}}(\mathbf{t}) = \prod_{k=1}^d e^{-t_k^2/2} = e^{-\mathbf{t}^T \mathbf{t}/2}, \quad \mathbf{t} = (t_1, \dots, t_d)^T.$$

Definition 2.14. Ist $\mu \in \mathbb{R}^d$ und $\Sigma \in \mathbb{R}^{d \times d}$ symmetrisch und positiv definit, dann heißt die von der

Dichte

$$f(\mathbf{x}; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right) \quad (2.3)$$

definierte Verteilung die **multivariate bzw. d-variate Normalverteilung** mit Parametern μ und Σ . Ist ein Zufallsvektor \mathbf{X} so verteilt, dann schreiben wir $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$.

Bemerkung. Die d -variate Standard-Normalverteilung ist also gerade die Verteilung $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

Wir zeigen zunächst, dass (2.3) tatsächlich eine Dichte definiert. Aus der Analysis ist der folgende Satz bekannt, der allgemeiner in der Maß- und Integrationstheorie behandelt wird.

Satz 2.15 (Transformationsformel – multivariate Substitutionsregel). Sei $U \subseteq \mathbb{R}^n$ offen, $T : U \rightarrow \mathbb{R}^n$ stetig differenzierbar und injektiv mit stetig differenzierbarer Inverser, und sei $f : T(U) \rightarrow \mathbb{R}$ integrierbar, dann ist

$$\int_{T(U)} f(\mathbf{y}) d\mathbf{y} = \int_U f(T(\mathbf{x})) \cdot |\det DT(\mathbf{x})| d\mathbf{x}, \quad DT(\mathbf{x}) = \left(\frac{\partial T_i}{\partial x_j} \right)_{i,j=1,\dots,n}, \quad T = (T_1, \dots, T_n)^T.$$

Bemerkung. Ist insbesondere T linear-affin, also $T(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{t}$ für $\mathbf{A} \in \mathbb{R}^{n \times n}$ invertierbar und $\mathbf{t} \in \mathbb{R}^n$, dann ist

$$\int_{\mathbf{A}U + \mathbf{t}} f(\mathbf{y}) d\mathbf{y} = \int_U f(\mathbf{A}\mathbf{x} + \mathbf{t}) \cdot |\det \mathbf{A}| d\mathbf{x}.$$

Lemma 2.16. Ist $\mu \in \mathbb{R}^d$ und $\Sigma \in \mathbb{R}^{d \times d}$ symmetrisch und positiv definit, dann ist

$$\int_{\mathbb{R}^d} f(\mathbf{x}; \mu, \Sigma) d\mathbf{x} = 1$$

Beweis. Zunächst diagonalisieren wir Σ , finden also eine orthogonale Matrix \mathbf{Q} und die Eigenwerte $\lambda_1, \dots, \lambda_d \in \mathbb{R}^+$ von Σ (positiv, da Σ positiv definit), so dass

$$\Sigma = \mathbf{Q} \text{diag}(\lambda_1, \dots, \lambda_d) \mathbf{Q}^T.$$

Dann definieren wir die Wurzel aus Σ durch

$$\Sigma^{1/2} = \mathbf{Q} \text{diag}(\lambda_1^{1/2}, \dots, \lambda_d^{1/2}) \mathbf{Q}^T.$$

Diese ist eindeutig bestimmt mit $\Sigma^{1/2} \cdot \Sigma^{1/2} = \Sigma$. Außerdem gelten

$$\begin{aligned} \Sigma^{-1} &= \mathbf{Q} \text{diag}(\lambda_1^{-1}, \dots, \lambda_d^{-1}) \mathbf{Q}^T, \\ \Sigma^{-1/2} &= \mathbf{Q} \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_d^{-1/2}) \mathbf{Q}^T. \end{aligned}$$

Mit der Substitution $\mathbf{y} = \Sigma^{-1/2}(\mathbf{x} - \mu) \Leftrightarrow \mathbf{x} = \Sigma^{1/2}\mathbf{y} + \mu$ ist nun nach der Transformationsformel

$$\begin{aligned} \int_{\mathbb{R}^d} f(\mathbf{x}; \mu, \Sigma) d\mathbf{x} &= \int_{\mathbb{R}^d} f(\Sigma^{1/2}\mathbf{y} + \mu; \mu, \Sigma) \sqrt{\det \Sigma} d\mathbf{y} \\ &= \frac{1}{\sqrt{(2\pi)^d}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}(\Sigma^{1/2}\mathbf{y})^T \Sigma^{-1} \Sigma^{1/2}\mathbf{y}\right) d\mathbf{y} \\ &= \frac{1}{\sqrt{(2\pi)^d}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\mathbf{y}^T \Sigma^{1/2} \Sigma^{-1/2} \Sigma^{-1/2} \Sigma^{1/2}\mathbf{y}\right) d\mathbf{y} \\ &= \frac{1}{\sqrt{(2\pi)^d}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\mathbf{y}^T \mathbf{y}\right) d\mathbf{y} \end{aligned}$$

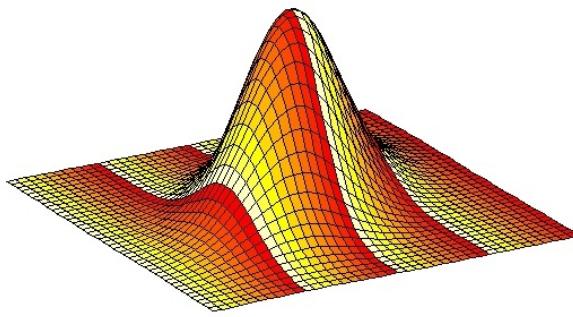


Abbildung 2.1: Zweidimensionale Dichte einer Standardnormalverteilung

$$= \int_{\mathbb{R}^d} f(\mathbf{y}; \mathbf{0}, \mathbf{I}_d) d\mathbf{y} = 1.$$

Dies gilt, da die Dichte von $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ als Produktdichte von univariaten Standard-Normalverteilungen zu 1 integriert. ■

Ist $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, so ist $Z = \Sigma^{1/2}X + \mu \sim \mathcal{N}(\mu, \Sigma)$, wobei die entsprechende Dichte aus einer Dichtetransformation hergeleitet werden kann. Wir wissen auch, wie sich charakteristische Funktionen unter linearen Transformation transformieren. Demnach ist ein Zufallsvektor \mathbf{X} in \mathbb{R}^d multivariat normalverteilt, $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, falls $\mu \in \mathbb{R}^d$ und $\Sigma \in \mathbb{R}^{d \times d}$, $\Sigma \geq 0$ (positiv semidefinit) existieren, so dass \mathbf{X} folgende charakteristische Funktion hat

$$\varphi_{\mathbf{X}}(\mathbf{t}) = e^{i\mu^T \mathbf{t}} e^{-\mathbf{t}^T \Sigma \mathbf{t}/2}. \quad (2.4)$$

Da hier Σ^{-1} nicht vorkommt, kann diese Definition etwas allgemeiner verwendet werden. Ist Σ singulär, so nennt man die Verteilung von \mathbf{X} degeneriert.

Satz 2.17 (Lineare Transformation). Sei $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, $\mu \in \mathbb{R}^d$, $\mathbf{A} \in \mathbb{R}^{p \times d}$ mit vollem Rang und $\mathbf{b} \in \mathbb{R}^p$. Dann ist $\mathbf{Y} = \mathbf{AX} + \mathbf{b} \sim \mathcal{N}(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}\Sigma\mathbf{A}^T)$. ■

Beweis. Wir benutzen (2.4). Für $s \in \mathbb{R}^n$ ist

$$\mathbb{E}[e^{is^T(\mathbf{AX}+\mathbf{b})}] = e^{i\mathbf{b}^T s} \mathbb{E}[e^{i(s^T \mathbf{A})\mathbf{X}}] = e^{is^T \mathbf{b}} e^{i(s^T \mathbf{A})\mu} e^{-(s^T \mathbf{A})\Sigma \mathbf{A}^T s/2}.$$

Es folgt insbesondere, dass **Marginalverteilungen einer multivariaten Normalverteilung wieder Normalverteilungen** sind. Dass jede lineare Transformation eines normalverteilten Zufallsvektors wieder normalverteilt ist, ist die wohl wichtigste Eigenschaft der multivariaten Normalverteilung. Eine ebenfalls sehr bedeutende Eigenschaft folgt, falls Σ diagonal ist: **Gemeinsam normalverteilte Zufallsvariablen, die unkorreliert sind, sind unabhängig**. Dies lässt sich daran sehen, dass die gemeinsame Dichte in dem Fall die Produktdichte der marginalen Normalverteilungsdichten ist.

Satz 2.18 (Weitere Eigenschaften der multivariaten Normalverteilung). Seien $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ mit $\Sigma \geq 0$ und $\mathbf{X} = (X_1, \dots, X_d)^T \sim \mathcal{N}(\mu, \Sigma)$.

1. **Momente:** Es ist $\mathbb{E}[X_i^2] < \infty$, $i = 1, \dots, d$, und $\mathbb{E}[\mathbf{X}] = \mu$, $\text{Cov}(\mathbf{X}) = \Sigma$.

Sei nun $\mathbf{X} = (\mathbf{Y}^T, \mathbf{Z}^T)^T$, $\mathbf{Y} \in \mathbb{R}^r$, $\mathbf{Z} \in \mathbb{R}^{d-r}$, für $1 \leq r < d$, und sei

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21}^T & \Sigma_{22} \end{pmatrix}$$

mit $\mu_1 \in \mathbb{R}^r$, $\mu_2 \in \mathbb{R}^{d-r}$, $\Sigma_{11} \in \mathbb{R}^{r \times r}$, $\Sigma_{22} \in \mathbb{R}^{(d-r) \times (d-r)}$, $\Sigma_{12} \in \mathbb{R}^{r \times (d-r)}$.

2. **Marginalverteilungen:** Es sind $\mathbf{Y} \sim \mathcal{N}(\mu_1, \Sigma_{11})$, $\mathbf{Z} \sim \mathcal{N}(\mu_2, \Sigma_{22})$, und $\text{Cov}(\mathbf{Y}, \mathbf{Z}) = \Sigma_{12}$.

3. **Unabhängigkeit:** \mathbf{Y} und \mathbf{Z} sind genau dann unabhängig, wenn $\Sigma_{12} = 0$. ■

Beweis. Wir nutzen stets (2.4).

Zu 1.: Ist $\mathbf{X} \sim \mathcal{N}(0, I_d)$, so ist die Behauptung richtig. Andernfalls ist für $\mathbf{Y} \sim \mathcal{N}(0, I_d)$ der Zufallsvektor $\mathbf{X} = A\mathbf{Y} + \mu \sim \mathcal{N}(\mu, \Sigma)$, falls $AA^T = \Sigma$. Daher folgt die Behauptung aus den Rechenregeln für Kovarianzmatrizen.

Zu 2.: Nutze Satz 2.17 mit $\mathbf{A} = (I_r, 0) \in \mathbb{R}^{r \times d}$ für \mathbf{Y} bzw. $\mathbf{A} = (0, I_{d-r}) \in \mathbb{R}^{(d-r) \times d}$ für \mathbf{Z} , sowie 1. für die Kovarianzmatrix.

Zu 3.: Aus der Unabhängigkeit folgt, dass alle Kovarianzen 0 sind. Ist umgekehrt $\Sigma_{12} = 0$, so folgt für $\mathbf{t}^T = (\mathbf{y}^T, \mathbf{z}^T)$,

$$\varphi_{\mathbf{X}}(\mathbf{t}) = e^{i\mu_1^T \mathbf{y} + \mu_2^T \mathbf{z}} \exp \left(-(\mathbf{y}^T, \mathbf{z}^T) \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} / 2 \right) = \varphi_{\mathbf{Y}}(\mathbf{y}) \varphi_{\mathbf{Z}}(\mathbf{z}).$$

Wir schließen mit Satz 1.15. ■

Satz 2.19. Ein Zufallsvektor \mathbf{X} im \mathbb{R}^d ist genau dann normalverteilt, wenn jede Linearkombination $\mathbf{a}^T \mathbf{X}$, $\mathbf{a} \in \mathbb{R}^d$ (univariat) normalverteilt ist. ■

Beweis. Ist \mathbf{X} multivariat normal, so ist jede Linearkombination nach Satz 2.17 univariat normal.

Angenommen, jede Linearkombination sei normalverteilt. Für $\mathbf{a} = \mathbf{e}_i$, den i -ten Einheitsvektor, erhält man die Komponenten von \mathbf{X} , die insbesondere endliche Varianz haben, also existieren Erwartungswertvektor $\mu = \mathbb{E}[\mathbf{X}]$ und Kovarianzmatrix $\Sigma = \text{Cov}(\mathbf{X})$ von X . Für $\mathbf{a} \in \mathbb{R}^d$ muss dann gelten, dass

$$\mathbf{a}^T \mathbf{X} \sim \mathcal{N}(\mathbf{a}^T \mu, \mathbf{a}^T \Sigma \mathbf{a}).$$

Da $\varphi_{\mathbf{X}}(\mathbf{t}) = \varphi_{\mathbf{t}^T \mathbf{X}}(1)$, ergibt sich die Behauptung. ■

Bemerkung (Folgerung). Ist $\mathbf{X} \sim \mathcal{N}(0, \mathbf{I}_n)$, $\mu \in \mathbb{R}^n$ und $\Sigma \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit mit $\Sigma = \mathbf{L}\mathbf{L}^T$. Dann ist $\mathbf{L}\mathbf{X} + \mu \sim \mathcal{N}(\mu, \Sigma)$.

Beispiel 2.20. Die multivariate Normalverteilung ist grundlegend für die Modellierung multivariater zufälliger Größen. Eine wichtige Rolle hat sie etwa im sogenannten *Varianz-Kovarianz Ansatz* im Risikocontrolling. Wir nehmen an, der Verlust L eines Portfolios (über einen festen Zeitraum T) setze sich zusammen als eine Linearkombination einzelner Risikofaktoren R_i mit Gewichten $w_i \in \mathbb{R}$, $i = 1, \dots, m$, also

$$L = \sum_{i=1}^m w_i R_i = \mathbf{w}^T \mathbf{R}, \quad \mathbf{w} = (w_1, \dots, w_m)^T, \quad \mathbf{R} = (R_1, \dots, R_m)^T.$$

Ist nun \mathbf{R} multivariat normalverteilt, also $\mathbf{R} \sim \mathcal{N}(\mu, \Sigma)$, so ist L univariat normalverteilt mit $L = \mathbf{w}^T \mathbf{R} \sim \mathcal{N}(\mathbf{w}^T \mu, \mathbf{w}^T \Sigma \mathbf{w})$. Für den *Value at risk* von L zum Niveau α gilt dann

$$\text{VaR}_\alpha(L) = (\mathbf{w}^T \Sigma \mathbf{w})^{1/2} q_\alpha + \mathbf{w}^T \mu,$$

wobei q_α das α -Quantil von $\mathcal{N}(0, 1)$ ist.

2.5 Bedingte Verteilungen

2.5.1 Diskrete bedingte Verteilungen

Beispiel 2.21. Seien X_1, \dots, X_n unabhängig und $\text{Ber}(p)$ -verteilt, $p \in (0, 1)$. Angenommen, für die Anzahl der Erfolge $Y = X_1 + \dots + X_n$ wird $Y = y \in \{0, 1, \dots, n\}$ beobachtet. Was kann dann über die Verteilung der k Erfolge auf die möglichen n Positionen ausgesagt werden? Sei dazu $(x_1, \dots, x_n)^\top \in \{0, 1\}^n$. Ist $x_1 + \dots + x_n \neq y$, so ist

$$\{X_1 = x_1, \dots, X_n = x_n, Y = y\} = \emptyset,$$

und

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n | Y = y) = 0.$$

Gilt dagegen $x_1 + \dots + x_n = y$, so ist $\{Y = y\} \supseteq \{X_1 = x_1, \dots, X_n = x_n\}$, und da $Y \sim \text{Bin}(n, p)$ ergibt sich

$$\begin{aligned}\mathbb{P}(X_1 = x_1, \dots, X_n = x_n | Y = y) &= \frac{\mathbb{P}(X_1 = x_1, \dots, X_n = x_n)}{\mathbb{P}(Y = y)} \\ &= \frac{p^y (1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{1}{\binom{n}{y}}.\end{aligned}$$

Also erhält man für die bedingte Verteilung der Positionen der Erfolge gegeben $Y = y$ Erfolge die uniforme Verteilung auf

$$\{(x_1, \dots, x_n)^\top \in \{0, 1\}^n \mid x_1 + \dots + x_n = y\}.$$

Seien $X : \Omega \rightarrow \Omega_X$ eine Ω_X -wertige diskrete Zufallsvariable und $Y : \Omega \rightarrow \Omega_Y$ eine Ω_Y -wertige diskrete Zufallsvariable.

Definition 2.22. Ist $y \in Y(\Omega)$ mit $p_Y(y) = \mathbb{P}(Y = y) > 0$, so heißt die Wahrscheinlichkeitsverteilung auf $X(\Omega)$ mit

$$A \mapsto \mathbb{P}(X \in A \mid Y = y) = \mathbb{P}(\{X \in A\} \mid \{Y = y\})$$

die **bedingte Verteilung** von X gegeben $Y = y$. Wir schreiben auch $\mathbb{P}_{X|Y}(A \mid y)$. Die zugehörige Wahrscheinlichkeitsfunktion auf $X(\Omega)$ ist

$$x \mapsto \mathbb{P}(X = x \mid Y = y) =: p_{X|Y}(x \mid y).$$

Bemerkung. 1. Die bedingte Verteilung ist die Verteilung von X bzgl. des bedingten Wahrscheinlichkeitsmaßes $\mathbb{P}(\cdot \mid \{Y = y\})$ auf Ω .

2. Ist $\mathbb{P}_{X,Y}$ die gemeinsame Verteilung von X und Y , so ist für $p_Y(y)$

$$\mathbb{P}_{X|Y}(A \mid y) = \frac{\mathbb{P}(\{X \in A\} \cap \{Y = y\})}{\mathbb{P}(\{Y = y\})} = \frac{\mathbb{P}_{X,Y}(A \times \{y\})}{\mathbb{P}_Y(\{y\})},$$

bzw.

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

Die bedingte Verteilung ist also aus der gemeinsamen Verteilung bestimmt.

Beispiel 2.23. Sind $X \sim \text{Poi}(\lambda)$ und $Y \sim \text{Poi}(\mu)$ unabhängig, so gilt $X + Y \sim \text{Poi}(\lambda + \mu)$ und für $k, z \in \mathbb{N}_0$, $k \leq z$, dann

$$p_{X|X+Y}(k \mid z) = \frac{\mathbb{P}(X = k) \cdot \mathbb{P}(Y = z - k)}{\mathbb{P}(X + Y = z)} = \frac{e^{-\lambda} \frac{\lambda^k}{k!} \cdot e^{-\mu} \frac{\mu^{z-k}}{(z-k)!}}{e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^z}{z!}}$$

$$= \binom{z}{k} \left(\frac{\lambda}{\lambda + \mu} \right)^k \left(\frac{\mu}{\lambda + \mu} \right)^{z-k},$$

also $\mathbb{P}_{X|Y} = \text{Bin}(z, \frac{\lambda}{\lambda + \mu})$.

Satz 2.24. Seien $X : \Omega \rightarrow \Omega_X$ und $Y : \Omega \rightarrow \Omega_Y$ diskrete Zufallsvariablen. Es sind äquivalent:

1. X und Y sind unabhängig.
2. Die bedingte Verteilung von X gegeben $Y = y$ hängt für $p_Y(y) > 0$ nicht von y ab.

■

Beweis. 1. \Rightarrow 2. Sind X und Y unabhängig, so ist für $\mathbb{P}(Y = y) > 0$ und $A \subseteq X(\Omega)$ stets

$$\mathbb{P}_{X|Y}(A | y) = \frac{\mathbb{P}(X \in A, Y = y)}{\mathbb{P}(Y = y)} = \frac{\mathbb{P}(X \in A) \cdot \mathbb{P}(Y = y)}{\mathbb{P}(Y = y)} = \mathbb{P}(X \in A),$$

die Verteilung hängt also nicht von y ab.

2. \Rightarrow 1. Ist $f(A) = \mathbb{P}_{X|Y}(A | y)$ für alle y mit $p_Y(y) > 0$, so gilt

$$\begin{aligned} \mathbb{P}(X \in A) &= \sum_{y \in Y(\Omega)} \mathbb{P}(X \in A, Y = y) = \sum_{\substack{y \in Y(\Omega) \\ p_Y(y) > 0}} \mathbb{P}(X \in A, Y = y) \\ &= \sum_{\substack{y \in Y(\Omega) \\ p_Y(y) > 0}} \mathbb{P}_{X|Y}(A | y) \cdot \mathbb{P}(Y = y) = f(A) \cdot \overbrace{\sum_{\substack{y \in Y(\Omega) \\ p_Y(y) > 0}} \mathbb{P}(Y = y)}^{=1} = f(A). \end{aligned}$$

Somit gilt

$$\mathbb{P}(X = x) = \mathbb{P}_{X|Y}(\{x\} | y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} \text{ für } \mathbb{P}(Y = y) > 0,$$

was mit $\mathbb{P}(X = x) \cdot \mathbb{P}(Y = y) = \mathbb{P}(X = x, Y = y)$ Unabhängigkeit liefert, da die Formel für $p_Y(y) = 0$ klar ist. ■

Bemerkung. Sind $X : \Omega \rightarrow \mathbb{R}$ eine diskrete reellwertige Zufallsvariable mit $\mathbb{E}[|X|] < \infty$ und $Y : \Omega \rightarrow \Omega_Y$, $y \in Y(\Omega)$ mit $p_Y(y) > 0$, so gilt

$$\sum_{x \in X(\Omega)} |x| \cdot \mathbb{P}_{X|Y}(x | y) < \infty, \quad (2.5)$$

da gilt $\mathbb{P}_{X,Y}(x, y) \leq \mathbb{P}_X(x)$ und somit

$$\sum_{x \in X(\Omega)} |x| \cdot \mathbb{P}_{X,Y}(x, y) \leq \sum_{x \in X(\Omega)} |x| \cdot \mathbb{P}_X(x) = \mathbb{E}[|X|] < \infty.$$

Division durch $p_Y(y)$ liefert (2.5).

Definition 2.25. Sind $X : \Omega \rightarrow \mathbb{R}$ eine diskrete reellwertige Zufallsvariable mit $\mathbb{E}[|X|] < \infty$ und $Y : \Omega \rightarrow \Omega_Y$, $y \in Y(\Omega)$ mit $p_Y(y) > 0$, so heißt

$$\mathbb{E}[X | Y = y] := \sum_{x \in X(\Omega)} x \cdot \mathbb{P}_{X|Y}(x | y)$$

der **bedingte Erwartungswert** von X gegeben $Y = y$.

Bemerkung. Dies ist der Erwartungswert von X bezüglich des bedingten Wahrscheinlichkeitsmaßes $\mathbb{P}(\cdot | \{Y = y\})$ auf Ω .

Beispiel 2.26 (2.23 fort.). Sind $X \sim \text{Poi}(\lambda)$ und $Y \sim \text{Poi}(\mu)$ unabhängig, so ist für $z \in \mathbb{N}_0$

$$\mathbb{E}[X | X + Y = z] = z \cdot \frac{\lambda}{\lambda + \mu}.$$

Definition 2.27. Seien $X : \Omega \rightarrow \mathbb{R}$ und $Y : \Omega \rightarrow \Omega_Y$ diskrete Zufallsvariablen und $\mathbb{E}[|X|] < \infty$. Wir setzen $g : Y(\Omega) \rightarrow \mathbb{R}$,

$$g(y) := \begin{cases} \mathbb{E}[X | Y = y], & \text{falls } p_Y(y) > 0, \\ 0, & \text{sonst.} \end{cases}$$

Die Zufallsvariable $\mathbb{E}[X | Y] : \Omega \rightarrow \mathbb{R}$, $\mathbb{E}[X | Y](\omega) = g(Y(\omega))$ heißt **bedingte Erwartung** von X gegeben Y .

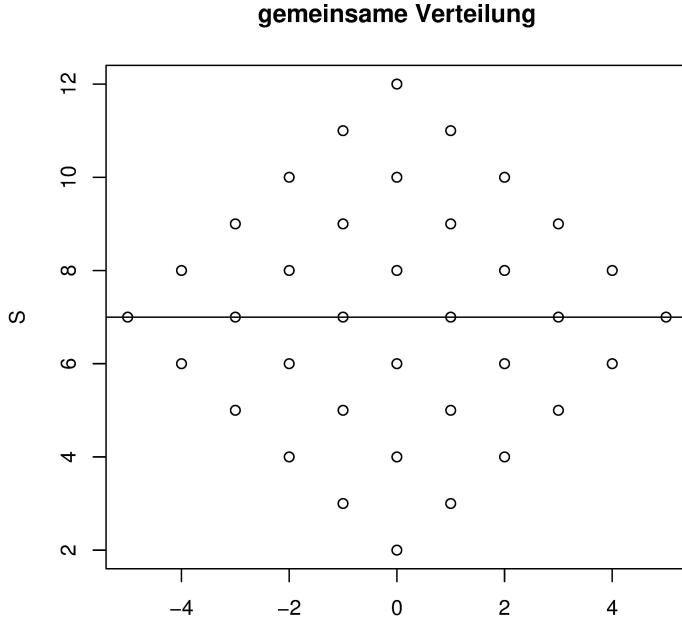
Beispiel 2.28 (2.23 fort.). Sind $X \sim \text{Poi}(\lambda)$ und $Y \sim \text{Poi}(\mu)$ unabhängig, so ist

$$\mathbb{E}[X | X + Y] = (X + Y) \cdot \frac{\lambda}{\lambda + \mu}.$$

Beispiel 2.29. Betrachte beim zweifachen Würfeln die **Summe S** und die **Differenz D** der beiden geworfenen Augenzahlen. Die gemeinsame Verteilung ergibt sich aus folgendem Schema,

S/D	2	3	4	5	6	7	8	9	10	11	12
0	(1,1)	-	(2,2)	-	(3,3)	-	(4,4)	-	(5,5)	-	(6,6)
1	-	(2,1)	-	(3,2)	-	(4,3)	-	(5,4)	-	(6,5)	-
		(1,2)	-	(2,3)	-	(3,4)	-	(4,5)	-	(5,6)	
2	-	-	(3,1)	-	(4,2)	-	(5,3)	-	(6,4)	-	-
			(1,3)	-	(2,4)	-	(3,5)	-	(4,6)		
3	-	-	-	(4,1)	-	(5,2)	-	(6,3)	-	-	-
				(1,4)	-	(2,5)	-	(3,6)			
4	-	-	-	-	(5,1)	-	(6,2)	-	-	-	-
					(1,5)	-	(2,6)	-			
5	-	-	-	-	-	(6,1)	-	-	-	-	-
						(1,6)					

veranschaulicht in der nachfolgenden Grafik.



Als bedingte Erwartungswerte ergeben sich:

$$\begin{aligned}\mathbb{E}[S \mid D = 0] &= \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 6 + \frac{1}{6} \cdot 8 + \frac{1}{6} \cdot 10 + \frac{1}{6} \cdot 12 = 7, \\ \mathbb{E}[S \mid D = 1] &= \frac{1}{5} \cdot 3 + \frac{1}{5} \cdot 5 + \frac{1}{5} \cdot 7 + \frac{1}{5} \cdot 9 + \frac{1}{5} \cdot 11 = 7, \\ \mathbb{E}[S \mid D = 2] &= \frac{1}{4} \cdot 4 + \frac{1}{4} \cdot 6 + \frac{1}{4} \cdot 8 + \frac{1}{4} \cdot 10 = 7, \\ \mathbb{E}[S \mid D = 3] &= \frac{1}{3} \cdot 5 + \frac{1}{3} \cdot 7 + \frac{1}{3} \cdot 9 = 7, \\ \mathbb{E}[S \mid D = 4] &= \frac{1}{2} \cdot 6 + \frac{1}{2} \cdot 8 = 7, \\ \mathbb{E}[S \mid D = 5] &= 7.\end{aligned}$$

Also gilt offensichtlich nicht nur $\text{Cov}(S, D) = 0$, sondern auch $\mathbb{P}(\mathbb{E}[S \mid D] = \mathbb{E}[S] = 7) = 1$. Wir wissen aber, dass S und D **nicht unabhängig** sind.

Satz 2.30 (totale Erwartung, tower rule). Seien $X : \Omega \rightarrow \mathbb{R}$ und $Y : \Omega \rightarrow \Omega_Y$ diskrete Zufallsvariablen und $\mathbb{E}[|X|] < \infty$. Dann gilt

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X].$$

■

Beweis. Es seien $g : Y(\Omega) \rightarrow \mathbb{R}$, $g(Y) = \mathbb{E}[X \mid Y]$, wie in der Definition der bedingten Erwartung und analog

$$\tilde{g}(y) := \begin{cases} \mathbb{E}[|X| \mid Y = y], & \text{falls } p_Y(y) > 0, \\ 0, & \text{sonst,} \end{cases}$$

so dass $\tilde{g}(Y) = \mathbb{E}[|X| \mid Y]$. Da $g(y)$ bzw. $\tilde{g}(y)$ für $p_Y(y) > 0$ der Erwartungswert von X bzw. $|X|$ unter $\mathbb{P}(\cdot \mid Y = y)$ ist, folgt $|g| \leq \tilde{g}$. Es ist

$$\begin{aligned}\mathbb{E}[\tilde{g}(Y)] &= \sum_{y \in Y(\Omega), p_Y(y) > 0} \tilde{g}(y) \cdot p_Y(y) = \sum_{y \in Y(\Omega), p_Y(y) > 0} \sum_{x \in X(\Omega)} |x| \cdot p_{X|Y}(x \mid y) \cdot p_Y(y) \\ &= \sum_{x \in X(\Omega)} |x| \sum_{y \in Y(\Omega), p_Y(y) > 0} p_{X,Y}(x, y) = \sum_{x \in X(\Omega)} |x| \cdot p_X(x) = \mathbb{E}[|X|] < \infty.\end{aligned}$$

Wegen $|g(y)| \leq \tilde{g}(y)$ ist $\mathbb{E}[|g(Y)|] \leq \mathbb{E}[|X|] < \infty$. Die gleiche Rechnung wie oben ohne Beträge liefert dann die Behauptung. ■

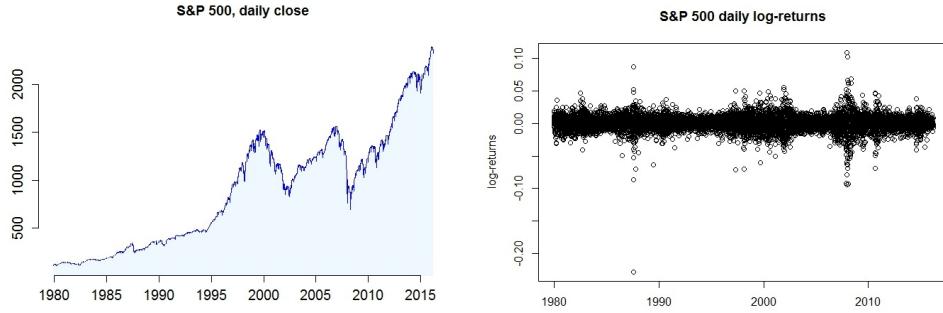


Abbildung 2.2: Tagesschlusswerte mit zugehörigen Log-Retruns des S&P 500

Beispiel 2.31 (2.23 fort.). Sind $X \sim \text{Poi}(\lambda)$ und $Y \sim \text{Poi}(\mu)$ unabhängig, so ist $X + Y \sim \text{Poi}(\lambda + \mu)$, also

$$\mathbb{E}[\mathbb{E}[X | X + Y]] = \mathbb{E}\left[(X + Y) \cdot \frac{\lambda}{\lambda + \mu}\right] = \frac{\lambda}{\lambda + \mu} \cdot (\lambda + \mu) = \lambda = \mathbb{E}[X].$$

Manchmal sind andere Charakteristika der bedingten Verteilung als der bedingte Erwartungswert von Interesse, insbesondere deren Varianz und Standardabweichung, die sogenannte **bedingte Varianz** bzw. **bedingte Standardabweichung**.

Die bedingte Standardabweichung spielt in der Finanzmathematik und insbesondere im Risikomanagement eine große Rolle. Sei (S_t) der Tagesschlusskurs einer Aktie am Handelstag t , dann heißt die relative Wertänderung der Aktie $R_t = (S_t - S_{t-1})/S_{t-1}$ die **Rendite** oder der **Return** der Aktie auf Tagesbasis. Plottet man (R_t) , so stellt man fest, dass es Phasen großer Streuung sowie Phasen geringerer Streuung gibt, sogenannte **Volatilitätscluster**, vgl. Abb. 2.2. Die bedingte Standardabweichung von R_t gegeben die Historie R_{t-1}, R_{t-2}, \dots heißt **historische Volatilität**, diese variiert offenbar in der Zeit. Dagegen ist die bedingte Erwartung stets nahe bei 0.

Definition 2.32. Sind $X : \Omega \rightarrow \mathbb{R}$ und $Y : \Omega \rightarrow \Omega_Y$, $y \in Y(\Omega)$ diskrete reellwertige Zufallsvariablen mit $\mathbb{E}[X^2] < \infty$ und $p_Y(y) > 0$, so heißt

$$\text{Var}(X | Y = y) := \mathbb{E}[(X - \mathbb{E}[X | Y = y])^2 | Y = y]$$

die **bedingte Varianz** von X gegeben $Y = y$ sowie

$$\sigma(X | Y = y) = (\text{Var}(X | Y = y))^{1/2}$$

die **bedingte Standardabweichung**. Die **Zufallsvariable** $\text{Var}(X | Y) = h(Y)$ mit

$$h(y) := \begin{cases} \text{Var}(X | Y = y), & \text{falls } p_Y(y) > 0, \\ 0, & \text{sonst,} \end{cases}$$

heißt die **bedingte Varianz** von X gegeben Y .

Bemerkung. $\text{Var}(X | Y = y)$ ist die Varianz von X bezüglich des Wahrscheinlichkeitsmaßes $\mathbb{P}(\cdot | \{Y = y\})$ auf Ω .

Beispiel 2.33 (2.23 fort.). Sind $X \sim \text{Poi}(\lambda)$ und $Y \sim \text{Poi}(\mu)$ unabhängig und $z \in \mathbb{N}_0$, so ist

$$\text{Var}(X | X + Y = z) = z \cdot \frac{\lambda}{\lambda + \mu} \cdot \frac{\mu}{\lambda + \mu}.$$

Satz 2.34. Sind $X : \Omega \rightarrow \mathbb{R}$ und $Y : \Omega \rightarrow \Omega_Y$ diskrete reellwertige Zufallsvariablen und $\mathbb{E}[X^2] < \infty$, so ist

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X | Y)] + \text{Var}(\mathbb{E}[X | Y]).$$

■

Beweis. Für Konstante $a \in \mathbb{R}$ gilt

$$\mathbb{E}[(X - a)^2] = \text{Var}(X) + (a - \mathbb{E}[X])^2.$$

Setzen wir $a = \mathbb{E}[X]$ und wenden das bezüglich der Verteilung $\mathbb{P}(\cdot | Y = y)$ an, so erhalten wir

$$\mathbb{E}[(X - \mathbb{E}[X])^2 | Y = y] = \text{Var}(X | Y = y) + (\mathbb{E}[X] - \mathbb{E}[X | Y = y])^2.$$

Gehen wir weiter zur bedingten Erwartung über, so gilt

$$\mathbb{E}[(X - \mathbb{E}[X])^2 | Y] = \text{Var}(X | Y) + (\mathbb{E}[X | Y] - \mathbb{E}[X])^2.$$

Bilden wir davon den Erwartungswert, so erhalten wir

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X | Y)] + \text{Var}(\mathbb{E}[X | Y]).$$

■

Beispiel 2.35 (2.23 fort.). Sind $X \sim \text{Poi}(\lambda)$ und $Y \sim \text{Poi}(\mu)$ unabhängig, so ist

$$\begin{aligned}\mathbb{E}[\text{Var}(X | X + Y)] &= (\lambda + \mu) \cdot \frac{\lambda \cdot \mu}{(\lambda + \mu)^2} = \frac{\lambda \cdot \mu}{\lambda + \mu}, \\ \text{Var}(\mathbb{E}[X | X + Y]) &= \text{Var}\left(\frac{\lambda}{\lambda + \mu} \cdot (X + Y)\right) = \frac{\lambda^2}{(\lambda + \mu)^2} \cdot (\lambda + \mu) = \frac{\lambda^2}{\lambda + \mu}, \\ \mathbb{E}[\text{Var}(X | X + Y)] + \text{Var}(\mathbb{E}[X | X + Y]) &= \frac{\lambda \cdot \mu}{\lambda + \mu} + \frac{\lambda^2}{\lambda + \mu} = \frac{\lambda \cdot (\lambda + \mu)}{\lambda + \mu} = \lambda = \text{Var}(X).\end{aligned}$$

2.5.2 Bedingte absolutstetige Verteilungen mit Dichten

Im Kontext dieses Abschnitts seien $\mathbf{Y} : \Omega \rightarrow \mathbb{R}^r$, $\mathbf{Z} : \Omega \rightarrow \mathbb{R}^{n-r}$ und $\mathbf{X} = (\mathbf{Y}^T, \mathbf{Z}^T)^T$ absolutstetige Zufallsvektoren. Dabei sei $f_{\mathbf{Y}, \mathbf{Z}}$ die Dichte von \mathbf{X} und seien $f_{\mathbf{Y}}$ bzw. $f_{\mathbf{Z}}$ die marginalen Dichten von \mathbf{Y} bzw. \mathbf{Z} .

Definition 2.36. Ist $\mathbf{z} \in \mathbb{R}^{n-r}$ mit $f_{\mathbf{Z}}(\mathbf{z}) > 0$, so heißt

$$f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} | \mathbf{z}) = \frac{f_{\mathbf{Y}, \mathbf{Z}}(\mathbf{y}, \mathbf{z})}{f_{\mathbf{Z}}(\mathbf{z})}, \quad \mathbf{y} \in \mathbb{R}^r$$

die **bedingte Dichte** von \mathbf{Y} gegeben $\mathbf{Z} = \mathbf{z}$. Die von dieser repräsentierte Verteilung heißt **bedingte Verteilung** von \mathbf{Y} gegeben $\mathbf{Z} = \mathbf{z}$.

Bemerkung (Motivation). Ist $f_{\mathbf{Z}}(z_0) > 0$, $B \in \mathcal{B}$ und $r = n - r = 1$, dann ist

$$\begin{aligned}\mathbb{P}(\mathbf{Y} \in B | \mathbf{Z} = z_0) &= \lim_{\varepsilon \rightarrow 0} \mathbb{P}(\mathbf{Y} \in B | z_0 \leq \mathbf{Z} \leq z_0 + \varepsilon) = \lim_{\varepsilon \rightarrow 0} \frac{\int_B \int_{z_0}^{z_0+\varepsilon} f_{\mathbf{Y}, \mathbf{Z}}(y, z) dz dy}{\int_{z_0}^{z_0+\varepsilon} f_{\mathbf{Z}}(z) dz} \\ &= \lim_{\varepsilon \rightarrow 0} \int_B \frac{\int_{z_0}^{z_0+\varepsilon} f_{\mathbf{Y}, \mathbf{Z}}(y, z) dz}{\int_{z_0}^{z_0+\varepsilon} f_{\mathbf{Z}}(z) dz} dy = \int_B \frac{f_{\mathbf{Y}|\mathbf{Z}}(y, z_0)}{f_{\mathbf{Z}}(z_0)} dy = \int_B f_{\mathbf{Y}|\mathbf{Z}}(y | z_0) dy.\end{aligned}$$

Satz 2.37. Sind die Zufallsvariablen normalverteilt mit

$$\mathbf{Y} \sim \mathcal{N}(\mu_1, \Sigma_1), \quad \mathbf{Z} \sim \mathcal{N}(\mu_2, \Sigma_2), \quad \mathbf{X} = \begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \Sigma_{1,2} \\ \Sigma_{1,2}^T & \Sigma_2 \end{pmatrix}\right) = \mathcal{N}(\mu, \Sigma),$$

$\mu_1 \in \mathbb{R}^r, \mu_2 \in \mathbb{R}^{n-r}, \Sigma_1 \in \mathbb{R}^{r \times r}, \Sigma_2 \in \mathbb{R}^{(n-r) \times (n-r)}$ und $\Sigma_{1,2} \in \mathbb{R}^{r \times (n-r)}$, so ist

$$(\mathbf{Y} | \mathbf{Z} = \mathbf{z}) \sim \mathcal{N}(\mu_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}}, \Sigma_{1|2}) = \mathcal{N}(\mu_1 + \Sigma_{1,2}\Sigma_2^{-1}(\mathbf{z} - \mu_2), \Sigma_1 - \Sigma_{1,2}\Sigma_2^{-1}\Sigma_{1,2}^T).$$

Beweis. Nach Definition ist (mit $\mathbf{x} = (\mathbf{y}^T, \mathbf{z}^T)^T$)

$$\begin{aligned} f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} | \mathbf{z}) &= \frac{(\det \Sigma (2\pi)^n)^{-1/2} \exp(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu))}{(\det \Sigma_2 (2\pi)^{n-r})^{-1/2} \exp(-\frac{1}{2}(\mathbf{z} - \mu_2)^T \Sigma_2^{-1}(\mathbf{z} - \mu_2))} \\ &= \frac{\exp(-\frac{1}{2}((\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu) - (\mathbf{z} - \mu_2)^T \Sigma_2^{-1}(\mathbf{z} - \mu_2))}{\sqrt{\frac{\det \Sigma}{\det \Sigma_2} (2\pi)^r}}. \end{aligned}$$

Zunächst berechnen wir Σ^{-1} :

$$\begin{aligned} \Sigma \Sigma^{-1} &= \mathbf{I}_n \\ \Rightarrow \begin{pmatrix} \Sigma_1 & \Sigma_{1,2} \\ \Sigma_{1,2}^T & \Sigma_2 \end{pmatrix} \Sigma^{-1} &= \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{pmatrix} \\ \Rightarrow \begin{pmatrix} \Sigma_1 & \Sigma_{1,2} \\ \Sigma_2^{-1}\Sigma_{1,2}^T & \mathbf{I}_{n-r} \end{pmatrix} \Sigma^{-1} &= \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \Sigma_2^{-1} \end{pmatrix} \\ \Rightarrow \begin{pmatrix} \Sigma_{1|2} & \mathbf{0} \\ \Sigma_2^{-1}\Sigma_{1,2}^T & \mathbf{I}_{n-r} \end{pmatrix} \Sigma^{-1} &= \begin{pmatrix} \mathbf{I}_r & -\Sigma_{1,2}\Sigma_2^{-1} \\ \mathbf{0} & \Sigma_2^{-1} \end{pmatrix} \\ \Rightarrow \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \Sigma_2^{-1}\Sigma_{1,2}^T & \mathbf{I}_{n-r} \end{pmatrix} \Sigma^{-1} &= \begin{pmatrix} \Sigma_{1|2}^{-1} & -\Sigma_{1|2}^{-1}\Sigma_{1,2}\Sigma_2^{-1} \\ \mathbf{0} & \Sigma_2^{-1} \end{pmatrix} \\ \Rightarrow \Sigma^{-1} &= \begin{pmatrix} \Sigma_{1|2}^{-1} & -\Sigma_{1|2}^{-1}\Sigma_{1,2}\Sigma_2^{-1} \\ -\Sigma_2^{-1}\Sigma_{1,2}^T\Sigma_{1|2}^{-1} & \Sigma_2^{-1}\Sigma_{1,2}^T\Sigma_{1|2}^{-1}\Sigma_{1,2}\Sigma_2^{-1} + \Sigma_2^{-1} \end{pmatrix}. \end{aligned}$$

Es folgt unter mehrfacher Anwendung des Distributivgesetzes

$$\begin{aligned} &(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu) - (\mathbf{z} - \mu_2)^T \Sigma_2^{-1}(\mathbf{z} - \mu_2) \\ &= (\mathbf{y} - \mu_1)^T \begin{pmatrix} \Sigma_{1|2}^{-1} & -\Sigma_{1|2}^{-1}\Sigma_{1,2}\Sigma_2^{-1} \\ -\Sigma_2^{-1}\Sigma_{1,2}^T\Sigma_{1|2}^{-1} & \Sigma_2^{-1}\Sigma_{1,2}^T\Sigma_{1|2}^{-1}\Sigma_{1,2}\Sigma_2^{-1} + \Sigma_2^{-1} \end{pmatrix} (\mathbf{y} - \mu_1) \\ &\quad - (\mathbf{z} - \mu_2)^T \Sigma_2^{-1}(\mathbf{z} - \mu_2) \\ &= \dots = (\mathbf{y} - \mu_1 - \Sigma_{1,2}\Sigma_2^{-1}(\mathbf{z} - \mu_2))^T \Sigma_{1|2}^{-1}(\mathbf{y} - \mu_1 - \Sigma_{1,2}\Sigma_2^{-1}(\mathbf{z} - \mu_2)) \\ &= (\mathbf{y} - \mu_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}})^T \Sigma_{1|2}^{-1}(\mathbf{y} - \mu_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}}). \end{aligned}$$

Wegen $\det \Sigma = \det \Sigma_2 \cdot \det \Sigma_{1|2}$ folgt

$$f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} | \mathbf{z}) = \frac{1}{\sqrt{\det \Sigma_{1|2} (2\pi)^r}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mu_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}})^T \Sigma_{1|2}^{-1}(\mathbf{y} - \mu_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}})\right),$$

also der Satz. ■

Bemerkung. Sei $(X, Y)^T \sim \mathcal{N}(\mu, \Sigma)$ mit

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \quad -1 < \rho < 1.$$

Dann ist

$$(Y | X = x) \sim \mathcal{N}\left(\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1), (1 - \rho^2)\sigma_2^2\right).$$

Speziell ist für $\sigma_1^2 = \sigma_2^2 = \sigma^2$ und $\mu_1 = \mu_2 = \mu$

$$(Y | X = x) \sim \mathcal{N}(\mu + \rho(x - \mu), (1 - \rho^2)\sigma^2), \quad \mathbb{E}(Y | X = x) = \mu + \rho(x - \mu),$$

also für $x \neq \mu$ und $|\rho| < 1$:

$$|\mathbb{E}(Y | X = x) - \mu| = |\rho| \cdot |x - \mu| < |x - \mu|.$$

Also ist unabhängig von der Korrelation, wenn $|\rho| < 1$, der Erwartungswert der abhängigen Variable näher am ursprünglichen Erwartungswert als das Ergebnis. Man spricht insbesondere bei $\rho > 0$ auch von **regression to the mean**.

3 Konvergenzarten der Stochastik

3.1 Konvergenz von Zufallsvariablen

In diesem Abschnitt diskutieren wir folgende Konvergenzbegriffe für Zufallsvariablen: fast sichere Konvergenz, stochastische Konvergenz, Konvergenz in L_p , sowie die Zusammenhänge zwischen diesen Konvergenzbegriffen.

Seien nachfolgend X_n, X, Y_n, Y Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$.

Fast sichere und stochastische Konvergenz

Definition 3.1. (Wdh.) Die Folge (X_n) konvergiert gegen X **fast sicher** oder kurz **fast sicher** ($X \rightarrow X$ \mathbb{P} -f.s.), falls

$$\mathbb{P}\left(\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\}\right) = 1.$$

Beachte, dass die betrachtete Menge als Schnitt

$$\left\{\liminf_{n \rightarrow \infty} X_n = X\right\} \cap \left\{\limsup_{n \rightarrow \infty} X_n = X\right\}$$

messbar ist. Für fast sichere Konvergenz genügt es, Konvergenz auf irgendeiner Menge mit Wahrscheinlichkeit 1 zu haben. Offenbar ist der fast sichere Limes fast sicher eindeutig bestimmt. Die **punktwise Konvergenz**, $X_n(\omega) \rightarrow X(\omega)$ für alle $\omega \in \Omega$, ist eine stärkere Bedingung und hinreichend für fast sichere Konvergenz.

Beispiel 3.2. Ist auf dem Wahrscheinlichkeitsraum $([0, 1], \mathcal{B}([0, 1]), \lambda_{[0,1]})$, mit dem Lebesgue-Maß λ , $X_n(\omega) = \omega^n$, so gilt $X_n \rightarrow 0$ λ -f.s., denn $X_n(\omega) \rightarrow 0$ für $\omega \neq 1$.

Definition 3.3. (Wdh.) Die Folge (X_n) konvergiert gegen X **stochastisch** oder in **\mathbb{P} -Wahrscheinlichkeit** ($X_n \xrightarrow{\mathbb{P}} X$), falls für alle $\varepsilon > 0$ gilt

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0. \quad (3.1)$$

Uns genügt die Definition für \mathbb{R}^d -wertige Zufallsvariablen. Für $d = 1$ bezeichnet $|\cdot|$ den Absolutbetrag, für $d > 1$ den Euklidischen Abstand. Die stochastische Konvergenz ist ein Spezialfall der **Konvergenz nach Maß** aus der Maßtheorie mit einem Wahrscheinlichkeitsmaß \mathbb{P} . Die fast sichere Konvergenz ist ein Spezialfall der **fast überall-Konvergenz** aus der Maßtheorie bezüglich einem Wahrscheinlichkeitsmaß \mathbb{P} .

Bemerkung. a. $X_n \xrightarrow{\mathbb{P}} X$ genau dann, falls für alle $\varepsilon > 0$ ein n_0 existiert, so dass für $n \geq n_0$ gilt

$$\mathbb{P}(|X_n - X| \geq \varepsilon) < \varepsilon.$$

b. Ist $X_n \xrightarrow{\mathbb{P}} X$ und $X_n \xrightarrow{\mathbb{P}} Y$, so folgt $X = Y$ \mathbb{P} -fast sicher (**fast sichere Eindeutigkeit des Grenzwertes**). Denn: Wegen $|X - Y| \leq |X_n - X| + |X_n - Y|$, gilt für beliebiges $n \in \mathbb{N}$ und $\varepsilon > 0$, dass

$$\{|X - Y| \geq \varepsilon\} \subseteq \{|X_n - X| \geq \varepsilon/2\} \cup \{|X_n - Y| \geq \varepsilon/2\},$$

also $\mathbb{P}(|X - Y| \geq \varepsilon) \leq \mathbb{P}(|X_n - X| \geq \varepsilon/2) + \mathbb{P}(|X_n - Y| \geq \varepsilon/2)$. Da die rechte Seite nach Voraussetzung für $n \rightarrow \infty$ gegen Null konvergiert, muss $\mathbb{P}(|X - Y| \geq \varepsilon) = 0$ gelten für beliebiges $\varepsilon > 0$, und damit

$$\mathbb{P}(X = Y) = 1.$$

c. Ist $X_n \xrightarrow{\mathbb{P}} X$, $Y_n \xrightarrow{\mathbb{P}} Y$, so gilt auch $X_n + Y_n \xrightarrow{\mathbb{P}} X + Y$. Dies folgt direkt aus der Ungleichung

$$\mathbb{P}(|X_n + Y_n - (X + Y)| > \varepsilon) \leq \mathbb{P}(|X_n - X| > \varepsilon/2) + \mathbb{P}(|Y_n - Y| > \varepsilon/2).$$

Es gibt folgende Verbindung zwischen dem fast sicheren und dem stochastischen Konvergenzbegriff:

Satz 3.4. Folgende Aussagen sind äquivalent:

1. $X_n \rightarrow X$ \mathbb{P} -fast sicher.
2. Für alle $\varepsilon > 0$ gilt

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \{|X_k - X| \geq \varepsilon\}\right) = 0.$$

3. Für alle $\varepsilon > 0$ gilt

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\bigcup_{k=n}^{\infty} \{|X_k - X| \geq \varepsilon\}\right) = 0.$$

■

Beweis. Es ist

$$\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \{|X_k - X| \geq \varepsilon\} = \{\omega \in \Omega : \forall n \geq 1 \exists k \geq n, \text{ s.d. } |X_k(\omega) - X(\omega)| \geq \varepsilon\} := A_{\varepsilon}.$$

Somit konvergiert $X_n(\omega) \rightarrow X(\omega)$ gerade für

$$\omega \in B := \bigcap_{\varepsilon > 0, \varepsilon \in \mathbb{Q}} A_{\varepsilon}^c = \left(\bigcup_{\varepsilon > 0, \varepsilon \in \mathbb{Q}} A_{\varepsilon} \right)^c.$$

2. \Rightarrow 1. Ist $\mathbb{P}(A_{\varepsilon}) = 0$, so auch das Maß der abzählbaren Vereinigung.

1. \Rightarrow 2. Da für $\omega \in B^c$ keine Konvergenz gilt, muss $\mathbb{P}(B^c) = 0$, also auch jedes $\mathbb{P}(A_{\varepsilon}) = 0$.
3. \Leftrightarrow 2. Folgt aus der Stetigkeit von oben für \mathbb{P} . ■

Korollar 3.5. (Wdh.) Konvergiert $X_n \rightarrow X$ \mathbb{P} -f.s., so konvergiert $X_n \xrightarrow{\mathbb{P}} X$.

fast sichere Konvergenz \Rightarrow stochastische Konvergenz

Beispiel 3.6. Stochastische Konvergenz impliziert umgekehrt nicht Konvergenz \mathbb{P} -f.s.: Definiere auf $([0, 1], \mathcal{B}([0, 1]), \lambda_{[0,1]})$, $X_n(\omega)$ durch

$$X_{2^n+k}(\omega) = \mathbf{1}_{[k/2^n, (k+1)/2^n]}(\omega), \quad n = 0, 1, 2, \dots, \quad k = 0, 1, \dots, 2^n - 1. \quad (3.2)$$

Dann ist $\lambda\{|X_{2^n+k}| > \varepsilon\} = 1/2^n$ (für $\varepsilon < 1$), und somit $X_n \xrightarrow{\lambda_{[0,1]}} 0$, aber für kein $\omega \in [0, 1]$ gilt $X_n(\omega) \rightarrow 0$.

Satz 3.7. Sei f eine stetige Abbildung.

1. Gilt $X_n \rightarrow X$ \mathbb{P} -f.s., so folgt $f(X_n) \rightarrow f(X)$ \mathbb{P} -f.s.
2. Gilt $X_n \xrightarrow{\mathbb{P}} X$, so folgt $f(X_n) \xrightarrow{\mathbb{P}} f(X)$.

■

Beweis. Dabei ist 1. klar. Wir beweisen 2. Zu $\varepsilon > 0$ existiert ein K , so dass $\mathbb{P}(|X| > K) < \varepsilon/3$. Es folgt mit der Dreiecksungleichung, Mengeninklusion und Subadditivität, dass

$$\mathbb{P}(|X_n| > K+1) \leq \mathbb{P}(|X| > K) + \mathbb{P}(|X_n - X| > 1) < \frac{\varepsilon}{3} + \frac{\varepsilon}{6},$$

für alle $n \geq n_0$, und ein hinreichend großes $n_0 \in \mathbb{N}$. Auf dem Kompaktum $[-K-1, K+1]$ ist f gleichmäßig stetig. Nach dem $\varepsilon - \delta$ -Kriterium existiert zu $\varepsilon > 0$ dann ein $\delta > 0$, sodass

$$|f(x) - f(y)| < \varepsilon, \forall x, y \text{ mit } |x - y| < \delta.$$

Aus der Mengeninklusion ergibt sich die Ungleichung

$$\begin{aligned} \mathbb{P}(|f(X_n) - f(X)| \geq \varepsilon) &\leq \mathbb{P}(|X_n| > K+1) + \mathbb{P}(|X_n - X| \geq \delta) + \mathbb{P}(|X| > K+1) \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{6} + \frac{\varepsilon}{6} + \frac{\varepsilon}{3} = \varepsilon, \end{aligned}$$

für alle $n \geq n_1$, und ein hinreichend großes $n_1 \in \mathbb{N}$. Daraus folgt die Behauptung. ■

Konvergenz in L_p (auch Konvergenz im p -ten Mittel)

Definition 3.8. Sei $p \geq 1$, und seien $X, X_n \in L_p(\mathbb{P})$, also $\mathbb{E}[|X_n|^p] < \infty$, $\mathbb{E}[|X|^p] < \infty$. Dann konvergiert die Folge (X_n) in L_p gegen X ($X_n \rightarrow X$ in L_p), falls

$$\mathbb{E}[|X_n - X|^p] \rightarrow 0.$$

Wir schreiben auch $X_n \xrightarrow{L_p} X$.

Dabei ist nach der Minkowski-Ungleichung (1.15) der Erwartungswert von $|X_n - X|^p$ endlich.

Satz 3.9. Sei $p \geq 1$, und $\mathbb{E}[|X_n|^p] < \infty$, $\mathbb{E}[|X|^p] < \infty$. Gilt $X_n \rightarrow X$ in L_p , so folgt $X_n \xrightarrow{\mathbb{P}} X$.

L_p -Konvergenz \Rightarrow stochastische Konvergenz

Beweis. Dies folgt direkt aus der Markov-Ungleichung: Für $\varepsilon > 0$ mit $g(x) = x^p$ folgt für $|X_n - X|$

$$\mathbb{P}(|X_n - X| \geq \varepsilon) \leq \frac{1}{\varepsilon^p} \mathbb{E}[|X_n - X|^p] \rightarrow 0.$$

Satz 3.10. Sei $p_1 < p_2$, dann gilt: $X_n \xrightarrow{L_{p_2}} X \Rightarrow X_n \xrightarrow{L_{p_1}} X$. ■

Beweis. Dies folgt direkt aus der Jensen-Ungleichung, da

$$(\mathbb{E}[|X_n - X|^{p_1}])^{p_2/p_1} \leq \mathbb{E}[|X_n - X|^{p_2}].$$

Beispiel 3.11. Im Allgemeinen impliziert weder die fast sichere Konvergenz die L_p -Konvergenz, noch umgekehrt. Dazu bemerken wir zunächst, dass die Limiten fast sicher übereinstimmen müssten, denn beide sind (falls existent) gleich dem Limes in stochastischer Konvergenz.

1. Betrachte die Zufallsvariablen in (3.2) Diese konvergieren für alle $p \geq 1$ gegen 0 in L_p , aber nirgends punktweise.

2. Für $X_n(x) = n^2 1_{[0,1/n]}(x)$ für $x \in [0, 1]$ mit dem Lebesgue Maß λ gilt $X_n \rightarrow 0$ λ -f.s., denn $X_n(x) \rightarrow 0$ für $x \neq 0$, aber $\mathbb{E}[|X_n|^p] \rightarrow \infty$, $p \geq 1$.

Definition 3.12. Eine Teilmenge $\mathcal{F} \subset \mathcal{L}_1(\Omega, \mathcal{A}, \mathbb{P})$, also eine Familie integrierbarer reellwertiger Zufallsvariablen, heißt **gleichgradig integrierbar**, falls

$$\lim_{n \rightarrow \infty} \sup_{Z \in \mathcal{F}} \int_{\{|Z| \geq n\}} |Z| d\mathbb{P} = 0.$$

Satz 3.13. Sei $(X_n)_n$ gleichgradig integrierbar sowie $X_n \xrightarrow{\mathbb{P}} X$, mit $X \in \mathcal{L}_1(\Omega, \mathcal{A}, \mathbb{P})$. Dann folgt dass $X_n \xrightarrow{L_1} X$. ■

Beweis. Für $k \in \mathbb{N}$ und $x \in \mathbb{R}$ sei

$$g_k(x) = -k \vee (x \wedge k) = \max(-k, \min(x, k)).$$

Für $\varepsilon > 0$ und $k, n \in \mathbb{N}$ gilt:

$$\begin{aligned} \|g_k(X_n) - g_k(X)\|_1 &= \mathbb{E}[|g_k(X_n) - g_k(X)|] \\ &= \mathbb{E}[|g_k(X_n) - g_k(X)|(\mathbf{1}\{|X_n - X| \leq \varepsilon\} + \mathbf{1}\{|X_n - X| > \varepsilon\})] \\ &\leq \varepsilon \int d\mathbb{P} + 2k \mathbb{P}(|X_n - X| > \varepsilon). \end{aligned}$$

Da $X_n \xrightarrow{\mathbb{P}} X$ folgt dass $\|g_k(X_n) - g_k(X)\|_1 \rightarrow 0$ für $n \rightarrow \infty$ und beliebiges k . Es gilt

$$\begin{aligned} \|X_n - X\|_1 &\leq \|X_n - g_k(X)\|_1 + \|g_k(X_n) - g_k(X)\|_1 + \|g_k(X) - X\|_1 \\ &\leq 2 \int_{\{|X_n| \geq k\}} |X_n| d\mathbb{P} + \|g_k(X_n) - g_k(X)\|_1 + 2 \int_{\{|X| \geq k\}} |X| d\mathbb{P}. \end{aligned}$$

Für $n \rightarrow \infty$ und danach $k \rightarrow \infty$ folgt aus der gleichgradigen Integrierbarkeit von (X_n) und der Integrierbarkeit von X , dass

$$\|X_n - X\|_1 \rightarrow 0.$$

■

Es gilt im obigen Satz auch die Äquivalenz. Definiert man analog zu Definition 3.12 eine p -gleichgradige Integrierbarkeit, $p \geq 2$, erhält man eine entsprechende ‘Brücke’ zwischen stochastischer und L_p -Konvergenz. Wir verweisen auf Kapitel 6.2 in Klenke (2008) für eine ausführlichere Behandlung der gleichgradigen Integrierbarkeit.

3.2 Konvergenz von Verteilungen

Die asymptotische Approximation von Wahrscheinlichkeitsverteilungen spielt in der Wahrscheinlichkeitstheorie und Statistik eine überragende Rolle. Beispiele sind der zentrale Grenzwertsatz oder die Poisson Approximation. Wir diskutieren in diesem Abschnitt die zugrundeliegende Theorie der schwachen Konvergenz von Wahrscheinlichkeitsmaßen auf \mathbb{R} bzw. \mathbb{R}^d . Die meisten Resultate werden nur über \mathbb{R} angegeben und/oder bewiesen, allerdings werden wir später gelegentlich von Versionen in \mathbb{R}^d Gebrauch machen. Ein zentrales Ergebnis ist die Charakterisierung von schwacher Konvergenz durch punktweise Konvergenz der charakteristischen Funktionen, der sogenannte Stetigkeitssatz von Lévy.

Definition 3.14. Eine Folge (F_n) von Verteilungsfunktionen auf \mathbb{R}^d konvergiert schwach gegen eine Verteilungsfunktion F auf \mathbb{R}^d , in Zeichen $F_n \xrightarrow{w} F$, falls F_n punktweise gegen F konvergiert an allen Stetigkeitsstellen von F , also falls für alle $x \in \mathbb{R}^d$ in denen F stetig ist, gilt

$$F_n(x) \rightarrow F(x), \quad n \rightarrow \infty.$$

Die Grenzfunktion F ist dann eindeutig bestimmt. Für $d = 1$ ist dies klar, da F nur abzählbar viele Unstetigkeitsstellen hat. Somit ist F zunächst bis auf an höchstens abzählbar vielen Stellen bestimmt, und wegen der Rechtsstetigkeit dann überall.

Definition 3.15. Sind μ_n, μ Wahrscheinlichkeitsmaße auf \mathbb{R}^d , so konvergiert μ_n schwach gegen μ , in Zeichen $\mu_n \xrightarrow{w} \mu$, falls $F_{\mu_n} \xrightarrow{w} F_\mu$. Sind schließlich X_n, X Zufallsvariablen, so konvergiert X_n in Verteilung gegen X , in Zeichen $X_n \xrightarrow{d} X$, falls $F_{X_n} \xrightarrow{w} F_X$.

Bemerkung. Für schwache Konvergenz müssen die Zufallsvariablen nicht auf einem gemeinsamen Wahrscheinlichkeitsraum definiert sein. Man kann daher auch von $X_n \xrightarrow{d} F$ sprechen für eine Verteilungsfunktion F .

Beispiele 3.16. 1. Sei X eine Zufallsvariable mit Verteilungsfunktion F , und sei $X_n = X + 1/n$. Dann hat X_n Verteilungsfunktion

$$F_n(x) = \mathbb{P}(X + 1/n \leq x) = F(x - 1/n),$$

also gilt $F_n(x) \rightarrow F(x-)$, $n \rightarrow \infty$. Insbesondere hat man keine punktweise Konvergenz in Unstetigkeitsstellen von F , aber dies wird bei schwacher Konvergenz auch nicht verlangt.

2. Seien $X_n \sim \text{Geo}(p_n)$ geometrisch verteilt, also $\mathbb{P}(X_n = k) = p_n(1 - p_n)^{k-1}$, $k \geq 1$, und $\mathbb{P}(X_n \geq k) = (1 - p_n)^{k-1}$, bzw. für $x > 0$

$$\mathbb{P}(X_n > x) = (1 - p_n)^{[x]},$$

wobei $[x]$ die kleinste natürliche Zahl größer oder gleich x ist. Für $p_n \rightarrow 0$ ist dann

$$\mathbb{P}(p_n X_n > x) = (1 - p_n)^{[x/p_n]} \rightarrow e^{-x}.$$

Also konvergiert $p_n X_n$ schwach gegen eine exponentialverteilte Zufallsvariable.

Falls die Grenzfunktion stetig ist, so hat man sogar bereits gleichmäßige Konvergenz: Gilt $F_n \xrightarrow{w} F$ und ist F stetig, so gilt sogar

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

Im nächsten Satz zeigen wir, dass bei Maßen mit Lebesgue-Dichten, die punktweise konvergieren, automatisch gleichmäßige Konvergenz folgt.

Satz 3.17 (Satz von Scheffé). Sind μ_n, μ Wahrscheinlichkeitsmaße auf $(\mathbb{R}^d, \mathcal{B}^d)$ mit Lebesgue-dichten f_n, f , und gilt $f_n \rightarrow f$ Lebesgue-fast überall, so folgt

$$\|f_n - f\|_1 = \int_{\mathbb{R}^d} |f_n(x) - f(x)| dx \rightarrow 0, \quad n \rightarrow \infty.$$

Insbesondere ist

$$\sup_{B \in \mathcal{B}^d} |\mu_n(B) - \mu(B)| \rightarrow 0,$$

und somit gilt insbesondere die schwache Konvergenz $\mu_n \xrightarrow{w} \mu$. ■

Beweis. Es ist $|f_n - f| = f_n - f + 2(f - f_n)^+$, und da $\int f_n = \int f = 1$, $0 \leq (f - f_n)^+ \leq f$, sowie $(f - f_n)^+ \rightarrow 0$ f.ü., folgt für $B \in \mathcal{B}^d$ mit dem Satz von der majorisierten Konvergenz

$$\left| \int_B f_n - \int_B f \right| \leq \int_{\mathbb{R}^d} |f_n - f| = 2 \int_{\mathbb{R}^d} (f - f_n)^+ \rightarrow 0.$$
■

Beispiel 3.18. (Ordnungsstatistiken) Sind U_1, \dots, U_n unabhängig und uniform verteilt auf $(0, 1)$, so heißen die der Größe nach geordneten Zufallsvariablen $U_{1:n} < \dots < U_{n:n}$ die n -ten Ordnungsstatistiken. Beachte, dass die U_i fast sicher alle nicht gleich sind. Man kann zeigen, dass $U_{k:n}$ folgende Verteilungsfunktion und Dichte hat

$$\begin{aligned} F_{k:n}(x) &= \sum_{j=k}^n \binom{n}{j} x^j (1-x)^{n-j}, \quad x \in (0, 1), \\ f_{k:n}(x) &= n \binom{n-1}{k-1} x^{k-1} (1-x)^{n-k}. \end{aligned}$$

Es gilt nämlich, dass

$$nF_n(t) = \sum_{i=1}^n \mathbf{1}_{(-\infty, t]}(X_i) \sim \text{Bin}(n, F(t)).$$

Da $\mathbb{P}(X_{(k)} \leq t) = \mathbb{P}(nF_n(t) \geq k)$, erhalten wir die Aussage über die Verteilungsfunktion. Gliedweise Differentiation der Summe, die Produktregel und eine Teleskopsumme ergeben

$$\begin{aligned} &\frac{\partial}{\partial x} \left(\sum_{m=k}^n \binom{n}{m} F^m(x) (1-F)^{n-m}(x) \right) \\ &= \sum_{m=k}^n \left(\binom{n}{m} m f(x) F^{m-1}(x) (1-F)^{n-m}(x) - \binom{n}{m} (n-m) f(x) F^m(x) (1-F)^{n-m-1}(x) \right) \\ &= \color{red} n f(x) F^{n-1}(x) + n(n-1) f(x) F^{n-2}(x) (1-F)(x) \\ &\quad - \color{red} n f(x) F^{n-1}(x) + \frac{1}{2} n(n-1)(n-2) f(x) F^{n-3}(x) (1-F)^2(x) - \color{red} n(n-1) f(x) F^{n-2}(x) (1-F)(x) \dots \\ &= k \binom{n}{k} f(x) F^{k-1}(x) (1-F)^{n-k}(x) \\ &= n \binom{n-1}{k-1} x^{k-1} (1-x)^{n-k}. \end{aligned}$$

Für die Dichte der zentralen Ordnungsstatistik $U_{(n+1):(2n+1)}$ gilt dann

$$g_n(x) = (2n+1) \binom{2n}{n} x^n (1-x)^n, \quad x \in (0, 1),$$

und für die Dichte von $V_n = 2(U_{(n+1):(2n+1)} - 1/2) \sqrt{2n}$

$$\begin{aligned} f_n(y) &= \frac{1}{2\sqrt{2n}} g_n\left(\frac{y}{2\sqrt{2n}} + \frac{1}{2}\right) \\ &= (2n+1) \binom{2n}{n} \frac{1}{2^{2n+1} \sqrt{2n}} \left(1 - \frac{y^2}{2n}\right)^n \mathbf{1}_{\{|y| \leq \sqrt{2n}\}}, \quad y \in \mathbb{R}. \end{aligned}$$

Mit Hilfe der Stirlingschen Formel $n! \sim \sqrt{2\pi n}(n/e)^n$, wobei $a_n \sim b_n$ für deterministische Folgen (a_n) und (b_n) bedeutet dass $a_n/b_n \rightarrow 1$, zeigt man $f_n(y) \rightarrow (\sqrt{2\pi})^{-1} e^{-y^2/2}$, und somit folgt die schwache Konvergenz von V_n gegen die Standardnormalverteilung.

Schwache und stochastische Konvergenz

Der folgende Satz besagt, dass stochastische Konvergenz stets Konvergenz in Verteilung impliziert.

Satz 3.19. (Wdh.) Sind $X_n, X : \Omega \rightarrow \mathbb{R}$ Zufallsvariablen mit $X_n \xrightarrow{\mathbb{P}} X$, so folgt $X_n \xrightarrow{d} X$. ■

Beweis. Für $x \in \mathbb{R}$ und $\varepsilon > 0$ gilt

$$\left(\{X \leq x - \varepsilon\} \cap \{|X - X_n| \leq \varepsilon\} \right) \subseteq \{X_n \leq x\} \subseteq \left(\{X \leq x + \varepsilon\} \cup \{|X - X_n| \geq \varepsilon\} \right)$$

und somit wegen der stochastischen Konvergenz

$$F(x - \varepsilon) \leq \liminf_{n \rightarrow \infty} F_n(x) \leq \limsup_{n \rightarrow \infty} F_n(x) \leq F(x + \varepsilon).$$

Ist x eine Stetigkeitsstelle von F , so folgt mit $\varepsilon \rightarrow 0$ die Konvergenz $F_n(x) \rightarrow F(x)$. ■

Für Konvergenz in Verteilung müssen die X_n, X nicht auf dem gleichen Wahrscheinlichkeitsraum definiert sein, daher lässt sich ohnehin allgemein nichts bzgl. stochastischer Konvergenz aussagen. Selbst wenn alle X_n und X auf dem gleichen Wahrscheinlichkeitsraum definiert sind, folgt aus schwacher Konvergenz nicht stochastische Konvergenz: Ist etwa $X \sim \mathcal{N}(0, 1)$ und $X_n = (-1)^n X$, so sind alle $X_n \sim \mathcal{N}(0, 1)$, und damit ist die Konvergenz in Verteilung klar, aber $\mathbb{P}(|X_n - X_{n+1}| \geq \varepsilon) = \mathbb{P}(X \geq \varepsilon/2 \text{ oder } X \leq -\varepsilon/2)$. Eine Zusammenfassung unserer Resultate über die Zusammenhänge zwischen den Konvergenzarten zeigt Abbildung 3.1. Für eine Konstante $a \in \mathbb{R}$ macht die Bedingung

$$\text{Für alle } \varepsilon > 0 \text{ gilt } P_{X_n}((a - \varepsilon, a + \varepsilon]^c) \rightarrow 0, \quad n \rightarrow \infty \quad (3.3)$$

auch Sinn, falls die X_n auf verschiedenen Wahrscheinlichkeitsräumen definiert sind, ansonsten entspricht sie $X_n \xrightarrow{\mathbb{P}} a$. Insbesondere impliziert (3.3) nach Satz 3.19 die schwache Konvergenz $X_n \xrightarrow{d} a$. Hier gilt auch die Umkehrung.

Satz 3.20. Gilt $X_n \xrightarrow{d} a$ für $a \in \mathbb{R}$, so folgt (3.3). ■

Beweis. Die Verteilungsfunktion der konstanten Zufallsvariable a ist $F_a(x) = \mathbf{1}_{x \geq a}$. Daher impliziert die schwache Konvergenz, dass für $\varepsilon > 0$ ein n_0 existiert, so dass

$$F_n(a - \varepsilon) \leq \varepsilon, \quad F_n(a + \varepsilon) \geq 1 - \varepsilon, \quad n \geq n_0.$$

Da $\{|X_n - a| > \varepsilon\} = (\{X_n < a - \varepsilon\} \cup \{X_n > a + \varepsilon\})$, folgt für $n \geq n_0$

$$\mathbb{P}(|X_n - a| > \varepsilon) \leq F_n(a - \varepsilon) + 1 - F_n(a + \varepsilon) \leq 2\varepsilon. \quad \blacksquare$$

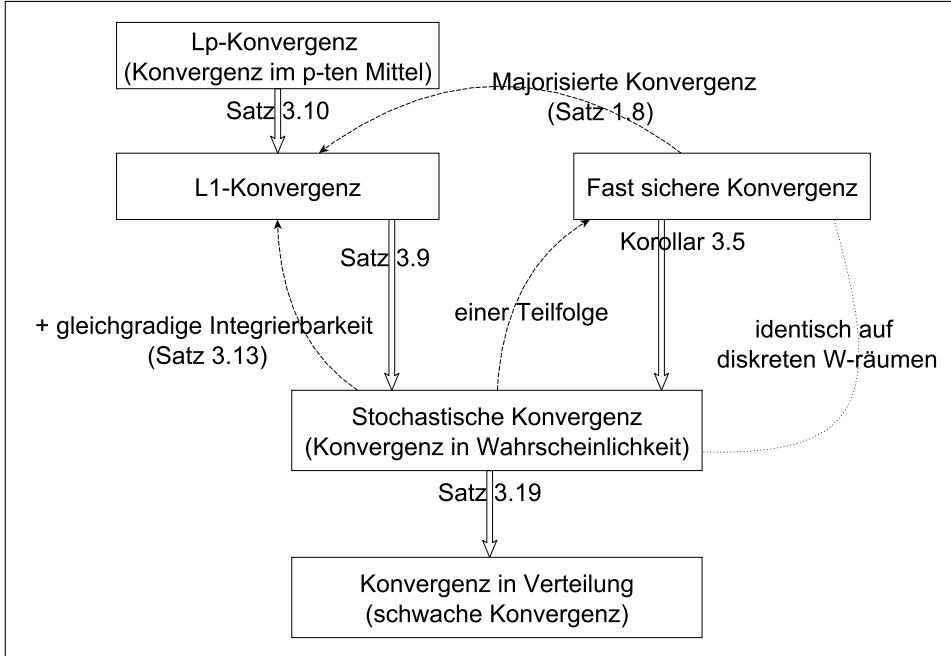


Abbildung 3.1: Die Konvergenzarten in der Stochastik mit Zusammenhängen.

Satz 3.21. Seien (X_n, Y_n) bivariate Zufallsvektoren und X eine Zufallsvariable mit $X_n \xrightarrow{d} X$ und $Y_n \xrightarrow{d} 0$.

1. Für $Z_n = X_n + Y_n$ (also $Z_n - X_n \xrightarrow{d} 0$) gilt dann $Z_n \xrightarrow{d} X$.
2. $Y_n X_n \xrightarrow{d} 0$. ■

Beweis. Zu 1.: Seien $x \in \mathbb{R}$, $\varepsilon > 0$. Dann ist

$$\left(\{X_n \leq x - \varepsilon\} \cap \{|Y_n| \leq \varepsilon\} \right) \subseteq \{Z_n \leq x\} \subseteq \left(\{X_n \leq x + \varepsilon\} \cup \{|Y_n| \geq \varepsilon\} \right).$$

Ist F die Verteilungsfunktion von X und $\varepsilon > 0$ derart, dass $x - \varepsilon$ und $x + \varepsilon$ Stetigkeitsstellen von F sind, so folgt

$$F(x - \varepsilon) \leq \liminf_{n \rightarrow \infty} \mathbb{P}(Z_n \leq x) \leq \limsup_{n \rightarrow \infty} \mathbb{P}(Z_n \leq x) \leq F(x + \varepsilon).$$

Ist x eine Stetigkeitsstelle von F , so folgt, indem $\varepsilon \rightarrow 0$ entlang von Stetigkeitsstellen $x + \varepsilon$, $x - \varepsilon$, die Behauptung.

Zu 2.: Für $\varepsilon > 0$ und $C > 0$ gilt stets

$$\{|X_n Y_n| \geq \varepsilon\} \subseteq \left(\{|X_n| > C\} \cup \{|Y_n| \geq \varepsilon/C\} \right).$$

Ist nun $C > 0$ Stetigkeitsstelle von F mit $F(-C) + 1 - F(C) < \varepsilon$. Wegen der schwachen Konvergenz existiert n_0 , so dass für $n \geq n_0$

$$\mathbb{P}(-C \leq X_n \leq C) \geq 1 - 2\varepsilon, \quad n \geq n_0.$$

Für ein $n_1 \geq n_0$ gilt

$$\mathbb{P}(|Y_n| \geq \varepsilon/C) \leq \varepsilon, \quad n \geq n_1.$$

Dann folgt für $n \geq n_1$ dass $\mathbb{P}(|X_n Y_n| \geq \varepsilon) \leq 3\varepsilon$, also die Behauptung. ■

Implikationen und äquivalente Bedingungen

Wir beginnen mit folgender Beobachtung: Sind X_n, X sowie Y_n, Y reellwertige Zufallsvariablen mit $Y_n \stackrel{d}{=} X_n$, $Y \stackrel{d}{=} X$, sowie $X_n \xrightarrow{d} X$, so folgt $Y_n \xrightarrow{d} Y$. Dabei können sich die gemeinsamen Verteilungen der Zufallsvariablen (Y_n) bzw. (X_n) stark unterscheiden, nur die univariaten Marginalverteilungen müssen gleich sein.

Satz 3.22 (Skorochod). Seien X_n, X reellwertige Zufallsvariablen mit $X_n \xrightarrow{d} X$. Dann existieren auf dem Wahrscheinlichkeitsraum $((0, 1), \mathcal{B}(0, 1), \lambda_{(0,1)})$ Zufallsvariablen $Y_n, Y : (0, 1) \rightarrow \mathbb{R}$ mit $Y_n \stackrel{d}{=} X_n$, $Y \stackrel{d}{=} X$ sowie $Y_n(\omega) \rightarrow Y(\omega)$ für λ -fast alle $\omega \in (0, 1)$. ■

Natürlich gilt in der Situation des Satzes auch $Y_n \xrightarrow{d} Y$, aber es gilt für diese Folge noch mehr.

Beweis. Sei F_n die Verteilungsfunktion von X_n und F die von X . Für $\omega \in (0, 1)$ sei $Y_n(\omega) = F_n^{-1}(\omega)$, $Y(\omega) = F^{-1}(\omega)$. Dann gilt für $\omega \in (0, 1)$ und $x \in \mathbb{R}$:

$$\omega \leq F(x) \Leftrightarrow Y(\omega) \leq x, \quad (3.4)$$

analog für F_n, Y_n .

Sei $\omega \in (0, 1)$, $\varepsilon > 0$, wähle x als Stetigkeitsstelle von F mit (möglich, da $\omega \neq 0, 1$)

$$Y(\omega) - \varepsilon < x < Y(\omega).$$

Nach (3.4) folgt $F(x) < \omega$, somit wegen $F_n(x) \rightarrow F(x)$ auch $F_n(x) < \omega$ für große n , und daher wieder mit (3.4) für große n

$$Y(\omega) - \varepsilon < x < Y_n(\omega), \quad \text{somit } \liminf_{n \rightarrow \infty} Y_n(\omega) \geq Y(\omega),$$

da $\varepsilon > 0$ beliebig war.

Ist $\omega < \omega' < 1$, und $y \in \mathbb{R}$ mit

$$Y(\omega) \leq Y(\omega') < y < Y(\omega') + \varepsilon,$$

und ist y Stetigkeitsstelle von F , so folgt mit (3.4) $\omega < \omega' \leq F(y)$, also für große n auch $\omega < F_n(y)$, daher

$$Y_n(\omega) \leq y < Y(\omega') + \varepsilon, \quad \text{somit } \limsup_{n \rightarrow \infty} Y_n(\omega) \leq Y(\omega').$$

Ist ω Stetigkeitsstelle von Y , so folgt $Y_n(\omega) \rightarrow Y(\omega)$. Y ist monoton wachsend, und hat daher höchstens abzählbar viele Unstetigkeitsstellen, insbesondere gilt $Y_n \rightarrow Y$ λ -fast sicher. ■

Wir nutzen diesen Darstellungssatz von Skorochod für die Beweise der folgenden Sätze.

Satz 3.23 (Continuous Mapping Theorem). Seien X_n, X reellwertige Zufallsvariablen mit $X_n \xrightarrow{d} X$, und sei $f : \mathbb{R} \rightarrow \mathbb{R}$ messbar mit $P_X(D_f) = 0$, wobei D_f die Menge der Unstetigkeitsstellen von f sei (insbesondere ist D_f messbar). Dann folgt

$$f(X_n) \xrightarrow{d} f(X).$$

Beweis. Seien Y_n, Y wie im Satz von Skorochod. Nach der einleitenden Bemerkung genügt es, $f(Y_n) \xrightarrow{d}$

$f(Y)$ zu zeigen. Dazu bemerken wir, dass

$$f(Y_n(\omega)) \rightarrow f(Y(\omega)) \quad \text{für } \{\omega : Y(\omega) \notin D_f\} \cap \{\omega : Y_n(\omega) \rightarrow Y(\omega)\}.$$

Da

$$\mathbb{P}(\{\omega : Y(\omega) \notin D_f\}) = P_X(D_f^c) = 1,$$

folgt $f(Y_n) \rightarrow f(Y)$ fast sicher und daher insbesondere in Verteilung (Satz 3.19).

D_f messbar: Für $\varepsilon, \delta > 0$ sei

$$A(\varepsilon, \delta) = \{x \in \mathbb{R} : \exists y, z \in \mathbb{R}, |x - y| < \delta, |x - z| < \delta, |f(y) - f(z)| \geq \varepsilon\}.$$

Dann ist $A(\varepsilon, \delta)$ offen, denn ist $x \in A(\varepsilon, \delta)$ mit y, z wie in der Bedingung, so erfüllen diese y, z die Bedingungen auch für \tilde{x} mit $|\tilde{x} - x| < \delta - \max(|x - y|, |x - z|)$. Somit ist $A(\varepsilon, \delta)$ messbar und

$$D_f = \bigcup_{n \geq 1} \bigcap_{m \geq 1} A(1/n, 1/m),$$

wie man durch Übergang zum Komplement, welches man mit den Stetigkeitsstellen identifiziert, sieht. ■

Folgendes Resultat ist als **Lemma von Slutsky** bekannt.

Korollar 3.24. Seien (X_n, A_n, B_n) dreidimensionale Zufallsvektoren und X eine reellwertige Zufallsvariable, $a, b \in \mathbb{R}$ mit $X_n \xrightarrow{d} X, A_n \xrightarrow{d} a, B_n \xrightarrow{d} b$, so folgt

$$A_n X_n + B_n \xrightarrow{d} aX + b.$$

Beweis. Zunächst gilt mit obigem Satz $aX_n + b \xrightarrow{d} aX + b$. Nach Satz 3.21 gilt

$$A_n X_n + B_n - (aX_n + b) \xrightarrow{d} 0,$$

und somit folgt die Behauptung mit Satz 3.19. ■

Satz 3.25 (Portmanteau–Theorem). Für reellwertige Zufallsvariablen X_n, X sind äquivalent:

1. $X_n \xrightarrow{d} X$.

2. Für alle $f \in C_b(\mathbb{R})$ gilt

$$\mathbb{E}[f(X_n)] \rightarrow \mathbb{E}[f(X)].$$

3. Für alle $U \subset \mathbb{R}$ offen gilt

$$\liminf_{n \rightarrow \infty} \mathbb{P}(X_n \in U) \geq \mathbb{P}(X \in U).$$

4. Für alle $K \subset \mathbb{R}$ abgeschlossen gilt

$$\limsup_{n \rightarrow \infty} \mathbb{P}(X_n \in K) \leq \mathbb{P}(X \in K).$$

5. Für alle $A \subset \mathbb{R}$ messbar mit $\mathbb{P}(X \in \partial A) = 0$ gilt

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A).$$

Beweis. 1. \Rightarrow 2.: Seien Y_n, Y wie im Satz von Skorochod. Dann gilt $f(Y_n) \rightarrow f(Y)$ für fast alle ω , und die Behauptung folgt mit dem Satz von der beschränkten Konvergenz.

2. \Rightarrow 1.: Für $x \in \mathbb{R}$, $\varepsilon > 0$ sei

$$g_{x,\varepsilon}(y) = \begin{cases} 0, & y \geq x + \varepsilon, \\ 1, & y \leq x, \\ (x + \varepsilon - y)/\varepsilon, & x \leq y \leq x + \varepsilon. \end{cases}$$

Dann ist $g_{\varepsilon,x}$ stetig und

$$\limsup_n \mathbb{P}(X_n \leq x) \leq \limsup_n \mathbb{E}[g_{\varepsilon,x}(X_n)] = \mathbb{E}[g_{\varepsilon,x}(X)] \leq \mathbb{P}(X \leq x + \varepsilon).$$

Mit $\varepsilon \rightarrow 0$ folgt mit der Stetigkeit von oben

$$\limsup_n \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x).$$

Analog ist

$$\liminf_n \mathbb{P}(X_n \leq x) \geq \liminf_n \mathbb{E}[g_{\varepsilon,x-\varepsilon}(X_n)] = \mathbb{E}[g_{\varepsilon,x-\varepsilon}(X)] \geq \mathbb{P}(X \leq x - \varepsilon)$$

also

$$\liminf_n \mathbb{P}(X_n \leq x) \geq \mathbb{P}(X < x).$$

Ist x Stetigkeitsstelle der Verteilungsfunktion von X , so folgt die Konvergenz.

1. \Rightarrow 3.: Seien Y_n, Y wie im Satz von Skorochod. Da U offen, ist

$$\liminf_n \mathbf{1}_U(Y_n) \geq \mathbf{1}_U(Y),$$

da für $Y(\omega) \in U$, $Y_n(\omega) \in U$ für große n folgt. Mit dem Lemma von Fatou folgt

$$\liminf_n \mathbb{P}(Y_n \in U) \geq \mathbb{E}[\liminf_n \mathbf{1}_U(Y_n)] \geq \mathbb{P}(Y \in U).$$

3. \Leftrightarrow 4.: Durch Übergang zum Komplement.

3+4. \Rightarrow 5.: Setze $K = \bar{A}$ und $U = \text{int}(A)$, das Innere von A . Dann ist wegen $P_X(\partial A) = 0$

$$\mathbb{P}(X \in U) = \mathbb{P}(X \in K) = \mathbb{P}(X \in A).$$

Somit nach 3.

$$\liminf_n \mathbb{P}(X_n \in A) \geq \liminf_n \mathbb{P}(X_n \in U) \geq \mathbb{P}(X \in U) = \mathbb{P}(X \in A),$$

und nach 4.

$$\limsup_n \mathbb{P}(X_n \in A) \leq \limsup_n \mathbb{P}(X_n \in K) \leq \mathbb{P}(X \in K) = \mathbb{P}(X \in A).$$

Daher folgt die behauptete Gleichheit.

5. \Rightarrow 1: Wähle A von der Form $A = (-\infty, x]$, für eine Stetigkeitsstelle x . ■

Auch hier muss in 1. nicht vorausgesetzt werden, dass die (X_n) auf demselben Wahrscheinlichkeitsraum definiert sind. Man kann Erwartungswerte und Wahrscheinlichkeiten in 2.-5. dann über die Verteilungen schreiben, was hier zur Vereinfachung nicht gemacht wurde.

3.3 Straffheit und schwache Konvergenz

Satz 3.26 (Helly's Selection Theorem). Sei (F_n) eine Folge von Verteilungsfunktionen. Dann existiert eine Teilfolge (F_{n_k}) sowie eine monoton wachsende, rechtsseitig stetige Funktion F mit $F_{n_k}(x) \rightarrow F(x)$, $k \rightarrow \infty$, für alle Stetigkeitsstellen x von F . ■

Lemma 3.27 (Diagonalfolgenargument). Für alle $m \geq 1$ sei $(a_{m,n})_{n \geq 1}$ eine beschränkte Folge von reellen Zahlen. Dann existiert eine Teilfolge (n_k) sowie eine Folge $(a_m)_{m \geq 1}$, so dass

$$a_{m,n_k} \rightarrow a_m, \quad k \rightarrow \infty \quad \text{für alle } m \geq 1.$$

Beweis. Nach Bolzano-Weierstrass existiert eine Teilfolge $n_{1,k}$ sowie $a_1 \in \mathbb{R}$ mit

$$a_{1,n_{1,k}} \rightarrow a_1, \quad k \rightarrow \infty.$$

Dann existiert eine Teilfolge $n_{2,k}$ von $n_{1,k}$ sowie $a_2 \in \mathbb{R}$ mit

$$a_{2,n_{2,k}} \rightarrow a_2, \quad k \rightarrow \infty.$$

Induktiv konstruiert man sukzessive Teilfolgen $n_{m,k}$ und a_m mit

$$a_{m,n_{m,k}} \rightarrow a_m, \quad k \rightarrow \infty.$$

Setzt man $n_k = n_{k,k}$, so folgt für alle m

$$a_{m,n_k} \rightarrow a_m, \quad k \rightarrow \infty,$$

da man für $k \geq m$ entlang einer Teilfolge der konvergenten Teilfolge $a_{m,n_{m,k}}$ geht. ■

Beweis. [von Satz 3.26] Mit dem Diagonalfolgenargument existiert eine Teilfolge n_k und ein $G : \mathbb{Q} \rightarrow [0, 1]$ mit $G(r) = \lim_k F_{n_k}(r)$ für alle $r \in \mathbb{Q}$. Setze

$$F(x) = \inf\{G(r), r > x\}.$$

Dann gilt

1. F ist monoton wachsend.

2. F ist rechtsseitig stetig: Sei $x \in \mathbb{R}$, $\varepsilon > 0$, wähle $r \in \mathbb{Q}$, $r > x$ mit $G(r) < F(x) + \varepsilon$. Für $y \leq x < r$ ist dann

$$F(x) \leq F(y) \leq G(r) < F(x) + \varepsilon,$$

dies impliziert die rechtsseitige Stetigkeit.

3. $F_{n_k}(x) \rightarrow F(x)$, $k \rightarrow \infty$, für alle Stetigkeitsstellen x von F : Für $\varepsilon > 0$, x Stetigkeitsstelle von F wähle $r_1, r_2, s \in \mathbb{Q}$ mit

$$r_1 < r_2 < x < s, \quad F(x) - \varepsilon < F(r_1) \leq F(r_2) \leq F(x) \leq G(s) < F(x) + \varepsilon.$$

Da $F_{n_k}(s) \rightarrow G(s)$ und $F_{n_k}(r_2) \rightarrow G(r_2) \geq F(r_1)$, folgt für große k mit der Monotonie der F_n

$$F(x) - \varepsilon < F_{n_k}(r_2) \leq F_{n_k}(x) \leq F_{n_k}(s) < F(x) + \varepsilon,$$

also die Behauptung. ■

Beispiel 3.28. Die Limesfunktion F muss keine Verteilungsfunktion sein: Ist $F_n(x) = \mathbf{1}_{[n, \infty)}(x)$, dann muss $F(x) = 0$ für alle $x \in \mathbb{R}$, falls $F_n(x) = \mathbf{1}_{[-n, \infty)}(x)$, dann $F(x) = 1$.

Bemerkung. Die Funktion G im Beweis ist monoton (da für $r < s$, $r, s \in \mathbb{Q}$ gilt $F_{n_k}(r) \leq F_{n_k}(s)$ für alle k wegen der Monotonie der F_{n_k} , und dies überträgt sich auf den Limes). Jedoch muss G nicht rechtsseitig stetig sein, etwa $F_n(x) = \mathbf{1}_{(-\infty, 1-1/n]}(x) \rightarrow \mathbf{1}_{(-\infty, 1)}(x)$, $n \rightarrow \infty$ für alle $x \in \mathbb{R}$.

Damit die Grenzfunktion wieder eine Verteilungsfunktion ist, muss die Folge zusätzliche Eigenschaften erfüllen.

Definition 3.29. Eine Familie $(\mu_i)_{i \in I}$ von Wahrscheinlichkeitsmaßen auf $(\mathbb{R}, \mathcal{B})$ heißt **straff**, falls für alle $\varepsilon > 0$ eine kompakte Menge $K \subset \mathbb{R}$ existiert, so dass

$$\mu_i(K) \geq 1 - \varepsilon \quad \forall i \in I.$$

Analog nennt man eine Familie von Zufallsvariablen $(X_i)_{i \in I}$ straff, wenn die zugehörigen Verteilungen straff sind, und eine Familie von Verteilungsfunktionen (F_i) , wenn (μ_{F_i}) straff sind.

Bemerkung. 1. Äquivalent zu $(\mu_i)_{i \in I}$ straff: Für alle $\varepsilon > 0$ existieren $a < b$ mit (F_i Verteilungsfunktion von μ_i)

$$\mu_i(a, b] = F_i(b) - F_i(a) \geq 1 - \varepsilon \quad \forall i \in I.$$

2. Jede endliche Familie von Wahrscheinlichkeitsmaßen ist straff: Zunächst ist ein einzelnes Wahrscheinlichkeitsmaß straff, da \mathbb{R} sich abzählbar monoton wachsend durch kompakte Mengen ausschöpfen lässt (dann mit Stetigkeit von unten). Da die Vereinigung endlich vieler kompakter Mengen wieder kompakt ist, bleiben endlich viele Wahrscheinlichkeitsmaße straff.

3. Ist $(\mu_i)_{i \in I}$ eine Familie von Wahrscheinlichkeitsmaßen auf \mathbb{R} und $I = I_1 \cup I_2$, und sind $(\mu_i)_{i \in I_1}$ und $(\mu_i)_{i \in I_2}$ straff, dann auch $(\mu_i)_{i \in I}$. Insbesondere genügt es für die Straffheit einer Folge $(\mu_n)_{n \geq 1}$ die Straffheit von $(\mu_n)_{n \geq n_0}$ für ein $n_0 \geq 1$ nachzuweisen. 4. Allgemein wird Straffheit von endlichen Maßen analog auf metrischen Räumen (\mathcal{X}, d) , mit deren Borelscher σ -Algebra $\mathcal{B}(\mathcal{X})$, definiert.

Lemma 3.30. Sind (μ_n) , μ Wahrscheinlichkeitsmaße mit $\mu_n \xrightarrow{w} \mu$, so ist $(\mu_n)_{n \geq 1}$ straff.

Beweis. Seien F_n und F die Verteilungsfunktionen von μ_n und μ . Sei $\varepsilon > 0$, und $a < b$ Stetigkeitsstellen von F mit $F(b) - F(a) \geq 1 - \varepsilon/2$. Da $F_n(a) \rightarrow F(a)$, $F_n(b) \rightarrow F(b)$, existiert n_0 , so dass $F_n(b) \geq F(b) - \varepsilon/4$, $F_n(a) \leq F(a) + \varepsilon/4$, und daher $F_n(b) - F_n(a) \geq 1 - \varepsilon$, $n \geq n_0$. Mit obigen Bemerkungen folgt die Behauptung. ■

Satz 3.31. Eine Familie $(\mu_i)_{i \in I}$ von Wahrscheinlichkeitsmaßen auf \mathbb{R} ist genau dann straff, wenn jede Folge $(\mu_{i_n})_{n \geq 1}$ aus $(\mu_i)_{i \in I}$ (also $i_1, i_2, \dots \in I$) eine schwach-konvergente Teilfolge hat.

Straff = relativ folgenkompakt bzgl. schwacher Konvergenz.

Beweis. Straff \Rightarrow relativ folgenkompakt: Sei $\mu_{i_n} =: \mu_n$ eine Folge aus (μ_i) , und seien F_n die Verteilungsfunktionen von μ_n . Nach dem Satz 3.26 von Helly existiert eine Teilfolge (F_{n_k}) und eine monoton wachsende, rechtsseitig stetige Funktion F mit $F_{n_k}(x) \rightarrow F(x)$ in Stetigkeitsstellen von F . Sei $\mu := \mu_F$ das zu F gehörige endliche Maß (also $\mu_F(a, b] = F(b) - F(a)$). Es bleibt zu zeigen, dass μ ein Wahrscheinlichkeitsmaß ist. Wegen $0 \leq F \leq 1$ ist $\mu(\mathbb{R}) \leq 1$. Da $(\mu_{n_k})_{k \geq 1}$ straff ist, existieren für $\varepsilon > 0$ $a < b$, o.E.d.A. Stetigkeitsstellen von F , mit $\mu_{n_k}(a, b] = F_{n_k}(b) - F_{n_k}(a) \geq 1 - \varepsilon$. Da $F_{n_k}(a) \rightarrow F(a)$, $F_{n_k}(b) \rightarrow F(b)$, $k \rightarrow \infty$, folgt auch $\mu(\mathbb{R}) \geq \mu(a, b] \geq 1 - \varepsilon$. Da $\varepsilon > 0$ beliebig, folgt $\mu(\mathbb{R}) = 1$.

Relativ folgenkompakt \Rightarrow straff: Angenommen, $(\mu_i)_{i \in I}$ ist nicht straff. Dann existiert $\varepsilon > 0$, so dass für alle n ein i_n existiert mit

$$\mu_{i_n}(-n, n] \leq 1 - \varepsilon.$$

Dann kann keine Teilfolge von $(\mu_{i_n})_{n \geq 1}$ straff sein, und insbesondere nicht schwach konvergent, vgl. Lemma 3.30. ■

Satz 3.32. Für Wahrscheinlichkeitsmaße $(\mu_n), \mu$ sind äquivalent:

1. $\mu_n \xrightarrow{w} \mu$.
2. a. (μ_n) ist straff
und
- b. Für alle schwach-konvergenten Teilfolgen (μ_{n_k}) von (μ_n) ist der schwache Limes notwendigerweise μ : $\mu_{n_k} \xrightarrow{w} \mu, k \rightarrow \infty$. ■

Beweis. 1. \Rightarrow 2.a: Lemma 3.30.

1. \Rightarrow 2.b: Jede Teilfolge einer schwach-konvergenten Folge konvergiert selbst schwach gegen denselben Limes, und dieser ist eindeutig bestimmt.

2.a. + 2.b. \Rightarrow 1.: Angenommen, μ_n konvergiert nicht schwach gegen μ . Dann existiert $\varepsilon > 0, f \in C_b(\mathbb{R})$, und eine Teilfolge μ_{n_k} mit

$$\left| \int_{\mathbb{R}} f d\mu_{n_k} - \int_{\mathbb{R}} f d\mu \right| \geq \varepsilon, \quad k \geq 1.$$

Da (μ_{n_k}) straff, existiert nach Satz 3.31 eine weitere schwach-konvergente Teilfolge, die nach Annahme gegen μ schwach konvergieren muss, welches aber obiger Ungleichung widerspricht. ■

Abschließend geben wir noch ein Kriterium für Straffheit über die gleichmäßige Beschränktheit geeigneter Momente an.

Satz 3.33. Existiert für eine Familie $(\mu_i)_{i \in I}$ von Wahrscheinlichkeitsmaßen auf \mathbb{R} eine messbare Funktion $\phi \geq 0$ mit $\phi(x) \rightarrow \infty, |x| \rightarrow \infty$, und ein $C > 0$ mit

$$\sup_{i \in I} \int_{\mathbb{R}} \phi d\mu_i \leq C < \infty,$$

dann ist $(\mu_i)_{i \in I}$ straff. ■

Beweis. Da

$$\mu_i[-M, M] \subseteq \inf_{|x| > M} \phi(x) \leq \int_{|x| > M} \phi d\mu_i \leq \int_{\mathbb{R}} \phi d\mu_i \leq C$$

für alle i , und $\inf_{|x| > M} \phi(x) \rightarrow \infty, M \rightarrow \infty$, folgt mit

$$\mu_i[-M, M] \leq \frac{C}{\inf_{|x| > M} \phi(x)}, \quad i \in I,$$

für hinreichend großes M die Behauptung. ■

Schwache Konvergenz und charakteristische Funktionen

Satz 3.34 (Stetigkeitssatz von Lévy). Seien μ_n, μ Wahrscheinlichkeitsmaße auf \mathbb{R}^d mit charakteristischen Funktionen $\varphi_{\mu_n} = \varphi_n$ und $\varphi_{\mu} = \varphi$. Dann sind äquivalent:

1. $\mu_n \xrightarrow{w} \mu$.
2. $\varphi_n(t) \rightarrow \varphi(t), n \rightarrow \infty$, für alle $t \in \mathbb{R}^d$. ■

Beweis. 1. \Rightarrow 2. Für festes $t \in \mathbb{R}^d$ ist $f_t(x) = e^{i \langle t, x \rangle}$ stetig und beschränkt. Nach Satz 3.25 folgt

$$\varphi_n(t) = \int_{\mathbb{R}^d} f_t(x) d\mu_n(x) \rightarrow \int_{\mathbb{R}^d} f_t(x) d\mu(x) = \varphi(t).$$

2. \Rightarrow 1. $d = 1$: Wir zeigen zunächst:

Behauptung: Die Wahrscheinlichkeitsmaße (μ_n) sind straff.

Da

$$\int_{-u}^u (1 - e^{itx}) dt = 2u - \frac{2 \sin(ux)}{x},$$

berechnen wir mit Fubini

$$\begin{aligned} \frac{1}{u} \int_{-u}^u (1 - \varphi_n(t)) dt &= \int_{\mathbb{R}} \frac{1}{u} \int_{-u}^u (1 - e^{itx}) dt d\mu_n(x) \\ &= 2 \int_{\mathbb{R}} \left(1 - \frac{\sin(ux)}{ux}\right) d\mu_n(x). \end{aligned}$$

Da $|\sin x| \leq |x|$, und $|\sin t| \leq 1$, folgt

$$\begin{aligned} \frac{1}{u} \int_{-u}^u (1 - \varphi_n(t)) dt &\geq 2 \int_{\{|x| \geq 2/u\}} \left(1 - \frac{\sin(ux)}{ux}\right) d\mu_n(x) \\ &\geq 2 \int_{\{|x| \geq 2/u\}} \left(1 - \left|\frac{\sin(ux)}{ux}\right|\right) d\mu_n(x) \\ &\geq 2 \int_{\{|x| \geq 2/u\}} \left(1 - \left|\frac{1}{ux}\right|\right) d\mu_n(x) \\ &\geq \mu_n(\{x : |x| \geq 2/u\}). \end{aligned} \tag{3.5}$$

Da $\varphi(t) \rightarrow \varphi(0) = 1$, $t \rightarrow 0$, folgt

$$\frac{1}{u} \int_{-u}^u (1 - \varphi(t)) dt \rightarrow 0, \quad u \rightarrow 0.$$

Gegeben ε , wähle u_0 , so dass die linke Seite $< \varepsilon$. Da $\varphi_n(t) \rightarrow \varphi(t)$, und $|\varphi_n(t)| \leq 1$, folgt mit dem Satz von der majorisierten Konvergenz

$$\frac{1}{u_0} \int_{-u_0}^{u_0} (1 - \varphi_n(t)) dt \rightarrow \frac{1}{u_0} \int_{-u_0}^{u_0} (1 - \varphi(t)) dt < \varepsilon.$$

und somit mit (3.5)

$$\mu_n(\{x : |x| \geq 2/u_0\}) < \varepsilon,$$

für n groß, also die Straffheit.

Nun zeigen wir noch 2.b. von Satz 3.32, dann folgt die Aussage des Satzes. Ist dazu (μ_{n_k}) eine schwach konvergente Teilfolge, etwa $\mu_{n_k} \xrightarrow{w} \nu$, so folgt $\varphi_{n_k}(t) \rightarrow \varphi_\nu(t)$, $t \in \mathbb{R}$, also $\varphi_\nu(t) = \varphi_\mu(t)$ für alle t und somit $\nu = \mu$.

$d \geq 1$: Sind X_n , X Zufallsvektoren mit $X_n \sim \mu_n$, $X \sim \mu$, und ist $\theta \in \mathbb{R}^d$ fest, so hat $\theta^T X_n$ die charakteristische Funktion

$$\varphi_{\theta^T X_n}(s) = \varphi_n(s\theta), \quad s \in \mathbb{R}.$$

Offenbar gilt dann $\varphi_{\theta^T X_n}(s) \rightarrow \varphi_{\theta^T X}(s)$, und somit ist $\theta^T X_n$ straff für alle $\theta \in \mathbb{R}^d$. Mit dem folgenden Lemma 3.35 folgt, dass auch (X_n) straff ist. Die Konvergenz ergibt sich dann wie im Fall $d = 1$. ■

Lemma 3.35. Sind X_n Zufallsvektoren in \mathbb{R}^d , so sind (X_n) genau dann straff, wenn $(\theta^T X_n)$ straff sind für alle $\theta \in \mathbb{R}^d$.

Beweis. 1. Angenommen, (X_n) ist straff. Sei $\varepsilon > 0$, $\theta \in \mathbb{R}^d$, und $K \subset \mathbb{R}^d$ kompakt mit $\mathbb{P}(X_n \in K) \geq 1 - \varepsilon$ für alle n . Dann ist auch $\theta(K) \subset \mathbb{R}$ kompakt (Bild einer kompakten Menge unter einer stetigen Abbildung ist kompakt), und $\{\theta^T X_n \in \theta(K)\} \supseteq \{X_n \in K\}$. Somit folgt

$$\mathbb{P}(\theta^T X_n \in \theta(K)) \geq \mathbb{P}(X_n \in K) \geq 1 - \varepsilon.$$

2. Angenommen, $(\theta^T X_n)$ ist straff für alle $\theta \in \mathbb{R}^d$. Für $\varepsilon > 0$ und $\theta = e_i$ (i -te Einheitsvektor), $i = 1, \dots, d$, existiert dann $C > 0$ mit $(e_i^T X_n = X_{i,n})$

$$\mathbb{P}(-C \leq X_{i,n} \leq C) \geq 1 - \varepsilon/d, \quad \text{für alle } i = 1, \dots, d, n \geq 1.$$

Dann folgt

$$\mathbb{P}(X_n \in [-C, C]^d) \geq 1 - \varepsilon, \quad \text{für alle } n \geq 1. \quad \blacksquare$$

Aus dem Beweis von Satz 3.34 entnehmen wir noch folgendes Korollar.

Korollar 3.36. Seien μ_n Wahrscheinlichkeitsmaße auf \mathbb{R}^d mit charakteristischen Funktionen $\varphi_{\mu_n} = \varphi_n$ und existiert eine Funktion φ mit $\varphi_{\mu_n}(t) \rightarrow \varphi(t)$, $t \in \mathbb{R}$, und ist φ stetig in 0, so ist φ charakteristische Funktion eines Wahrscheinlichkeitsmaßes μ , und es gilt $\mu_n \xrightarrow{w} \mu$.

Beweis. Da $1 = \varphi_n(0) \rightarrow \varphi(0)$, folgt $\varphi(0) = 1$. Dann zeigt das obige Argument, dass (μ_n) straff ist. Da (μ_n) nach Satz 3.31 eine schwach konvergente Teilfolge hat, etwa $\mu_{n_k} \rightarrow \nu$, folgt $\varphi_\nu = \varphi$, insbesondere ist φ eine charakteristische Funktion, und die Aussage des Korollars folgt. \blacksquare

Beispiel 3.37. Sind $X_n \sim \mathcal{N}(0, \sigma_n^2)$ und konvergieren $X_n \xrightarrow{d} X$, so folgt $\sigma_n^2 \rightarrow \sigma^2 \in [0, \infty)$ und $X \sim \mathcal{N}(0, \sigma^2)$: Die charakteristische Funktion von X_n ist $\varphi_n(t) = e^{-\sigma_n^2 t^2/2}$, und aus der Konvergenz von $\varphi_n(1) = e^{-\sigma_n^2/2}$ folgt die Konvergenz von σ_n^2 gegen ein $\sigma^2 \in [0, \infty)$. Dann muss $\varphi_n(t) \rightarrow e^{-\sigma^2 t^2/2}$, wie behauptet.

Die Aussage bleibt analog für $X_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$ richtig.

Dimensionsreduktion

Mit folgendem Resultat kann man Aussagen über schwache Konvergenz von Zufallsvektoren auf Aussagen über schwache Konvergenz von Zufallsvariablen reduzieren, welches insbesondere für Konvergenz gegen die Normalverteilung nützlich ist.

Satz 3.38 (Cramér-Wold). Sind (X_n) , X Zufallsvektoren auf \mathbb{R}^d , so sind äquivalent:

1. $X_n \xrightarrow{d} X$.
2. Für alle $\theta \in \mathbb{R}^d$ gilt $\theta^T X_n \xrightarrow{d} \theta^T X$. \blacksquare

Beweis. 1. \Rightarrow 2.: $x \mapsto \theta^T x$, $x \in \mathbb{R}^d$, ist stetig, also klar mit Satz 3.23.
2. \Rightarrow 1.: Es gilt für $t \in \mathbb{R}^d$

$$\varphi_{X_n}(t) = \mathbb{E}[e^{t^T X_n}] = \varphi_{t^T X_n}(1) \rightarrow \varphi_{t^T X}(1) = \varphi_X(t). \quad \blacksquare$$

Bemerkung. Die oben für Wahrscheinlichkeitsmaße über \mathbb{R} dargestellte Theorie lässt sich wesentlich allgemeiner entwickeln. Man vergleiche etwa

- über \mathbb{R}^d : A. van der Vaart (1998), *Asymptotic Statistics*, Cambridge University Press, Kapitel 2. Der Satz von Skorokhod, auf dem obige Darstellung stark beruht, gilt zwar im \mathbb{R}^d immer noch, aber ist wesentlich schwerer zu beweisen. Die Darstellung verzichtet daher auf dieses Hilfsmittel.
 - über metrischen Räumen: Billingsley, Patrick (1999) *Convergence of Probability Measures*, 2nd Edition, Wiley.
- Die Theorie ist hier ähnlich wie im \mathbb{R}^d , man muss aber die benötigten topologischen Eigenschaften (z.B. Separabilität und Vollständigkeit) klar formulieren.
- eine noch allgemeinere Theorie für nicht notwendigerweise messbare Abbildungen wird etwa in van der Vaart (1998), Kapitel 18, dargestellt.

4 Grundlagen der Statistik

4.1 Statistische Modelle und Statistiken

Wahrscheinlichkeitstheoretische Modellierung besteht in erster Linie in der Angabe eines geeigneten Wahrscheinlichkeitsraums $(\Omega, \mathcal{A}, \mathbb{P})$ für ein Zufallsexperiment. Als Beispiel denken wir an die Modellierung eines n -fachen Münzwurfs durch ein n -faches Produkt von Bernoulli-Experimenten mit Erfolgswahrscheinlichkeit $1/2$. Aufgaben sind dann die explizite Berechnung von Wahrscheinlichkeiten bestimmter Ereignisse, oder die Bestimmung von Verteilungen von Zufallsvariablen.

Statistische Modellierung besteht dagegen in der Angabe eines Ergebnisraums, dem **Stichprobenraum**, den wir hier mit \mathcal{X} bezeichnen wollen, sowie einer ganzen Familie von Wahrscheinlichkeitsverteilungen auf \mathcal{X} . Wir denken an das Werfen einer Reißzwecke, die mit Spitze schräg nach unten (als 0 kodiert) sowie Spitze nach oben (als 1 kodiert) fallen kann. Man hat wieder ein Bernoulli-Experiment, aber mit unbekannter Erfolgswahrscheinlichkeit $p \in (0, 1)$, also eine ganze Familie von Verteilungen $(\text{Ber}(p))_{p \in (0,1)}$ auf $\mathcal{X} = \{0, 1\}$. Aufgrund von Beobachtungen möchte man nun Rückschlüsse auf den Parameter p ziehen.

Definition 4.1. Für einen **Stichprobenraum** \mathcal{X} , mit σ -Algebra \mathcal{F} , heißt eine Familie $(\mathbb{P}_\theta)_{\theta \in \Theta}$ von Wahrscheinlichkeitsverteilungen auf $(\mathcal{X}, \mathcal{F})$ ein **statistisches Modell**.

Die Beobachtungen werden durch Ergebnisse in \mathcal{X} beschrieben und häufig als Realisierungen von Zufallsvariablen modelliert.

Beispiele 4.2. 1. Das **Bernoulli-Modell**: Hier ist $\mathcal{X} = \{0, 1\}$ und

$$(\mathbb{P}_\theta)_{\theta \in \Theta} = \{\text{Ber}(p) \mid p \in (0, 1)\}$$

mit $\theta = p$ und $\Theta = (0, 1)$. Dieses Modell wird verwendet, wenn je Versuchsdurchgang zwei mögliche Kategorien zu beobachten sind.

Beispiele sind das Werfen einer Reißzwecke, oder die „tea tasting lady“. In diesem Experiment werden einer englischen Lady zwei Tassen Tee gereicht, in denen einmal die Milch vor dem Tee und einmal die Milch nach dem Tee zugegeben wurde. Es wird jeweils geprüft, ob die Lady diese richtig zuordnen kann.

2. Das **Multinomialmodell** (mit festem $k \in \mathbb{N}$): Es sind k mögliche Versuchsausgänge eines Zufallsexperiments möglich, wir setzen daher $\mathcal{X} = \{1, \dots, k\}$ und

$$(\mathbb{P}_\theta)_{\theta \in \Theta} = \{\text{Mult}(1; p_1, \dots, p_k) \mid p_i > 0, p_1 + \dots + p_k = 1\}$$

mit $\theta = (p_1, \dots, p_k)$ und dem k -Simplex Θ .

Beispiele sind der Modelltyp eines zufällig ausgewählten Autos, die Augenfarbe einer zufällig ausgewählten Person oder die Augenzahl eines (nicht notwendigerweise fairen) Würfels.

3. Das **Poisson-Modell**: Hier ist $\mathcal{X} = \mathbb{N}_0$ und

$$(\mathbb{P}_\theta)_{\theta \in \Theta} = \{\text{Poi}(\lambda) \mid \lambda > 0\}$$

mit $\theta = \lambda$ und $\Theta = (0, \infty)$. Man benutzt dies als ein Modell für Zähldaten, bei denen viele, einzeln relativ unwahrscheinliche Ereignisse addiert werden.

Ein Beispiel ist das Rutherford-Geiger-Experiment, bei dem über 2608 Zeitintervalle von je 7,5 Sekunden die Anzahl der Zerfälle in einem radioaktiven Präparat gemessen werden.

4. Das **hypergeometrische Modell** (mit festem $n \in \mathbb{N}$): Das hypergeometrische Modell kann in zwei Varianten verwendet werden, je nachdem, welcher Parameter unbekannt ist. Ist dies die Anzahl der betroffenen Individuen R , so betrachtet man

$$(\mathbb{P}_\theta)_{\theta \in \Theta} = \{\text{Hyp}(;n, R, N), \quad R \in \mathbb{N}, R \leq N\}$$

als Familie von Verteilungen auf $\mathcal{X} = \{0, \dots, n\}$. Hier ist $\Theta = \{1, \dots, N\}$. Ein Beispiel ist ein Experiment, bei dem bei n verschiedenen Personen der Blutdruck untersucht wird und bei x Personen erhöhter Blutdruck festgestellt wurde. Hier ist N die bekannte Größe der Population und R die unbekannte Zahl der Personen mit erhöhtem Blutdruck.

Ist dagegen die Anzahl R der betroffenen Individuen bekannt aber die Gesamtzahl N unbekannt, so hat man das Modell

$$(\mathbb{P}_\theta)_{\theta \in \Theta} = \{\text{Hyp}(;n, R, N), \quad N \in \mathbb{N}, N \geq R\}$$

auf $\mathcal{X} = \{0, \dots, \min(n, R)\}$. Hier ist $\Theta = \mathbb{N} \setminus \{1, \dots, R-1\}$. Ein typische Beispiel ist ein **Capture-Recapture**-Experiment: Es soll die Anzahl der Fische in einem Teich bestimmt werden. Dazu fängt man zunächst R Fische, markiert diese und setzt sie wieder aus. Danach fängt man n Fische und betrachtet die Anzahl der markierten (also bereits vorher gefangenen) Fische x unter diesen n Fischen.

Während man für alle diskreten Beispiele jeweils \mathcal{F} als die Potenzmenge des diskreten Stichprobenraumes \mathcal{X} wählen kann, werden wir für \mathbb{R}^d -wertige, absolutstetig verteilte Zufallsvariablen jeweils den Messraum mit der Borel- σ -Algebra $\mathcal{F} = \mathcal{B}$ betrachten.

- Beispiele 4.3.**
1. Auf $(\mathbb{R}, \mathcal{B})$ interessiert man sich zum Beispiel für das **Exponentialverteilungsmodell** $(\text{Exp}(\lambda))_{\lambda \in \mathbb{R}^+}$, für kontinuierliche Wartezeiten mit unbekannter Intensität $\lambda > 0$, also $\Theta = (0, \infty)$.
 2. Betrachten wir eine eindimensionale Normalverteilung, $\mathcal{N}(\mu, \sigma^2)$. Hier sind unterschiedliche statistische Modelle denkbar, z. Bsp. $(\mathcal{N}(\mu, 1))_{\mu \in \mathbb{R}}$ bei einer Normalverteilungsfamilie mit bekannter Varianz 1 und unbekanntem Lageparameter μ mit $\Theta = \mathbb{R}$. Falls der Erwartungswert und die Varianz unbekannt sind setzen wir $\theta = (\mu, \sigma^2)$ und $\Theta = \mathbb{R} \times \mathbb{R}_+$.

Oft hat man nicht nur eine Beobachtung, sondern es sind unabhängig identisch verteilte Wiederholungen beobachtbar.

Definition 4.4. Für ein statistisches Modell $(\mathbb{P}_\theta)_{\theta \in \Theta}$ auf dem Stichprobenraum \mathcal{X} heißt das Modell $(\mathbb{P}_\theta^{\otimes n})_{\theta \in \Theta}$ der n -fachen Produktverteilungen auf \mathcal{X}^n , mit der Produkt- σ -Algebra, das **n -fache Produktmodell**.

Im diskreten Fall geben wir äquivalent die n -fachen Produkte $(p^{\otimes n}(\cdot; \theta))_{\theta \in \Theta}$ der Wahrscheinlichkeitsfunktionen p an, also

$$p^{\otimes n}(\mathbf{x}; \theta) = \prod_{i=1}^n p(x_i; \theta), \quad \mathbf{x} = (x_1, \dots, x_n)^T \in \mathcal{X}^n.$$

Im absolutstetigen Fall wird die Produktverteilung durch die Produktdichte

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

charakterisiert, wenn f die Dichte der Verteilung für $n = 1$ bezeichnet.

Bemerkung. Im Produktmodell sind die Koordinatenprojektionen $X_i : \mathcal{X}^n \rightarrow \mathcal{X}$, $i = 1, \dots, n$ für jedes feste $\theta \in \Theta$ unter $\mathbb{P}_\theta^{\otimes n}$ u.i.v. mit Verteilung \mathbb{P}_θ für jedes X_i . Umgekehrt ist die gemeinsame Verteilung

von n u.i.v. Zufallsvariablen das n -fache Produkt. Statt formal das Produktexperiment anzugeben, sagt man daher manchmal auch etwas informell: Seien X_1, \dots, X_n u.i.v. mit Verteilung \mathbb{P}_θ , $\theta \in \Theta$.

Definition 4.5. Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ ein statistisches Modell auf dem Stichprobenraum \mathcal{X} und sei \mathcal{S} eine Menge mit Messraum $(\mathcal{S}, \mathcal{F}_{\mathcal{S}})$. Eine $(\mathcal{F} - \mathcal{F}_{\mathcal{S}})$ -messbare Abbildung $T : \mathcal{X} \rightarrow \mathcal{S}$ heißt eine \mathcal{S} -wertige Statistik. Ist eine Statistik auf einem Produktmodell $(\mathbb{P}_\theta^{\otimes n})_{\theta \in \Theta}$ auf \mathcal{X}^n definiert, so kennzeichnen wir dies manchmal durch einen Index, also $T_n : \mathcal{X}^n \rightarrow \mathcal{S}$.

Definition 4.6. Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ ein statistisches Modell auf dem Stichprobenraum \mathcal{X} und $T : \mathcal{X} \rightarrow \mathcal{S}$ eine \mathcal{S} -wertige Statistik. Die Familie $(\mathbb{P}_{\theta, T})_{\theta \in \Theta}$ der Verteilungen $P_{\theta, T}(S) = P_\theta(T \in S)$, $S \subseteq \mathcal{S}$ auf \mathcal{S} heißt von T induziertes statistisches Modell (auf \mathcal{S}).

Statistiken und die induzierten Modelle sind für die Statistik ähnlich grundlegend wie Zufallsvariablen für die Wahrscheinlichkeitstheorie. Statistiken dienen unter anderem der Datenkompression oder Datenzusammenfassung.

Beispiele 4.7. 1. Im n -fachen Produkt von Bernoulli-Experimenten ist

$$T_n : \{0, 1\}^n \rightarrow \{0, 1, \dots, n\} =: \mathcal{S}, \quad T_n(\mathbf{x}) = x_1 + \dots + x_n, \quad \mathbf{x} = (x_1, \dots, x_n)^T \in \{0, 1\}^n,$$

unter $\mathbb{P}_p^{\otimes n}$ binomialverteilt, $\text{Bin}(n, p)$ mit $p \in (0, 1)$.

2. Ähnlich ist im n -fachen Produkt von Multinomialexperimenten

$$T_n : \{1, \dots, k\}^n \rightarrow \mathbb{N}_0^k, \quad T_n(\mathbf{x}) = \left(\sum_{i=1}^n 1_{\{x_i=1\}}, \dots, \sum_{i=1}^n 1_{\{x_i=k\}} \right)^T, \quad \mathbf{x} = (x_1, \dots, x_n)^T \in \{1, \dots, k\}^n$$

unter $\mathbb{P}_{(p_1, \dots, p_k)}^{\otimes n}$ multinomial $\text{Mult}(n; p_1, \dots, p_k)$ -verteilt.

3. Im n -fachen Produkt von Poisson-Experimenten ist

$$T_n : \mathbb{N}_0^n \rightarrow \mathbb{N}_0, \quad T_n(\mathbf{x}) = x_1 + \dots + x_n, \quad \mathbf{x} = (x_1, \dots, x_n)^T \in \mathbb{N}_0^n,$$

unter $\mathbb{P}_\lambda^{\otimes n}$ Poisson $\text{Poi}(n\lambda)$ -verteilt.

4.2 Parameterschätzung

4.2.1 Die Maximum-Likelihood-Methode

Wir betrachten ein statistisches Modell $(\mathbb{P}_\theta)_{\theta \in \Theta}$ mit dem Stichprobenraum \mathcal{X} . Aus der Beobachtung $x \in \mathcal{X}$ wollen wir Rückschlüsse auf den Parameter $\theta \in \Theta$ ziehen.

Definition 4.8. Ein Schätzer für den Parameter $\theta \in \Theta$ ist eine Statistik $\hat{\theta} : \mathcal{X} \rightarrow \overline{\Theta}$. In einem n -fachen Produktmodell $(\mathbb{P}_\theta^{\otimes n})_{\theta \in \Theta}$ auf \mathcal{X}^n schreiben wir $\hat{\theta}_n : \mathcal{X}^n \rightarrow \overline{\Theta}$.

Bemerkung. Zunächst werden alle messbaren Abbildungen $\hat{\theta}_n : \mathcal{X}^n \rightarrow \overline{\Theta}$ als Schätzer zugelassen, z.B. $\hat{\theta}_n : \{0, 1\}^n \rightarrow [0, 1]$, $\hat{\theta}_n(\mathbf{x}) = 1/2$ für alle $\mathbf{x} \in \{0, 1\}^n$. Offenbar ist ein solcher ‘Schätzer’ nicht besonders sinnvoll, da er die Beobachtung(en) nicht berücksichtigt, und nur gut im Fall, dass die wahre Erfolgswahrscheinlichkeit $1/2$ ist. Eine wesentliche Forderung an Schätzer ist, dass sie für jeden möglichen zugrundeliegenden Wert des Parameters sinnvolle Schätzungen geben sollen. Wir werden dies weiter unten formalisieren.

Bemerkung. Es ist im weiteren günstig als Wertebereich eines Schätzers $\hat{\theta}$ den (topologischen) Abschluss der Parametermenge $\bar{\Theta} \supseteq \Theta$ zu verwenden, also beispielsweise $[0, 1]$ für die auf $(0, 1)$ eingeschränkte Erfolgswahrscheinlichkeit eines Bernoulli-Experimentes. Für offene Mengen Θ ist $\bar{\Theta}$ dann eine echte Obermenge von Θ .

Wir wollen zunächst ein allgemeines Prinzip zur Parameterschätzung einführen, und danach Eigenschaften von Schätzern diskutieren.

Das Maximum-Likelihood-Prinzip

Gegeben die Beobachtung $x \in \mathcal{X}$ soll der Parameter θ gewählt werden, so dass die Wahrscheinlichkeit, x unter \mathbb{P}_θ zu beobachten, maximal wird.

Definition 4.9. Sei $\mathbb{P}_\theta \ll \mu$, mit einem σ -endlichen Maß μ , für alle $\theta \in \Theta$, mit der Dichte $f_\theta = d\mathbb{P}_\theta/d\mu$. Wir nennen

$$L : \Theta \times \mathcal{X} \rightarrow [0, 1], \quad L(\theta; x) = f_\theta(x)$$

die **Likelihood-Funktion** (des statistischen Modells). Der Logarithmus der Likelihood-Funktion

$$\mathcal{L}(\theta; x) = \log L(\theta; x)$$

heißt **Log-Likelihood-Funktion**.

Im diskreten Fall mit einer Wahrscheinlichkeitsfunktion p , ist die Likelihood-Funktion also

$$L : \Theta \times \mathcal{X} \rightarrow [0, 1], \quad L(\theta; x) = p(x; \theta).$$

Im absolutstetigen Fall entspricht die Likelihood-Funktion der Lebesguedichte, aber ebenfalls als Funktion im Parameter (und in x) und aufgefasst.

Definition 4.10. Ein Schätzer $\hat{\theta}^{\text{ML}} : \mathcal{X} \rightarrow \bar{\Theta}$ heißt ein **Maximum-Likelihood-Schätzer (ML-Schätzer)**, falls gilt

$$\forall x \in \mathcal{X} : L(\hat{\theta}^{\text{ML}}(x); x) = \sup_{\theta \in \Theta} L(\theta, x).$$

Bemerkung. Wendet man die Definition in einem diskreten Produktmodell $(p^{\otimes n}(\cdot; \theta))_{\theta \in \Theta}$ auf \mathcal{X}^n an, so erhält man

$$L_n(\theta; \mathbf{x}) = \prod_{i=1}^n p(x_i; \theta), \quad \mathcal{L}_n(\theta; \mathbf{x}) = \sum_{i=1}^n \log p(x_i; \theta), \quad \mathbf{x} = (x_1, \dots, x_n)^T \in \mathcal{X}^n,$$

und $\hat{\theta}_n^{\text{ML}} : \mathcal{X}^n \rightarrow \bar{\Theta}$ muss erfüllen

$$\begin{aligned} \forall \mathbf{x} \in \mathcal{X}^n : L_n(\hat{\theta}_n^{\text{ML}}(\mathbf{x}); \mathbf{x}) &= \sup_{\theta \in \Theta} L_n(\theta, \mathbf{x}) \\ \Leftrightarrow \forall \mathbf{x} \in \mathcal{X}^n : \mathcal{L}_n(\hat{\theta}_n^{\text{ML}}(\mathbf{x}); \mathbf{x}) &= \sup_{\theta \in \Theta} \mathcal{L}_n(\theta, \mathbf{x}) = \sup_{\theta \in \Theta} \sum_{i=1}^n \mathcal{L}(\theta, x_i). \end{aligned}$$

Analog hat die Likelihood in einem absolutstetigen Produktmodell die Form eines Produktes und die Log-Likelihood einer Summe.

Beispiele 4.11. 1. Ist $(\mathbb{P}_\theta)_{\theta \in \Theta} = \{\text{Ber}(p) \mid p \in (0, 1)\}$, so ist

$$L_n(p; x_1, \dots, x_n) = \prod_{k=1}^n p^{x_k} (1-p)^{1-x_k}$$

Tabelle 4.1: Absolute Häufigkeiten im Rutherford-Geiger-Experiment.

Anz. Zerfälle	0	1	2	3	4	5	6	7
Anz. Zeitinterv.	57	203	383	525	532	408	273	139
Anz. Zerfälle	8	9	10	11	12	13	14	
Anz. Zeitinterv.	45	27	10	4	0	1	1	

die Likelihood-Funktion. Um diese über p zu maximieren, bilden wir zunächst die Log-Likelihood-Funktion

$$\mathcal{L}_n(p; x_1, \dots, x_n) = \sum_{k=1}^n (x_k \ln p + (1 - x_k) \ln(1 - p)).$$

Wir bestimmen die Ableitung und erhalten als ML-Schätzer die Stelle, an der

$$\frac{d\mathcal{L}_n(p; x_1, \dots, x_n)}{dp} = \frac{1}{p} \sum_{k=1}^n x_k + \frac{1}{1-p} \sum_{k=1}^n x_k - \frac{n}{1-p} = 0$$

gilt und somit

$$\sum_{k=1}^n x_k = n \hat{p}_n^{\text{ML}}(x_1, \dots, x_n) \Leftrightarrow \hat{p}_n^{\text{ML}}(x_1, \dots, x_n) = \frac{1}{n} \sum_{k=1}^n x_k = \bar{x}_n.$$

Mit dem Monotonieverhalten der Ableitung, $\frac{d\mathcal{L}_n}{dp} < 0$ für $p > \bar{x}_n$ und $\frac{d\mathcal{L}_n}{dp} > 0$ für $p < \bar{x}_n$, prüft man leicht nach, dass es sich tatsächlich um ein Maximum handelt. Da weiter gilt $L_n(p; x_1, \dots, x_n) \rightarrow 0$ für $p \rightarrow 0$ und $p \rightarrow 1$, ist $\hat{p}_n^{\text{ML}}(x_1, \dots, x_n) = \bar{x}_n$ eindeutiger ML-Schätzer.

- Ist $(\mathbb{P}_\theta)_{\theta \in \Theta} = \{\text{Poi}(\lambda) \mid \lambda > 0\}$, so ist

$$L_n(\lambda; x_1, \dots, x_n) = \prod_{k=1}^n e^{-\lambda} \frac{\lambda^{x_k}}{x_k!}$$

die Likelihood-Funktion, und die Log-Likelihood-Funktion ist gegeben durch

$$\mathcal{L}_n(\lambda; x_1, \dots, x_n) = -n\lambda + \sum_{k=1}^n (x_k \ln \lambda - \ln(x_k!)).$$

Eine einfache Rechnung zeigt

$$\hat{\lambda}_n^{\text{ML}}(x_1, \dots, x_n) = \frac{1}{n} \sum_{k=1}^n x_k = \bar{x}_n.$$

Mit der zweiten Ableitung prüft man leicht nach, dass es sich tatsächlich um ein Maximum handelt.

Als Beispiel betrachten wir das Rutherford-Geiger-Experiment, bei dem wie oben erwähnt über 2608 Zeitintervalle von je 7,5 Sekunden die Anzahl der Zerfälle in einem radioaktiven Präparat gemessen werden. Tabelle 4.1 enthält die Ergebnisse. Der Parameter der Poisson-Verteilung wird mit $\hat{\lambda} = 3,87$ geschätzt. In Abbildung 4.1 sind die relativen Häufigkeiten sowie die Wahrscheinlichkeiten unter der geschätzten Poisson-Verteilung $\text{Poi}(3,87)$ abgebildet, die Anpassung erscheint relativ gut.

- Betrachte die diskrete uniforme Verteilung mit $\mathcal{X} = \{1, \dots, N\}$ und $\mathbb{P}_\theta = U(\{1, 2, \dots, N\})$, mit $\theta = N$ und $\Theta = \mathbb{N}$ und mit der Wahrscheinlichkeitsfunktion $p(x; N) = 1/N$ für alle $x \in \{1, \dots, N\}$.

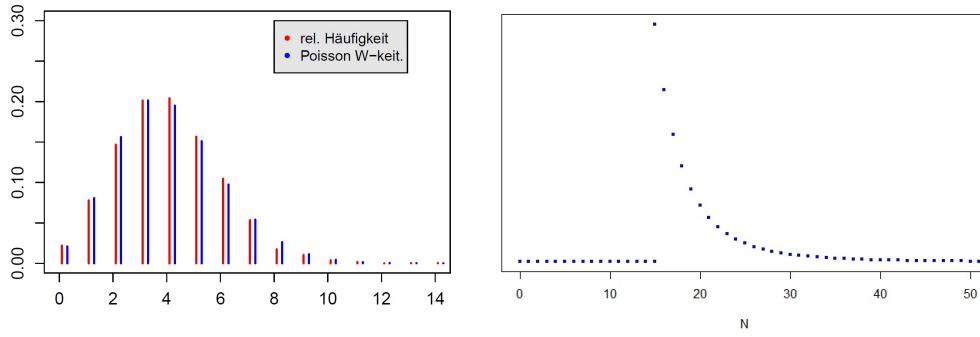


Abbildung 4.1: Links: Relative Häufigkeiten sowie Wahrscheinlichkeiten unter der geschätzten Poisson-Verteilung im Rutherford-Geiger-Experiment. Rechts: Veranschaulichung von $L(N; x_1, \dots, x_n)$ in Beispiel 4.11 3. für $\max_{1 \leq i \leq n} x_i = 15$ und $n = 5$.

Mit $L(N; x_1, \dots, x_n) = \prod_{i=1}^n 1(x_i \in \{1, \dots, N\})/N$ ergibt sich

$$\begin{aligned} L(N; x_1, \dots, x_n) &= \prod_{i=1}^n 1(x_i \in \{1, \dots, N\})/N \\ &= \begin{cases} N^{-n}, & 1 \leq x_1, \dots, x_n \leq N, \\ 0, & \text{sonst,} \end{cases} \\ &= \begin{cases} N^{-n}, & \min_{1 \leq i \leq n} x_i \geq 1, \max_{1 \leq i \leq n} x_i \leq N, \\ 0, & \text{sonst.} \end{cases} \\ \Rightarrow \hat{N}_n^{\text{ML}} &= \max_{1 \leq i \leq n} x_i. \end{aligned}$$

Hier wird das Maximum am Rand des Trägers der Likelihood angenommen und wird über die Monotonie der Likelihood gefunden, siehe Abbildung 4.1. Wir finden also $\hat{N}_n^{\text{ML}} = \max_{1 \leq i \leq n} x_i$.

Bei den Beispielen 1. und 2. müssen wir den Fall $\sum_{k=1}^n x_k = 0$ separat betrachten. In dem Fall wird jeweils das Supremum für $\theta = 0$ angenommen, auf dem Rand von Θ . Beispiel 1. verallgemeinert sich direkt auf das Binomialmodell, da der zusätzliche Binomialkoeffizient in der Wahrscheinlichkeitsfunktion nicht von p abhängt. In der Literatur wird Definition 4.10 manchmal auf das Vorliegen eines eindeutigen Maximums eingeschränkt. Wie bereits in den Beispielen, hat unsere allgemeinere Definition Vorteile. Eindeutigkeit des Maximums setzen wir dabei nicht voraus.

Bemerkung. Basiert man die Schätzung auf den induzierten Modellen $\{\text{Bin}(n, p), p \in (0, 1)\}$ auf $\mathcal{S} = \{0, 1, \dots, n\}$ oder $\{\text{Poi}(n\lambda), \lambda \in (0, \infty)\}$ auf $\mathcal{S} = \mathbb{N}_0$ in Beispiel 4.7, so erhält man dieselben ML-Schätzer.

Beispiel 4.12. Wir betrachten folgendes Modell: $\mathcal{X} = \mathbb{N}_0$, $\Theta = [0, 1] \times (0, \infty)^2$, $\theta = (p, \lambda_1, \lambda_2)^T$ und

$$p(x; \theta) = p e^{-\lambda_1} \frac{\lambda_1^{x_1}}{x_1!} + (1-p) e^{-\lambda_2} \frac{\lambda_2^{x_2}}{x_2!}.$$

Dieses Modell wäre etwa nützlich für ein Experiment, in dem radioaktiver Zerfall einer Stichprobe gemessen wird mit Anteilen zweier unterschiedlicher radioaktiver Präparate. Die Log-Likelihood im n -fachen Produktexperiment ist

$$\mathcal{L}_n(\theta; x_1, \dots, x_n) = \sum_{k=1}^n \ln \left(p e^{-\lambda_1} \frac{\lambda_1^{x_k}}{x_k!} + (1-p) e^{-\lambda_2} \frac{\lambda_2^{x_k}}{x_k!} \right),$$

diese kann nur numerisch maximiert werden. Dies ist ein typischer Fall, ML-Schätzer sind in der Regel nicht explizit.

4.2.2 Die Momentenmethode

Sogenannte **Momentenschätzer** basieren auf folgendem einfachen Konstruktionsprinzip. Man betrachte ein statistisches Modell mit X_1, \dots, X_n unabhängig identisch verteilten Zufallsvariablen mit einer Verteilungsfamilie $\{\mathbb{P}_\theta, \theta \in \Theta \subseteq \mathbb{R}^k\}$ und $\mathbb{E}[|X_1|^k] < \infty$. Bezeichne $\mu_j = \mathbb{E}[X_1^j]$ das j -te Moment wobei $j \leq k$. Das empirische j -te Moment ist

$$\hat{\mu}_j = n^{-1} \sum_{i=1}^n X_i^j,$$

und dieses liefert einen natürlichen Schätzer für $\mu_j, j = 1, \dots, k$. Insbesondere gilt $\mathbb{E}[\hat{\mu}_j] = \mu_j$. Häufig interessiert man sich für einen Parameter θ , für den gilt $h_j(\theta) = \mu_j$ mit Funktionen $h_j, j = 1, \dots, k$. Die Idee ist nun in dieser Gleichung (bzw. diesen Gleichungen), μ_j durch die empirischen Momente $\hat{\mu}_j$ zu ersetzen. Man erhält dann einen Momentenschätzer $\hat{\theta}$ durch Lösen der Gleichung(en) nach θ . Diese Vorgehensweise nennt man die **Momentenmethode**.

Beispiele 4.13. 1. Es seien $X_1, \dots, X_n \sim \text{Exp}(\lambda)$ unabhängig identisch exponentialverteilte Zufallsvariablen mit unbekanntem Parameter $\lambda > 0$. Betrachte $h_k(\lambda) = \mu_k = \mathbb{E}_\lambda[X_i^k] = \lambda^{-k}k!$. Für jedes $k \in \mathbb{N}$ erhalten wir einen Momentenschätzer

$$\hat{\lambda}_{k,n} := \left(\frac{k!}{\frac{1}{n} \sum_{i=1}^n X_i^k} \right)^{1/k}.$$

2. Betrachte einen **autoregressiven Prozess** der Ordnung 1 (AR(1)-Prozess):

$$X_n = aX_{n-1} + \varepsilon_n, \quad n \geq 1,$$

mit (ε_n) unabhängig identisch verteilt, $\mathbb{E}[\varepsilon_n] = 0$, $\text{Var}(\varepsilon_n) = \sigma^2 < \infty$ und $X_0 = x_0 \in \mathbb{R}$.

Um den Parameter a zu schätzen, nutzen wir die folgende Identität

$$\mathbb{E}[X_{n-1}X_n | \varepsilon_1, \dots, \varepsilon_{n-1}] = aX_{n-1}^2.$$

Dies führt auf den (modifizierten) Momentenschätzer (**Yule-Walker-Schätzer**):

$$\hat{a}_n := \frac{\frac{1}{n} \sum_{k=1}^n X_{k-1}X_k}{\frac{1}{n} \sum_{k=1}^n X_{k-1}^2} = a + \frac{\sum_{k=1}^n X_{k-1}\varepsilon_k}{\sum_{k=1}^n X_{k-1}^2}.$$

Bemerkung. In beiden Fällen von Beispiel 4.11 1. und 2. ist der Parameter p bzw. λ gleich dem Erwartungswert, und als Maximum-Likelihood-Schätzer ergab sich der Mittelwert. Hier entsprechen die ML-Schätzer also Momentenschätzern und sind somit explizit. In sogenannten Exponentialfamilien (worunter unter anderem Exponential-, Normal-, Bernoulli- und Poisson-Verteilung fallen) kann man allgemein zeigen, dass Maximum-Likelihood-Schätzer bestimmten Momentenschätzern entsprechen.

4.2.3 Eigenschaften von Schätzern

Wir wollen nun formale Konzepte dafür einführen, dass Schätzer für jeden möglichen Parameterwert "funktionieren" sollen und um ihre Güte zu bewerten.

Definition 4.14. Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ ein statistisches Modell mit $\Theta \subseteq \mathbb{R}$ auf dem Ergebnisraum \mathcal{X} . Ein Schätzer $\hat{\theta} : \mathcal{X} \rightarrow \overline{\Theta}$ heißt **erwartungstreu** für $\theta \in \Theta$, falls gilt:

$$\forall \theta \in \Theta : \mathbb{E}_\theta[\hat{\theta}] = \theta,$$

wobei $\mathbb{E}_\theta[\hat{\theta}]$ der Erwartungswert unter \mathbb{P}_θ ist, und vorausgesetzt wird, dass dieser für jedes θ existiert.

Beispiel 4.15. Wichtig ist, dass die Gleichheit $\mathbb{E}_\theta[\hat{\theta}] = \theta$ für **jedes** $\theta \in \Theta$ gefordert wird. Betrachten wir das Bernoulli-Modell mit n Wiederholungen, also $\mathcal{X}^n = \{0,1\}^n$ mit dem n -fachen Produkt von Bernoulli- θ -Verteilungen, $\theta \in (0,1)$.

- a. $\hat{\theta}_{n,1}(x_1, \dots, x_n) = 1/2$ ist nicht erwartungstreue, da $\mathbb{E}_\theta[\hat{\theta}_{n,1}] = 1/2 \neq \theta$ für alle $\theta \in (0,1) \setminus \{1/2\}$.
- b. $\hat{\theta}_{n,2}(x_1, \dots, x_n) = x_1$ und $\hat{\theta}_{n,3}(x_1, \dots, x_n) = (x_1 + \dots + x_n)/n$ sind dagegen offenbar erwartungstreue.

Definition 4.16. Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ ein statistisches Modell mit $\Theta \subseteq \mathbb{R}$ auf dem Ergebnisraum \mathcal{X} , und sei $\hat{\theta} : \mathcal{X} \rightarrow \overline{\Theta}$ ein Schätzer, für den gilt

$$\forall \theta \in \Theta : \mathbb{E}_\theta[\hat{\theta}^2] < \infty.$$

Die Abbildung

$$MSE_{\hat{\theta}} : \Theta \rightarrow [0, \infty), MSE_{\hat{\theta}}(\theta) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2]$$

heißt **mittlerer quadratischer Fehler (mean squared error)** von $\hat{\theta}$.

Bemerkung. Es gilt für alle $\theta \in \Theta$ folgende **Bias-Varianz-Zerlegung**

$$MSE_{\hat{\theta}}(\theta) = \text{Var}_\theta(\hat{\theta}) + (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2.$$

Insbesondere ist für einen erwartungstreuen Schätzer

$$MSE_{\hat{\theta}}(\theta) = \text{Var}_\theta(\hat{\theta}).$$

Beispiel 4.17. In Beispiel 4.15 sind

$$MSE_{\hat{\theta}_{n,1}}(\theta) = \left(\frac{1}{2} - \theta\right)^2, \quad MSE_{\hat{\theta}_{n,2}}(\theta) = \theta(1-\theta), \quad MSE_{\hat{\theta}_{n,3}}(\theta) = \frac{\theta(1-\theta)}{n}.$$

Gerade der Schätzer $\hat{\theta}_{n,1}$, dessen MSE in $\theta = 1/2$ verschwindet, zeigt auf dass man den MSE nicht gleichmäßig in Θ durch einen Schätzer minimieren kann.

Behauptung: $\hat{\theta}_{n,3}$ hat jedoch die kleinste Varianz unter allen erwartungstreuen Schätzern.

Beweis: Betrachte den Fall $Z = \sum_{i=1}^n X_i \sim \text{Bin}(n, \theta)$ mit Werten in $\{0, \dots, n\}$. Sei $\hat{\theta}(Z)$ ein Schätzer mit $\mathbb{E}[(\hat{\theta}(Z))^2] < \infty$. Die Cauchy-Schwarz-Ungleichung ergibt, dass

$$\left(\text{Cov}\left(\frac{d\mathcal{L}(\theta; Z)}{d\theta}, \hat{\theta}(Z)\right)\right)^2 \leq \text{Var}\left(\frac{d\mathcal{L}(\theta; Z)}{d\theta}\right) \text{Var}(\hat{\theta}(Z)).$$

Es ist $d\mathcal{L}(\theta; Z)/d\theta = Z/(\theta(1-\theta)) - n/(1-\theta)$ und daher $\mathbb{E}_\theta[d\mathcal{L}(\theta; Z)/d\theta] = 0$. Mit

$$\begin{aligned} \mathbb{E}_\theta\left[\frac{d\mathcal{L}(\theta; Z)}{d\theta} \hat{\theta}(Z)\right] &= \sum_{i=0}^n \frac{1}{L(\theta; i)} \frac{dL(\theta; i)}{d\theta} \hat{\theta}(i) L(\theta; i) \\ &= \frac{d}{d\theta} \sum_{i=0}^n \hat{\theta}(i) L(\theta; i) = \frac{d}{d\theta} \mathbb{E}_\theta[\hat{\theta}(Z)] = \frac{d\theta}{d\theta} = 1, \end{aligned}$$

sowie $\text{Var}\left(\frac{d\mathcal{L}(\theta; Z)}{d\theta}\right) = \text{Var}\left(\frac{Z}{\theta(1-\theta)}\right) = \frac{n}{\theta(1-\theta)}$, erhalten wir für erwartungstreue Schätzer die untere Schranke

$$\text{Var}(\hat{\theta}(Z)) \geq \frac{\theta(1-\theta)}{n}.$$

Definition 4.18. Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ ein statistisches Modell mit $\Theta \subseteq \mathbb{R}^k$ auf dem Stichprobenraum \mathcal{X} . Für jedes $n \in \mathbb{N}$ sei $\hat{\theta}_n : \mathcal{X}^n \rightarrow \Theta$ ein Schätzer für θ auf dem n -fachen Produktmodell. Dann heißt die Folge der Schätzer $(\hat{\theta}_n)_{n \geq 1}$ **konsistent** für $\theta \in \Theta$, falls gilt:

$$\forall \theta \in \Theta, \varepsilon > 0 : \quad \mathbb{P}_\theta^{\otimes n}(\{\mathbf{x} \in \mathcal{X}^n : \|\hat{\theta}_n(\mathbf{x}) - \theta\| \geq \varepsilon\}) \rightarrow 0, \quad n \rightarrow \infty,$$

wobei $\|\cdot\|$ eine Norm auf \mathbb{R}^k bezeichne.

Bemerkung. Konsistenz einer Folge von Schätzern $(\hat{\theta}_n)$ heißt also, dass für alle θ die Folge $(\hat{\theta}_n)$ stochastisch gegen den wahren Parameter θ konvergiert. Dies bedeutet, dass der Parameter für eine wachsende Anzahl an Beobachtungen immer präziser geschätzt werden kann. Wiederum ist wesentlich, dass die Konvergenz für jedes $\theta \in \Theta$, und nicht nur für spezielle Parameterwerte erfüllt ist.

Nach dem schwachen Gesetz der großen Zahlen sind die ML-Schätzer p_n^{ML} und λ_n^{ML} konsistent. Konsistenz und asymptotische Normalität (ein zentraler Grenzwertsatz) von Momentenschätzern lassen sich sehr allgemein zeigen. Dies werden wir in Kapitel 5.2 behandeln. Allgemeine Resultate zur Asymptotik von Maximum-Likelihood-Schätzern sind schwieriger herzuleiten. In der mathematischen Statistik wird sich zeigen, dass letztere häufig konsistent und sogar asymptotisch effizient sind im Sinne einer kleinst möglichen asymptotischen Varianz. Letzteres haben wir (nur) in Beispiel 4.17 gezeigt.

4.3 Statistische Tests

4.3.1 Einführendes Beispiel und Begriffe

Beispiel 4.19 (Einführung). Eine englische Lady trinkt Tee mit Milch und behauptet, erkennen zu können, ob die Milch vor oder nach dem Tee in die Tasse gegeben wurde. Wir vermuten, dass das nicht stimmt und wollen testen, ob Grund zur Annahme besteht, sie hätte tatsächlich recht. Dazu starten wir einen Versuch (**tea tasting lady-Versuch**). Wir servieren der Lady n Mal zwei Tassen Tee mit Milch, wobei einmal die Milch vor dem Tee und einmal nach dem Tee zugegeben wurde. Dies soll die Lady richtig zuordnen.

Wir modellieren das Experiment als Produkt von Bernoulli-Experimenten mit unbekannter Erfolgswahrscheinlichkeit p , wobei die Ergebnisse 1 und 0 korrekte bzw. falsche Zuordnungen darstellen.

Wir stellen nun die **Hypothese H** auf, dass die Lady nur rät. Diese hat die **Alternative K** , dass die Lady tendenziell die Tassen korrekt klassifizieren kann. Dies beschreibt man formal als

$$H : p = \frac{1}{2} \quad \text{gegen} \quad K : p > \frac{1}{2}.$$

Sind nun x_1, \dots, x_n beobachtet und $s_n = x_1 + \dots + x_n$ die Anzahl der Erfolge, so wird man sich gegen H und für K entscheiden, wenn s_n sehr groß (etwa im Vergleich zum unter H erwarteten Wert $n/2$) ist.

Präziser betrachtet man die Wahrscheinlichkeit unter der Hypothese $p = 1/2$ für einen Wert $\geq s_n$, den sogenannten **p-Wert**. Falls dieser klein genug ist (typisch $\leq 5\%$), falls also die Wahrscheinlichkeit für eine Fehlentscheidung zugunsten von K hinreichend klein ist, verwirft man H zugunsten von K . Wir beschreiben die Vorgehensweise noch einmal explizit als Anleitung des Ablaufs eines Testverfahrens:

1. **Wahl eines geeigneten statistischen Modells**
In unserem Beispiel $s_n \sim \text{Bin}(n, p)$ mit unbekanntem Parameter p .
2. **Formulierung von Hypothese und Alternative**
Im Beispiel die Hypothese $H : p = 1/2$ gegen die Alternative $K : p \in (1/2, 1]$.
Es wird also die Hypothese (auch Nullhypothese) $H : p = 1/2$ gegen die Alternative (auch Alternativhypothese) $K : p > 1/2$ getestet.

3. Festlegung des Niveaus

Wähle ein $\alpha \in (0, 1)$, zum Beispiel $\alpha = 0,05$, und beschränke die Wahrscheinlichkeit einer Entscheidung für die Alternative obwohl die Hypothese vorliegt (Fehler 1. Art) durch α .

4. Statistische Entscheidungsregel

Man wählt eine Statistik, $\varphi : \mathcal{X} \rightarrow [0, 1]$ mit folgender Bedeutung. Tritt $x \in \mathcal{X}$ ein, so bedeutet

$$\begin{aligned}\varphi(x) &= 0, H \text{ wird nicht abgelehnt}, \\ \varphi(x) &= 1, H \text{ wird abgelehnt}, \\ \varphi(x) &\in (0, 1), \text{ eine Randomisierung entscheidet}.\end{aligned}$$

Bei der Randomisierung wird die Entscheidung mittels eines zusätzlichen unabhängigen Zufallsexperiments bestimmt, wobei mit Wahrscheinlichkeit $\varphi(x)$ die Wahl gegen H ausfällt. Die letzte Möglichkeit (randomisierter Test) ist in der Anwendung oft nicht gewollt, aber nützlich zur mathematischen Behandlung und der Konstruktion optimaler Tests. Ohne Randomisierung ist ein Test eine Statistik $\varphi : \mathcal{X} \rightarrow \{0, 1\}$, mit einem zwei-elementigen Bildraum. Ein Test wird meist so formuliert, dass die Hypothese abgelehnt werden soll, man also dadurch eher auf die Alternative schließt. Das Niveau beschränkt aber lediglich die Wahrscheinlichkeit des Fehlers die Hypothese fälschlicherweise zu verwerfen und erlaubt nicht den Schluss für $\varphi(x) = 0$ mit diesem Irrtumsniveau die Hypothese für wahr zu erklären. Daher formulieren wir oben “ H wird nicht abgelehnt”, statt etwa “Entscheidung für H ”. Der dritten Schritt wird heute praktisch oft so abgewandelt, dass man statt dessen (mit dem PC) den p-Wert bestimmt, das kleinste mögliche (Irrtums-)Niveau, zu dem die Hypothese abgelehnt werden kann.

Definition 4.20. Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ ein statistisches Modell auf dem Ergebnisraum \mathcal{X} . Zerlegt man den Parameterraum disjunkt in $\Theta = \Theta_0 \cup \Theta_1$, so heißt die Aussage $H : \theta \in \Theta_0$ die **Hypothese** (oder **Nullhypothese**) und die Aussage $K : \theta \in \Theta_1$ die **Alternative** (oder **Alternativhypothese**). Ist $\text{card } H = 1$, so heißt H **einfach**, sonst heißt H **zusammengesetzt**, analog für K .

Definition 4.21. Eine Statistik $T : \mathcal{X} \rightarrow \{0, 1\}$ heißt (nicht-randomisierter) **Test**. In einem Produktmodell schreiben wir wiederum $T_n : \mathcal{X}^n \rightarrow \{0, 1\}$.

Bemerkung (Interpretation). Die Ergebnisse eines Tests werden so interpretiert, dass anzunehmen ist, dass $\theta \in \Theta_{T(x)}$ gilt. Man behält also bei $T(x) = 0$ die Hypothese H bei und verwirft sie bei $T(x) = 1$.

Definition 4.22. Verwirft man H fälschlicherweise, d.h. wird die Beobachtung von einem \mathbb{P}_θ für $\theta \in \Theta_0$ generiert aber $T(x) = 1$, so spricht man von einem **Fehler erster Art**. Behält man H fälschlicherweise bei, d.h. wird die Beobachtung von einem \mathbb{P}_θ für $\theta \in \Theta_1$ generiert aber $T(x) = 0$, so spricht man von einem **Fehler zweiter Art**.

Definition 4.23. Sei $T : \mathcal{X} \rightarrow \{0, 1\}$ ein Test. Dann heißt die Funktion

$$\beta_T : \Theta \rightarrow [0, 1], \quad \beta_T(\theta) = \mathbb{P}_\theta(\{x \in \mathcal{X} : T(x) = 1\})$$

die **Gütefunktion** von T . Ist $\theta \in \Theta_1$, so bezeichnet man $\beta_T(\theta)$ auch als **Güte**, oder **Macht** (engl. **power**) des Tests.

Bemerkung. 1. Für $\theta \in \Theta_0$ ist $\beta_T(\theta)$ die Wahrscheinlichkeit, H fälschlicherweise zu verwerfen, also für einen Fehler erster Art. Also sollte $\beta_T(\theta)$ möglichst klein sein.

2. Für $\theta \in \Theta_1$ ist $\beta_T(\theta)$ die Wahrscheinlichkeit, K tatsächlich zu erkennen. Die Wahrscheinlichkeit für einen Fehler zweiter Art ist dann $1 - \beta_T(\theta)$. Also sollte $\beta_T(\theta)$ möglichst groß sein.

Definition 4.24. Ein Test T hat **Signifikanzniveau** oder **Niveau** $\alpha \in (0, 1)$, falls für seine Gütefunktion β_T gilt

$$\sup_{\theta \in \Theta_0} \beta_T(\theta) \leq \alpha.$$

Gilt sogar

$$\sup_{\theta \in \Theta_0} \beta_T(\theta) < \alpha,$$

so heißt der Test T **konservativ** zum Niveau α .

Bemerkung. Hier gibt man also eine Schranke α (etwa 5% oder 1%) für die Wahrscheinlichkeit eines Fehlers erster Art vor.

Definition 4.25. Hat der Test $T : \mathcal{X} \rightarrow \{0, 1\}$ die Form $T(x) = 1_I(S(x))$ für eine Statistik $S : \mathcal{X} \rightarrow \mathbb{R}$ und $I \subseteq \mathbb{R}$, so heißen S die **Prüfgröße** des Tests und I (oder auch $I \cap S(\mathcal{X})$) deren **Verwerfungsbereich**. Man spricht bei $I = [t, \infty)$ oder $I = (-\infty, t]$ mit einem festen $t \in \mathbb{R}$ von einem **oberen** bzw. **unteren Verwerfungsbereich** (zusammengefasst ein **einseitiger Verwerfungsbereich**) und bei $I = (-\infty, t_1] \cup [t_2, \infty)$ mit $t_1 < t_2$ von einem **zweiseitigen Verwerfungsbereich**.

Weiter heißen für eine Beobachtung $x_0 \in \mathcal{X}$

$$\begin{aligned} PVal_l(S; x_0) &:= \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\{x \in \mathcal{X} : S(x) \leq S(x_0)\}) \quad \text{linksseitiger } p\text{-Wert,} \\ PVal_r(S; x_0) &:= \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\{x \in \mathcal{X} : S(x) \geq S(x_0)\}) \quad \text{rechtsseitiger } p\text{-Wert.} \end{aligned}$$

Die p -Werte hängen von x_0 nur über $s_0 = S(x_0)$ ab, man schreibt daher auch $PVal_l(S; s_0)$ bzw. $PVal_r(S; s_0)$.

Bemerkung. Die Wahrscheinlichkeiten können über die Verteilungen von S , also das induzierte Modell berechnet werden, etwa

$$PVal_r(S; x_0) := \sup_{\theta \in \Theta_0} \mathbb{P}_{\theta, S}(\{s \in S(\mathcal{X}) : s \geq S(x_0)\}).$$

4.3.2 Einseitige Tests

Satz 4.26. Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ ein statistisches Modell auf dem Ergebnisraum \mathcal{X} , $\Theta = \Theta_0 \cup \Theta_1$ das gegebene Testproblem und $S : \mathcal{X} \rightarrow \mathbb{R}$ eine reellwertige Statistik als Prüfgröße mit induziertem Modell $P_{\theta, S}(A) = P_\theta(S \in A)$, $A \subseteq S(\mathcal{X})$ und $0 < \alpha < 1/2$ ein gegebenes Testniveau. Wir nehmen an, dass $\min(A) \in A$ für alle nach unten beschränkten $A \subseteq S(\mathcal{X})$ gilt.

1. Wählt man

$$t_r = \min \left(t \in S(\mathcal{X}) : \sup_{\theta \in \Theta_0} \mathbb{P}_{\theta, S}(\{s \in S(\mathcal{X}) : s \geq t\}) \leq \alpha \right),$$

so ist $T(x) = 1_{[t_r, \infty)}(S(x))$ ein Test zum Niveau α .

2. Es ist $T(x_0) = 1$ genau dann, wenn $PVal_r(S; x_0) \leq \alpha$.

Wir setzen dabei $\min \emptyset = \infty$, dann ist $T(x) = 0$ für alle $x \in \mathcal{X}$. Analoges gilt für ein linksseitiges Testproblem.

Beweis. Zu 1.: Es ist

$$\sup_{\theta \in \Theta_0} P_\theta(T = 1) = \sup_{\theta \in \Theta_0} P_\theta(S \geq t_r) = \sup_{\theta \in \Theta_0} \mathbb{P}_{\theta,S}(\{s \in S(\mathcal{X}) : s \geq t_r\}) \leq \alpha$$

nach Wahl von t_r .

Zu 2.: Es ist

$$t \mapsto \sup_{\theta \in \Theta_0} \mathbb{P}_{\theta,S}(\{s \in S(\mathcal{X}) : s \geq t\})$$

monoton fallend in t . Daher haben wir

$$\begin{aligned} T(x_0) = 1 &\Leftrightarrow S(x_0) \geq t_r \Leftrightarrow \sup_{\theta \in \Theta_0} \mathbb{P}_{\theta,S}(\{s \in S(\mathcal{X}) : s \geq S(x_0)\}) \leq \alpha \\ &\Leftrightarrow \text{PVal}_r(S; x_0) \leq \alpha. \end{aligned}$$

■

Problematisch bei der Anwendung des Satzes ist, dass das Niveau bzw. der p-Wert bei einer zusammengesetzten Hypothese über ganz Θ_0 kontrolliert werden muss. Manchmal kann man sich mit Monotonieargumenten auf einen “extremen” Wert in Θ_0 zu beschränken.

Beispiel 4.27 (Binomialtest). Wir betrachten das Problem der tea tasting lady. Bilden wir $S_n(x_1, \dots, x_n) = x_1 + \dots + x_n$, so erhalten wir das induzierte Binomialmodell auf $\{0, 1, \dots, n\}$ mit Verteilungen $\text{Bin}(n, p)$, $p \in (0, 1)$.

Für die tea tasting lady hatten wir das Testproblem $H_0 : p = 1/2$ gegen $K : p > 1/2$ formuliert. Möchte man keine Parameter einfach ausschließen, wäre für allgemeines $p_0 \in (0, 1)$ fest (etwa $p_0 = 0,5$) das einseitige (obere) Testproblem gegeben durch

$$H_0 : p \leq p_0 \quad \text{gegen} \quad K : p > p_0. \quad (4.1)$$

Um mit der zusammengesetzten Hypothese umzugehen, benötigen wir folgendes

Lemma 4.28. Sei $S_n \sim \text{Bin}(n, p)$, $p \in (0, 1)$. Für $x \in \{0, 1, \dots, n\}$ setze

$$F(p; n, x) = \mathbb{P}(S_n \leq x) = \sum_{t=0}^x \binom{n}{t} p^t (1-p)^{n-t}.$$

Dann ist für $x \in \{0, \dots, n-1\}$

$$F(p; n, x) = 1 - \frac{n!}{x!(n-x-1)!} \int_0^p t^x (1-t)^{n-x-1} dt. \quad (4.2)$$

streng monoton fallend in p . Das Bild von $(0, 1)$ unter $F(\cdot; n, x)$ ist $(0, 1)$.

Beweis. Wir zeigen zunächst (4.2) für festes $n \in \mathbb{N}$ durch Induktion über $x \in \{0, \dots, n-1\}$. Für $x = 0$ ist

$$\begin{aligned} \mathbb{P}(S_n = 0) &= (1-p)^n = 1 - (1 - (1-p)^n) \\ &= 1 - n \int_0^p (1-t)^{n-1} dt = F(p; n, 0). \end{aligned}$$

Mit partieller Integration folgt für $x \in \{0, \dots, n-2\}$

$$\int_0^p t^x (1-t)^{n-x-1} dt = \frac{t^{x+1}}{x+1} (1-t)^{n-x-1} \Big|_{t=0}^p + \frac{n-x-1}{x+1} \int_0^p t^{x+1} (1-t)^{n-(x+1)-1} dt$$

$$= \frac{p^{x+1} (1-p)^{n-x-1}}{x+1} + \frac{n-x-1}{x+1} \int_0^p t^{x+1} (1-t)^{n-(x+1)-1} dt$$

und somit

$$\begin{aligned} F(p : n, x+1) &= F(p : n, x) + \binom{n}{x+1} p^{x+1} (1-p)^{n-(x+1)} \\ &= 1 - \frac{n!}{x!(n-x-1)!} \left(\frac{p^{x+1} (1-p)^{n-x-1}}{x+1} + \frac{n-x-1}{x+1} \int_0^p t^{x+1} (1-t)^{n-(x+1)-1} dt \right) \\ &\quad + \binom{n}{x+1} p^{x+1} (1-p)^{n-(x+1)} \\ &= 1 - \frac{n!}{(x+1)! (n-(x+1)-1)!} \int_0^p t^{x+1} (1-t)^{n-(x+1)-1} dt. \end{aligned}$$

Da der Integrand $t^x (1-t)^{n-x-1}$ für $t \in (0, 1)$ positiv ist, ist $F(p : n, x)$ als Funktion von p streng monoton fallend. Aufgrund der Stetigkeit ist das Bild des Intervalls $(0, 1)$ unter $F(\cdot : n, x)$ ein Intervall. Aus (4.2) folgt $F(p; n, x) \rightarrow 1$, $p \rightarrow 0$, und aus der Definition und $x \leq n-1$ folgt $F(p; n, x) \rightarrow 0$, $p \rightarrow 1$. ■

Nach dem Lemma ist $1 - F(p; n, x)$ streng monoton wachsend in p , und für $\alpha \in (0, 1)$ ergibt sich

$$\begin{aligned} t_r &:= \min \left(t \in \{1, \dots, n\} : \sup_{0 < p \leq p_0} \mathbb{P}_p^{\otimes n} (S_n \geq t) \leq \alpha \right) \\ &= \min \left(t \in \{1, \dots, n\} : \sup_{0 < p \leq p_0} [1 - \mathbb{P}_p^{\otimes n} (S_n \leq t-1)] \leq \alpha \right) \\ &= \min \left(t \in \{1, \dots, n\} : \sup_{0 < p \leq p_0} (1 - F(p; n, t-1)) \leq \alpha \right) \\ &= \min \left(t \in \{1, \dots, n\} : (1 - F(p_0; n, t-1)) \leq \alpha \right), \end{aligned} \tag{4.3}$$

was leicht berechnet werden kann. Dabei ist $t_r = \infty$ und der Verwerfungsbereich leer, falls $1 - F(p_0; n, t-1) = p_0^n > \alpha$ ist. Analog ist für beobachtetes $s = S_n(x)$

$$\text{PVal}_r(S_n; s) = 1 - F(p_0; n, s-1),$$

und wir verwerfen zum gegebenen Niveau α , falls dieser Wert $\leq \alpha$ ist.

Aufgrund der strengen Monotonie in Lemma 4.28 gilt

$$\begin{aligned} \mathbb{P}_p^{\otimes n} (S_n \geq t_r) &= 1 - F(p; n, t_r - 1) \\ &> 1 - F(p_0; n, t_r - 1) = \mathbb{P}_{p_0}^{\otimes n} (S_n \geq t_r), \quad p > p_0. \end{aligned}$$

Die Güte ist also unter der Alternative strikt größer als unter der Hypothese.

Wir betrachten zwei konkrete Zahlenbeispiele im Zusammenhang mit der tea tasting lady ($p_0 = 1/2$): Angenommen, $n = 10$, und $\alpha = 0,05$. Dann ist $t_r = 9$, man kann also erst ab 9 Treffern der Lady die Hypothese verwerfen. Es ist $\text{PVal}_r(S_{10}; 8) = 0,05469 > \alpha$ und $\text{PVal}_r(S_{10}; 9) = 0,01074 < \alpha$.

Für $n = 20$ erhält man $t_r = 15$, und $\text{PVal}_r(S_{20}; 14) = 0,05766 > \alpha$ und $\text{PVal}_r(S_{20}; 15) = 0,02069 < \alpha$. ◇

Ein analoges Vorgehen ist etwa für die Poisson-Verteilung möglich.

Beispiel 4.29 (Multinomialmodell). Wir betrachten allgemeiner das Multinomialmodell mit n Wiederholungen und $k \geq 2$ möglichen Ausgängen, und wollen die Hypothese testen, dass

$$p_j = p_{j,0}, \quad j = 1, \dots, k,$$

für festes $p_{1,0} > 0, \dots, p_{k,0} > 0$ mit $p_{1,0} + \dots + p_{k,0} = 1$.

Konkret könnten wir etwa testen wollen, ob ein gegebener Würfel fair ist, also $k = 6$ und $p_{j,0} = 1/6$, $j = 1, \dots, 6$.

Wegen des mehrdimensionalen Parameterraums ist das Testproblem komplizierter als im Binomialmodell. Als Prüfgröße benutzt man etwa die χ^2 -Statistik

$$S_n(n_1, \dots, n_k) = \sum_{j=1}^k \frac{(n_j - p_{j,0}n)^2}{p_{j,0}n},$$

wobei $n_j \in \mathbb{N}_0$, $n_1 + \dots + n_k = n$ beobachtet sind. Man verwirft nun für große Werte von S_n , benutzt also den einseitigen Verwerfungsbereich aus Satz 4.26 für die Prüfgröße S_n . Um die Verteilung von S_n unter der Hypothese zu bestimmen werden Simulationsverfahren oder asymptotische Approximationen angewendet.

4.3.3 Optimalität von Tests

In der Statistik sind oft mehrere Test- oder Schätzmethoden möglich, daher versucht man, Optimalität zu definieren und optimale Verfahren zu konstruieren.

Definition 4.30. Seien $(\mathbb{P}_\theta)_{\theta \in \Theta}$ ein statistisches Modell auf dem Ergebnisraum \mathcal{X} , $\Theta = \Theta_0 \cup \Theta_1$ das gegebene Testproblem, $0 < \alpha < 1/2$ das Testniveau, und seien $\varphi, \tilde{\varphi} : \mathcal{X} \rightarrow \{0, 1\}$ zwei Tests mit Niveau α . Wir nennen φ gleichmäßig besser als $\tilde{\varphi}$, falls gilt

$$\forall \theta \in \Theta_1 : \beta_\varphi(\theta) \geq \beta_{\tilde{\varphi}}(\theta).$$

Es kann noch gefordert werden, dass ein $\theta \in \Theta_1$ existiert mit $\beta_T(\theta) > \beta_{\tilde{T}}(\theta)$.

Beispiel 4.31. Wir betrachten das einseitige Testproblem (4.1) zur Binomialverteilung. Wählt man einen Verwerfungsbereich mit unterer Grenze $s < t_r$ für t_r aus (4.3), so hat man offenbar keinen Test zum Niveau α mehr. Wählt man aber $s > t_r$ so hat der Test weiterhin Niveau α , aber die Wahl t_r führt zu einem gleichmäßig besseren Test zum Niveau α , wie man leicht einsieht.

Sind allgemeiner φ und $\tilde{\varphi}$ zwei Tests zum Niveau α mit $\{\tilde{\varphi} = 1\} \subseteq \{\varphi = 1\}$, so ist φ gleichmäßig besser als $\tilde{\varphi}$, da die Verwerfungswahrscheinlichkeit für φ stets höher ist. Man muss also den Verwerfungsbereich so groß wie möglich machen unter der Restriktion, dass das Niveau noch eingehalten wird.

Man interessiert sich speziell für „beste“ statistische Tests.

Definition 4.32. Ein Test φ der Hypothese $H_0 : \theta \in \Theta_0$ gegen die Alternative $K : \theta \in \Theta_1$ ist ein gleichmäßig bester Test^a zum Niveau α , falls φ Niveau α besitzt sowie für alle anderen Tests φ' zum Niveau α die Macht (power) nicht größer als die von φ ist:

$$\forall \theta \in \Theta_1 : \mathbb{E}_\theta[\varphi] \geq \mathbb{E}_\theta[\varphi'].$$

Ein Test φ ist unverfälscht zum Niveau α , falls φ Niveau α besitzt sowie auf der Alternative $\mathbb{E}_\theta[\varphi] \geq \alpha$, $\theta \in \Theta_1$, gilt. φ heißt gleichmäßig bester unverfälschter Test^b zum Niveau α , falls φ unverfälscht zum Niveau α ist sowie alle anderen unverfälschten Tests φ' zum Niveau α nicht größere Macht besitzen.

^auniformly most powerful = UMP

^buniformly most powerful unbiased = UMPU

Die **Neyman-Pearson-Theorie** liefert einem gleichmäßig beste Tests einfacher Hypothesen gegen einfache Alternativen, also UMP-Tests für einfache binäre Testprobleme. Sie beinhaltet auch ein Konstruktionsprinzip. Im Allgemeinen gibt es für komplexere Hypothesen keine gleichmäßig besten Tests, siehe Beispiel 4.36. Daher ist man für zweiseitige Testprobleme meist an UMPU-Tests interessiert. Die im Beispiel 4.31 heuristisch begründete Optimalität des einseitigen Binomialtests, kann aber mit der Neyman-Pearson-Theorie ebenfalls formal bewiesen werden.

Da uns Satz 4.26 und Satz 4.33 allgemeinere Konstruktionen von Tests ermöglichen, verzichten wir an dieser Stelle auf die Neyman-Pearson-Theorie. Weitergehend Interessierte finden diese im Buch von Krengel (§6.7) oder im Buch von Georgii (§10). Allgemeine Optimalitätsresultate werden in der mathematischen Statistik behandelt.

4.3.4 Zweiseitige Tests

Wir wenden uns jetzt zweiseitigen Testproblemen zu. Im Vergleich zu Satz 4.26 stellt sich die Frage, wie das Niveau α unter der Hypothese auf die beiden Seiten des Verwerfungsbereichs verteilt werden soll.

Ein plausibler Ansatz, der jedoch nicht zu optimalen Ergebnissen führen muss, ist die $\alpha/2$ -Methode, wie in folgendem Satz.

Satz 4.33. Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ ein statistisches Modell auf dem Ergebnisraum \mathcal{X} , $\Theta = \Theta_0 \cup \Theta_1$ mit $\Theta_0 = \{\theta_0\}$ das gegebene Testproblem und $S : \mathcal{X} \rightarrow \mathbb{R}$ eine reellwertige Teststatistik, und $0 < \alpha < 1/2$ ein gegebenes Testniveau. Wählt man

$$\begin{aligned} t_r &= \min \left(t \in S(\mathcal{X}) : \mathbb{P}_{\theta_0} (S \geq t) \leq \alpha/2 \right), \\ t_l &= \max \left(t \in S(\mathcal{X}) : \mathbb{P}_{\theta_0} (S \leq t) \leq \alpha/2 \right), \end{aligned}$$

so ist $T(x) = 1_{(-\infty, t_l] \cup [t_r, \infty)}(S(x))$ ein Test zum Niveau α , wobei $\min \emptyset = \infty$, $\max \emptyset = -\infty$ gesetzt wird.

■

Beispiel 4.34 (Binomialmodell). Wir betrachten das Binomialmodell aus Beispiel 4.27. Im zweiseitigen Testproblem

$$H : p = p_0 \quad \text{gegen} \quad K : p \neq p_0 \tag{4.4}$$

für ein festes $p = p_0$ besteht die Hypothese nur aus einem Punkt. Die Grenzen aus dem obigen Satz lauten somit

$$\begin{aligned} t_r &= \min \left(x \in \{1, \dots, n\} : (1 - F(p_0; n, x-1)) \leq \alpha/2 \right), \\ t_l &= \max \left(x \in \{0, 1, \dots, n-2\} : F(p_0; n, x) \leq \alpha/2 \right), \end{aligned}$$

mit resultierendem Test $T(x) = 1_{(-\infty, t_l] \cup [t_r, \infty)}(S_n(x))$.

Wir bemerken, dass der einseitige Test aus Satz 4.27 auch ein Test zum Niveau α zum zweiseitigen Testproblem (4.4) ist. Dieser wird in der Regel eine höhere Güte haben als der zweiseitige Test für Alternativen $p > p_0$. Allerdings liegt seine Güte für Alternativen $p < p_0$ unterhalb des Testniveaus. Der Test ist also verfälscht. Der zweiseitige Binomialtest ist UMPU.

Als Zahlenbeispiele bemerken wir, dass für $p_0 = 1/2$ bei $\alpha = 0,05$ aufgrund der Symmetrie für $n = 10$ gilt $t_r = 9$, $t_l = 1$ und für $n = 20$ $t_r = 15$ und $t_l = 5$.

Für $n = 30$ und $p_0 = 0,3$ ist im zweiseitigen Problem $t_l = 3$ und $t_r = 15$, dagegen ist die rechte Grenze im einseitigen Test = 14.

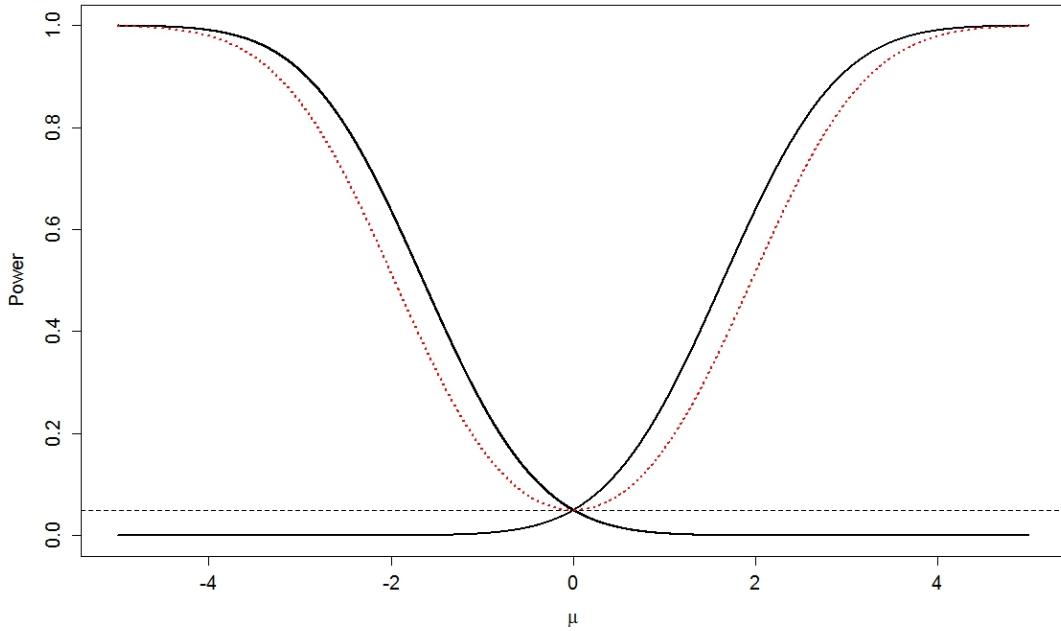


Abbildung 4.2: Gütfunktionen der einseitigen und des zweiseitigen Gauß-Tests mit $\mu_0 = 0$ und $\sigma^2 = 1$, also $\Theta = \mathbb{R}$ sowie $\Theta_0 = \{0\}$, jeweils zum Niveau $\alpha = 0,05$.

Beispiel 4.35 (Gauß-Test). Es seien $X_1, \dots, X_n \sim \mathcal{N}(\theta, \sigma^2)$ unabhängig identisch normalverteilte Zufallsvariablen mit $\theta \in \mathbb{R}$ unbekannt und $\sigma > 0$ bekannt. Wir geben einen gleichmäßig besten unverfälschten Test von $H_0 : \theta = \theta_0$ gegen $K : \theta \neq \theta_0$ an. Dieser basiert auf der unter $\mathcal{N}(\theta_0, \sigma_0^2)$ standardnormalverteilten Teststatistik (vgl. Satz 2.17)

$$T(X_1, \dots, X_n) = \sqrt{n} \frac{\bar{X} - \theta_0}{\sigma_0}.$$

Aus Symmetriegründen wähle kritische Werte $k_1 = -k$, $k_2 = k$ für den Verwerfungsbereich und verzichte wegen stetiger Verteilung auf Randomisierung, so dass $\varphi^* = \mathbf{1}(|T(x)| > k)$ gilt. Wählt man $k = \Phi^{-1}(1 - \alpha/2)$, das $(1 - \alpha/2)$ -Quantil von $\mathcal{N}(0, 1)$, so gilt $\mathbb{E}_{\theta_0}[\varphi^*] = \alpha$. Dies nennt man den **zweiseitigen Gauß-Test**.

Beispiel 4.36. Betrachte $X \sim \mathcal{N}(\mu, 1)$ mit unbekanntem Parameter $\mu \in \mathbb{R}$. Für ein Testproblem $H_0 : \mu = 0$ gegen $K : \mu = \mu_0 \neq 0$ ist der einseitige Gauß-Test, analog begründet wie für den Binomialtest in Beispiel 4.31 und formal nach Neyman-Pearson, ein gleichmäßig bester Test. Bezeichne diesen als φ falls $\mu_0 > 0$ bzw. $\tilde{\varphi}$ falls $\mu_0 < 0$.

Betrachte nun das Testproblem $H_0 : \mu = 0$ gegen die zwei-elementige Alternative $K : \mu \in \{\mu_0, \mu_1\}$ mit $\mu_0 < 0$ und $\mu_1 > 0$. Für einen gleichmäßig besten Test $\tilde{\varphi}$ müsste nun gelten:

$$\mathbb{E}_\mu[\tilde{\varphi}] \geq \max(\mathbb{E}_\mu[\varphi], \mathbb{E}_\mu[\tilde{\varphi}]) \text{ für } \mu = \mu_0 \text{ und } \mu = \mu_1.$$

So ein Test kann aber nicht existieren, vgl. Abbildung 4.2. Die einseitigen Gauß-Tests, die jeweils auf einem Teil der Alternative bessere Güte als der zweiseitige erreichen, sind jedoch verfälscht. Man kann zeigen, dass der zweiseitige Gauß-Test UMPU ist.

4.4 Konfidenzintervalle

In gewisser Weise ist die Angabe eines Punktschätzers für einen unbekannten reellen Parameter noch ein unbefriedigendes Resultat, wenn man nicht die Unsicherheit der Schätzung quantifizieren kann. Kennt man genauer die (asymptotische) Verteilung eines Schätzers, ist es möglich ein (asymptotisches) Konfidenzintervall zu bestimmen welches mit einer bestimmten großen Wahrscheinlichkeit den unbekannten Parameter einschließt.

Definition 4.37. Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ ein statistisches Modell mit $\Theta \subseteq \mathbb{R}^k$ auf dem Ergebnisraum \mathcal{X} . Ein reeller Parameter ist eine Abbildung $\gamma: \Theta \rightarrow \mathbb{R}$. Eine intervallwertige Abbildung

$$I: \mathcal{X} \rightarrow \mathcal{P}(\mathbb{R}), \quad I(x) = [U(x), O(x)]$$

mit den Statistiken $U, O: \mathcal{X} \rightarrow \mathbb{R}$ mit $U \leq O$ heißt (zweiseitige) **Intervallschätzung** für den Parameter γ . Analog kann man auch offene oder halboffene Intervalle verwenden.

Ist $I(x) = [U(x), \infty)$, so spricht man von einer rechtsseitigen Intervallschätzung, ist $I(x) = (-\infty, O(x)]$ von einer linksseitigen Intervallschätzung.

Bemerkung. Ist die Intervallschätzung auf einem n -fachen Produktmodell definiert, so schreiben wir wieder I_n . Ist $\Theta \subseteq \mathbb{R}$, so kann man $\gamma(\theta) = \theta$ nehmen.

Definition 4.38. Die Überdeckungswahrscheinlichkeit einer Intervallschätzung I für einen Parameter γ ist die Abbildung

$$\theta \mapsto \mathbb{P}_\theta(\{\gamma(\theta) \in I(x)\}), \quad \theta \in \Theta.$$

Als **Konfidenzniveau** einer Intervallschätzung bezeichnet man die minimale Überdeckungswahrscheinlichkeit

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta(\gamma(\theta) \in I(x)).$$

Definition 4.39. Eine Intervallschätzung I heißt (exaktes) **Konfidenzintervall** zum **Konfidenzniveau** $1 - \alpha$ (für ein festes $\alpha \in [0, 1]$), falls gilt

$$\forall \theta \in \Theta: \quad \mathbb{P}_\theta(\gamma(\theta) \in I(x)) \geq 1 - \alpha.$$

Konfidenzintervalle im Binomialmodell mit der Tschebyschev-Ungleichung

Satz 4.40. Sei $n \in \mathbb{N}$, $\mathcal{X} = \{0, 1, \dots, n\}$ und $p(x; p) = \binom{n}{x} p^x (1-p)^{n-x}$, $p \in (0, 1)$, das Binomialmodell. Für $\alpha \in (0, 1)$ setze

$$U(x) = \max\left(0, \frac{x}{n} - \frac{1}{2\sqrt{\alpha n}}\right), \quad O(x) = \min\left(1, \frac{x}{n} + \frac{1}{2\sqrt{\alpha n}}\right).$$

Dann ist $I(x) = (U(x), O(x))$, $x \in \mathcal{X}$, ein Konfidenzintervall zum Niveau $1 - \alpha$ für $p \in (0, 1)$. ■

Beweis. Für alle $p \in (0, 1)$ ist $p(1-p) \leq 1/4$. Nach der Tschebyschev-Ungleichung gilt daher

$$\forall p \in (0, 1): \quad \mathbb{P}_p\left(\left|x - \frac{x}{n}\right| \geq \varepsilon\right) \leq \frac{1}{n\varepsilon^2 4}.$$

Für $\varepsilon = (2\sqrt{\alpha n})^{-1}$ und Übergang zum Komplement ergibt sich

$$\forall p \in (0, 1) : \quad \mathbb{P}_p \left(\left\{ x \in \mathcal{X} : \left| \frac{x}{n} - p \right| < \frac{1}{2\sqrt{\alpha n}} \right\} \right) \geq 1 - \alpha.$$

Da gilt

$$\left| \frac{x}{n} - p \right| < \frac{1}{2\sqrt{\alpha n}} \Leftrightarrow p \in (U(x), O(x)),$$

ergibt sich die Behauptung. ■

Bemerkung. Die (nicht-symmetrische Version der) Hoeffding-Ungleichung ergibt

$$\forall p \in (0, 1) : \quad \mathbb{P}_p \left(\left\{ x \in \mathcal{X} : \left| \frac{x}{n} - p \right| \geq \varepsilon \right\} \right) \leq 2 \exp(-2n\varepsilon^2).$$

Hieraus kann man analog folgende Konfidenzintervalle gewinnen

$$U(x) = \max \left(0, \frac{x}{n} - \frac{\sqrt{\log(2/\alpha)}}{\sqrt{2n}} \right), \quad O(x) = \min \left(1, \frac{x}{n} + \frac{\sqrt{\log(2/\alpha)}}{\sqrt{2n}} \right).$$

Für $\alpha = 0,05$ ist

$$\frac{1}{2\sqrt{\alpha}} \approx 2,24 > 1,36 \approx \sqrt{\log(2/\alpha)}/\sqrt{2}.$$

Somit ist das resultierende Konfidenzintervall kleiner als das aus der Tschebyschev-Ungleichung, obwohl beide nach Konstruktion nur das Niveau 95% garantieren.

Konfidenzintervalle und Tests

Zwischen Konfidenzintervallen und Tests gibt es einen engen Zusammenhang, den wir nachfolgend genauer herausarbeiten wollen.

Satz 4.41. Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ ein statistisches Modell mit $\Theta \subseteq \mathbb{R}$ auf dem Ergebnisraum \mathcal{X} , $0 < \alpha < 1/2$ und $I : \mathcal{X} \rightarrow \mathcal{P}(\mathbb{R})$ ein Konfidenzintervall für $\theta \in \Theta$ zum Konfidenzniveau $1 - \alpha$.

Für alle $\theta_0 \in \Theta$ ist dann $T_{\theta_0}(x) = 1_{\{\theta_0 \notin I(x)\}}$ ein Test für

$$H : \theta = \theta_0 \quad \text{gegen} \quad K : \theta \neq \theta_0$$

zum Niveau $\alpha > 0$. ■

Beweis.

$$\mathbb{P}_{\theta_0} (T_{\theta_0} = 1) = \mathbb{P}_{\theta_0} (\{x \in \mathcal{X} : \theta_0 \notin I(x)\}) = 1 - \mathbb{P}_{\theta_0} (\{x \in \mathcal{X} : \theta_0 \in I(x)\}) \leq \alpha$$

da $\mathbb{P}_{\theta_0} (\{x \in \mathcal{X} : \theta_0 \in I(x)\}) \geq 1 - \alpha$. ■

Satz 4.42. Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ ein statistisches Modell mit $\Theta \subseteq \mathbb{R}$ auf dem Ergebnisraum \mathcal{X} , $0 < \alpha < 1/2$, und für jedes $\theta_0 \in \Theta$ existiere ein Test $T_{\theta_0} : \mathcal{X} \rightarrow \{0, 1\}$ für

$$H : \theta = \theta_0 \quad \text{gegen} \quad K : \theta \neq \theta_0$$

zum Niveau $\alpha > 0$. Dann ist

$$I(x) = \{ \theta_0 \in \Theta : T_{\theta_0}(x) = 0 \}$$

ein **Konfidenzbereich** für θ zum Konfidenzniveau $1 - \alpha$ (nicht unbedingt ein Intervall). ■

Beweis. Es ist $\theta \in I(x)$ nach Definition genau dann, wenn $T_\theta(x) = 0$, also

$$\mathbb{P}_\theta(\{x \in \mathcal{X} : \theta \in I(x)\}) = \mathbb{P}_\theta(T_\theta = 0) \geq 1 - \alpha.$$

■

Es gibt also einen direkten Zusammenhang, eine **Korrespondenz**, zwischen Test und Konfidenzintervall.

Asymptotische Konfidenzintervalle im Binomialmodell

Definition 4.43. Für jedes $n \geq n_0$ sei I_n eine Intervallschätzung auf dem n -fachen Produktmodell \mathcal{X}^n . Eine Folge $(I_n)_{n \geq 1}$ heißt **asymptotisches Konfidenzintervall** zum Konfidenzniveau $1 - \alpha$, falls gilt

$$\forall \theta \in \Theta : \liminf_{n \rightarrow \infty} \mathbb{P}_\theta^{\otimes n}(\{\mathbf{x} \in \mathcal{X}^n : \gamma(\theta) \in I_n(\mathbf{x})\}) \geq 1 - \alpha.$$

Wir betrachten das Produkt von Bernoulli-Experimenten, also $\mathcal{X}^n = \{0, 1\}^n$ mit Produktverteilung

$$p^{\otimes n}(\mathbf{x}; p) = p^{s_n(\mathbf{x})} (1-p)^{n-s_n(\mathbf{x})}, \quad \mathbf{x} \in \{0, 1\}^n, p \in (0, 1),$$

und $s_n(\mathbf{x}) = x_1 + \dots + x_n$. Wir setzen $\bar{x}_n = s_n(\mathbf{x})/n$.

Hier wollen wir asymptotische Konfidenzintervalle für den Parameter p konstruieren.

Satz 4.44. Für $\mathbf{x} \in \{0, 1\}^n$ und $0 < \alpha < 1/2$, mit $\bar{x}_n = s_n(\mathbf{x})/n$, seien

$$\begin{aligned} U_n(\mathbf{x}) &= \frac{n\bar{x}_n + q_{1-\alpha/2}^2/2}{n + q_{1-\alpha/2}^2} - \frac{q_{1-\alpha/2}(n\bar{x}_n(1-\bar{x}_n) + q_{1-\alpha/2}^2/4)^{1/2}}{n + q_{1-\alpha/2}^2} \\ O_n(\mathbf{x}) &= \frac{n\bar{x}_n + q_{1-\alpha/2}^2/2}{n + q_{1-\alpha/2}^2} + \frac{q_{1-\alpha/2}(n\bar{x}_n(1-\bar{x}_n) + q_{1-\alpha/2}^2/4)^{1/2}}{n + q_{1-\alpha/2}^2} \end{aligned} \tag{4.5}$$

wobei $q_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$, und Φ die Verteilungsfunktion der Standardnormalverteilung ist. Dann ist $I_n(\mathbf{x}) = [U_n(\mathbf{x}), O_n(\mathbf{x})]$ ein asymptotisches Konfidenzintervall zum Konfidenzniveau $1 - \alpha$, und genauer gilt

$$\forall p \in (0, 1) : \mathbb{P}^{\otimes n}\left(\{\mathbf{x} \in \{0, 1\}^n : p \in I_n(\mathbf{x})\}\right) \rightarrow 1 - \alpha, \quad n \rightarrow \infty.$$

■

Beweis. Nach dem zentralen Grenzwertsatz gilt

$$\begin{aligned} \forall p \in (0, 1) : \mathbb{P}^{\otimes n}\left(\{\mathbf{x} \in \{0, 1\}^n : -q_{1-\alpha/2} \leq \sqrt{n} \frac{\bar{x}_n - p}{\sqrt{p(1-p)}} \leq q_{1-\alpha/2}\}\right) \\ \rightarrow \Phi(\Phi^{-1}(1 - \alpha/2)) - \Phi(\Phi^{-1}(\alpha/2)) = 1 - \alpha, \quad n \rightarrow \infty. \end{aligned} \tag{4.6}$$

Durch Lösen einer quadratischen Gleichung können wir die Ungleichungen

$$-q_{1-\alpha/2} \leq \sqrt{n} \frac{\bar{x}_n - p}{\sqrt{p(1-p)}} \leq q_{1-\alpha/2}$$

Tabelle 4.2: Asymptotische Konfidenzintervalle $I_n(\mathbf{x})$ nach (4.5) und (4.7) zum Niveau 0,95 in Abhängigkeit von der beobachteten relativen Häufigkeit \bar{x}_n .

\bar{x}_n	$n = 20$				$n = 100$			
	$U_n(\mathbf{x})$	$O_n(\mathbf{x})$	$\tilde{U}_n(\mathbf{x})$	$\tilde{O}_n(\mathbf{x})$	$U_n(\mathbf{x})$	$O_n(\mathbf{x})$	$\tilde{U}_n(\mathbf{x})$	$\tilde{O}_n(\mathbf{x})$
0	0	0.161	0	0	0	0.037	0	0
0.05	0.009	0.236	-0.046	0.146	0.022	0.112	0.007	0.093
0.1	0.028	0.301	-0.031	0.231	0.055	0.174	0.041	0.159
0.15	0.052	0.36	-0.006	0.306	0.093	0.233	0.08	0.22
0.2	0.081	0.416	0.025	0.375	0.133	0.289	0.122	0.278
0.25	0.112	0.469	0.06	0.44	0.175	0.343	0.165	0.335
0.3	0.145	0.519	0.099	0.501	0.219	0.396	0.21	0.39
0.35	0.181	0.567	0.141	0.559	0.264	0.447	0.257	0.443
0.4	0.219	0.613	0.185	0.615	0.309	0.498	0.304	0.496
0.45	0.258	0.658	0.232	0.668	0.356	0.548	0.352	0.548
0.5	0.299	0.701	0.281	0.719	0.404	0.596	0.402	0.598
0.55	0.342	0.742	0.332	0.768	0.452	0.644	0.452	0.648
0.6	0.387	0.781	0.385	0.815	0.502	0.691	0.504	0.696
0.65	0.433	0.819	0.441	0.859	0.553	0.736	0.557	0.743
0.7	0.481	0.855	0.499	0.901	0.604	0.781	0.61	0.79
0.75	0.531	0.888	0.56	0.94	0.657	0.825	0.665	0.835
0.8	0.584	0.919	0.625	0.975	0.711	0.867	0.722	0.878
0.85	0.64	0.948	0.694	1.006	0.767	0.907	0.78	0.92
0.9	0.699	0.972	0.769	1.031	0.826	0.945	0.841	0.959
0.95	0.764	0.991	0.854	1.046	0.888	0.978	0.907	0.993
1	0.839	1	1	1	0.963	1	1	1

nach p auflösen und erhalten äquivalent $U_n(\mathbf{x}) \leq p \leq O_n(\mathbf{x})$. Dies ergibt unmittelbar die Aussage des Satzes. ■

Analog zum vorherigen Satz kann man einseitige Konfidenzintervalle konstruieren.

Bemerkung (Stetigkeitskorrektur). Nutzt man die Approximation des zentralen Grenzwertsatzes mit **Stetigkeitskorrektur**, erhält man ein Intervall $[U_{n,c}(\mathbf{x}), O_{n,c}(\mathbf{x})]$, wobei für $U_{n,c}(\mathbf{x})$ in (4.5) \bar{x}_n überall durch $\bar{x}_n - 1/(2n)$, und für $O_{n,c}(\mathbf{x})$ \bar{x}_n überall durch $\bar{x}_n + 1/(2n)$ ersetzt werden.

Einfachere und intuitivere Konfidenzintervalle ergeben sich, wenn man die Varianz $p(1-p)$ durch einen Schätzer ersetzt; die Aussage des zentralen Grenzwertsatzes gilt dann weiterhin.

Satz 4.45. Für $\mathbf{x} \in \{0,1\}^n$ und $0 < \alpha < 1/2$ seien

$$\tilde{U}_n(\mathbf{x}) = \bar{x}_n - \frac{q_{1-\alpha/2} \sqrt{\bar{x}_n(1-\bar{x}_n)}}{\sqrt{n}}, \quad \tilde{O}_n(\mathbf{x}) = \bar{x}_n + \frac{q_{1-\alpha/2} \sqrt{\bar{x}_n(1-\bar{x}_n)}}{\sqrt{n}}. \quad (4.7)$$

Dann ist $\tilde{I}_n(\mathbf{x}) = [\tilde{U}_n(\mathbf{x}), \tilde{O}_n(\mathbf{x})]$ ebenfalls ein asymptotisches Konfidenzintervall zum Konfidenzniveau $1 - \alpha$, und genauer gilt

$$\forall p \in (0,1) : \mathbb{P}^{\otimes n} \left(\{ \mathbf{x} \in \{0,1\}^n : p \in \tilde{I}_n(\mathbf{x}) \} \right) \rightarrow 1 - \alpha, \quad n \rightarrow \infty.$$

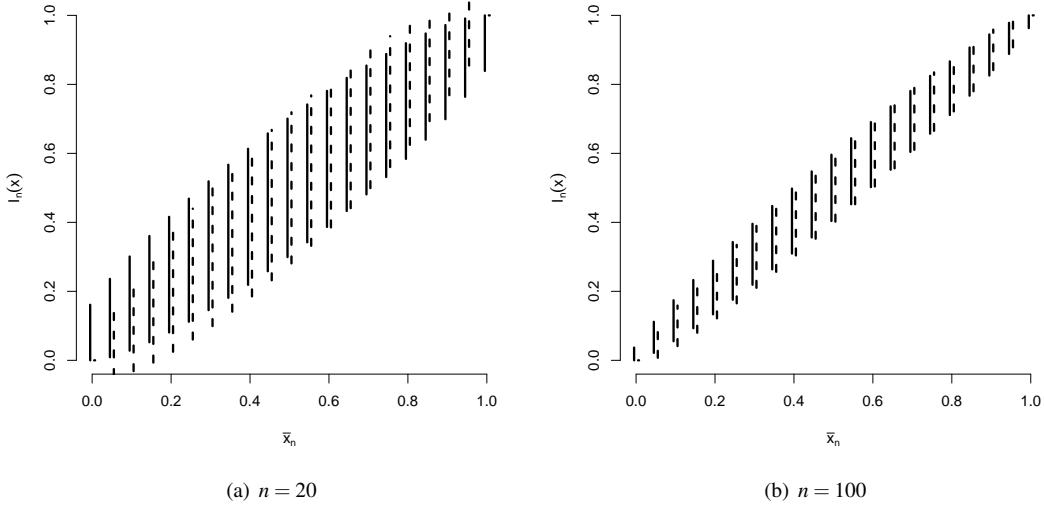


Abbildung 4.3: Asymptotische Konfidenzintervalle $I_n(\mathbf{x})$ nach (4.5) und (4.7) (gestrichelt) zum Niveau 0.95 in Abhängigkeit von der beobachteten relativen Häufigkeit \bar{x}_n .

Beweis. Mit dem Lemma von Slutsky folgt, dass wenn man in (4.6) den Ausdruck $p(1 - p)$ für die Varianz durch einen konsistenten Schätzer ersetzt, die Aussage des zentralen Grenzwertsatzes weiterhin gilt, also

$$\forall p \in (0, 1) : \mathbb{P}^{\otimes n} \left(\left\{ \mathbf{x} \in \{0, 1\}^n : -q_{1-\alpha/2} \leq \sqrt{n} \frac{\bar{x}_n - p}{\sqrt{\bar{x}_n(1 - \bar{x}_n)}} \leq q_{1-\alpha/2} \right\} \right) \rightarrow 1 - \alpha, \quad n \rightarrow \infty.$$

Löst man die Ungleichung

$$-q_{1-\alpha/2} \leq \sqrt{n} \frac{\bar{x}_n - p}{\sqrt{\bar{x}_n(1 - \bar{x}_n)}} \leq q_{1-\alpha/2}$$

nach p auf, erhält man die Grenzen in (4.7). ■

Bemerkung (Stetigkeitskorrektur). Nutzt man wiederum die Approximation des zentralen Grenzwertsatzes mit Stetigkeitskorrektur, erhält man ein Intervall $[\tilde{U}_{n,c}(\mathbf{x}), \tilde{O}_{n,c}(\mathbf{x})]$, wobei für $\tilde{U}_{n,c}(\mathbf{x})$ in (4.7) \bar{x}_n überall durch $\bar{x}_n - 1/(2n)$, und für $\tilde{O}_{n,c}(\mathbf{x})$ \bar{x}_n überall durch $\bar{x}_n + 1/(2n)$ ersetzt werden.

Bemerkung. Die obigen Intervalle hängen offenbar nur von der Summe der Erfolge s_n ab, und sind somit auch im Binomialmodell gültig.

Bemerkung. Man beachte dass \bar{x}_n auch der Maximum-Likelihood-Schätzer für p ist. Das asymptotische Konfidenzintervall erlaubt es, zusätzlich zur Schätzung deren Unsicherheit zu quantifizieren. Ein analoges Vorgehen ist möglich, wenn ein asymptotisch normalverteilter Schätzer für den Parameter existiert, etwa für das λ der Poisson-Verteilung.

Bemerkung (Qualität von Konfidenzintervallen). Konfidenzintervalle sollten, gegeben sie halten das Niveau (zumindest asymptotisch) ein, eine möglichst geringe Länge haben, damit sie informativ sind. Trivialerweise ist $I_n = (0, 1)$ ein Konfidenzintervall für p zu jedem Niveau, aber wenig informativ. Die Länge der oben konstruierten asymptotischen Konfidenzintervalle konvergiert mit wachsender Stichprobengröße mit der Geschwindigkeit $1/\sqrt{n}$ gegen 0, gleichmäßig in \mathbf{x} da $q(1 - q) \leq 1/4$, $q \in (0, 1)$.

Man kann daher die Stichprobengröße so wählen, dass zu gegebenem Niveau $1 - \alpha$ die Länge des Intervalls unterhalb einer festen Größe 2β liegt. Da $q(1 - q) \leq 1/4$, $q \in (0, 1)$, ist die Länge von

$[\tilde{U}_n(\mathbf{x}), \tilde{O}_n(\mathbf{x})]$ beschränkt durch $q_{1-\alpha/2}/\sqrt{n}$, so dass man $n \geq q_{1-\alpha/2}^2/4\beta^2$ wählen muss. So benötigt man bei einem Konfidenzniveau von 95% einen Stichprobenumfang von mindestens 2401, um eine Abweichung von höchstens 2% zu erhalten. Dieses Ergebnis wird regelmäßig bei der “Sonntagsfrage” nach der Zustimmung zu den einzelnen Parteien angewandt: Es soll garantiert werden, dass die Zustimmung bis auf 2% Abweichung genau ermittelt wurde.

4.5 Zweistichprobenprobleme und t -Test

Die oben diskutierten Tests beschränken sich auf nur eine Stichprobe und darauf, auf spezielle Werte eines Parameters zu testen. Statt dessen hat man in der Praxis aber oft zwei Stichproben, die man miteinander vergleichen möchte. Wir wollen dies an zwei typischen und zentralen Beispielen illustrieren.

Beispiel 4.46 (unverbundenes Zweistichprobenproblem). Der sogenannte **Pygmalioneffekt** (auch Rosenthal- oder Versuchsleitererwartungseffekt) beschreibt ein psychologisches Phänomen, bei dem sich Erwartungen und Vorurteile eines Versuchsleiters, zum Beispiel eines Lehrers, als „selbst-erfüllende“ Propheteiung bewahrheiten. Dabei soll sich die Erwartungshaltung so auswirken, dass dadurch die Behandlung und der Einfluss des Leiters zu dem erwarteten Ergebnis führen. Der Effekt geht auf ein Experiment zurück, in dem Lehrern zufällig ausgewählte Schüler als hochbegabte Schüler vorgestellt wurden (Rosenthal–Jacobsen 1968). Nach acht Monaten Unterricht wurde ein (vom Lehrer unabhängiger) Leistungstest durchgeführt und die vorher ausgesuchten angeblich hochbegabten Schüler mit den „gewöhnlichen“ verglichen. Die Testergebnisse bestehen aus einem Score (je höher desto besser) und sind in der folgenden Tabelle für Erstklässler aufgeführt.

hochbegabt	35 12	40 39	12 19	15 25	21 22	14 1	46 34	10 3	28 1	48 2	16 3	30 2	32 1	48 2	31
anderen	2 3	27 29	38 37	31 2	1 1	19 34	1 3	34 1	3 2	1 3	2 2	1 1	2 2	1 1	2 2

Wir nehmen an, dass die Scores $(X_i)_{1 \leq i \leq n}$ der ersten Gruppe u.i.v. $\mathcal{N}(\mu_1, \sigma_1^2)$, mit $\sigma_1 = 12,5$, und der zweiten Gruppe $(Y_j)_{1 \leq j \leq m}$ u.i.v. $\mathcal{N}(\mu_2, \sigma_2^2)$, mit $\sigma_2 = 14,5$, verteilt sind, und in beiden Gruppen die Scores unabhängig sind. Das zweiseitige Testproblem lautet

$$H : \mu_1 = \mu_2 \quad \text{gegen} \quad K : \mu_1 \neq \mu_2.$$

Analog könnte man einseitige Testprobleme formulieren. Dann ist die Prüfgröße

$$S := \frac{n^{-1} \sum_{i=1}^n X_i - m^{-1} \sum_{j=1}^m Y_j}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

unter H standardnormalverteilt. Ein Zweistichproben-Gauß-Test lehnt nun H ab, falls $|S| > q_{1-\alpha/2}$ und lehnt nicht ab, falls $|S| \leq q_{1-\alpha/2}$, mit $q_{1-\alpha/2}$ dem $(1 - \alpha/2)$ -Quantil der Standardnormalverteilung. Die Prüfgröße ergibt für obige Daten etwa 3,55. Es lässt sich also durch den Test zum Niveau $\alpha = 0,05$ die Hypothese gleicher Gruppenerwartungswerte verwerfen, der p-Wert liegt sogar deutlich unter 1%.

Beispiel 4.47 (Verbundene Stichproben). In einem Experiment wird der Abrieb von Schuhsohlen für zwei unterschiedliche Materialien (A und B) untersucht. Dabei werden für 10 Paar Schuhe jeweils die Sohle eines Schuhs mit Material A und des anderen Schuhs mit Material B versehen. Zufällig wird bestimmt, ob der linke oder rechte Schuh die Sohle mit Material A bekommt, um möglichst keinen Effekt der Seiten durch ein typisch menschliches Geh-Verhalten zu haben. Zehn Testpersonen tragen die Schuhe eine gewisse Zeit und anschließend wird der Abrieb gemessen. Es ergaben sich folgende Werte:

Abrieb Material A	14,0	8,8	11,2	14,2	11,8	6,4	9,8	11,3	9,3	13,6
Abrieb Material B	13,2	8,2	10,9	14,3	10,7	6,6	9,5	10,8	8,8	13,3

Untereinander stehende Werte sind jeweils der Schuh mit Sohle aus Material A und Material B der gleichen Testperson. Man spricht hier von verbundenen Stichproben, da die Werte Abrieb Material A/Abrieb Material B zwischen den beiden Stichproben im Allgemeinen nicht unabhängig sind. Verbundene Stichproben sind also Daten, die von den gleichen Versuchspersonen (z. Bsp. Patienten) erhoben wurden. Ein klassisches Beispiel sind Laborwerte einer bestimmten Kenngröße an Patienten vor und nach einer Behandlung mit den beiden Stichproben: (X_i) Laborwerte vor der Behandlung und (Y_i) Laborwerte nach der Behandlung. Diese beiden Stichproben sind verbunden, da jede Stichprobe von jedem Patienten einen Laborwert enthält. Unter Normalverteilungsannahme ist eine Lösung für dieses verbundene Zweistichprobenproblem simpel: Wir betrachten direkt die Differenzen der Abriebe bei Material A und der Abriebe bei Material B ($X_i - Y_i$) $_{1 \leq i \leq n}$, bei verbundenen Stichproben ist ja stets $n = m$, und testen in dem Modell $(X_i - Y_i) \stackrel{u.i.v.}{\sim} \mathcal{N}(\mu, \sigma^2)$ mit bekanntem $\sigma = 0,4$ die Hypothese

$$H : \mu = 0 \quad \text{gegen} \quad K : \mu \neq 0.$$

Die Prüfgröße $\sqrt{n}\sigma^{-1} \sum_{i=1}^n (X_i - Y_i)$ ist unter H standardnormalverteilt. Für vorliegende Daten erhalten wir etwa 3,24, und können H mit einem p-Wert von unter 1% ablehnen.

Bei beiden Beispielen haben wir relativ starke Annahmen vorausgesetzt.

1. Die **Normalverteilungsannahme**. Bemerke dazu dass sehr allgemein nach dem zentralen Grenzwertsatz dies zumindest approximativ für große Stichproben eine gute Näherung liefert. Es spricht hier zumindest nichts für ein anderes Modell oder gegen eine symmetrische Verteilung.
2. Die **Unabhängigkeit**. In beiden Beispielen scheint dies natürlich. Ohne die Voraussetzung der Unabhängigkeit ist eine statistische Analyse schwer bzw. nicht aussagekräftig. Es ist daher vor allem bei der Erhebung von Daten ein Hauptfokus nicht abhängige Messungen zu erheben.
3. Wir haben jeweils die **Varianzen** σ_1^2, σ_2^2 als **bekannt** vorausgesetzt. Die Werte entsprachen etwa den mittleren quadratischen Abweichungen der Werte von ihrem Mittelwert (den empirischen Varianzen). Tatsächlich würde man praktisch in den Beispielen besser den sogenannten **t-Test** verwenden. Man kann bei unbekannten Varianzen mit der geschätzten Varianz skalieren, was insbesondere für kleine Stichproben aber zu einer Abweichung der Verteilung der Prüfgröße von der Normalverteilung führt (nämlich zur sogenannten **t-Verteilung**). Die p-Werte sind dann größer, da die t-Verteilung langsamer abfallende Tails besitzt.

Den dritten Punkt wollen wir nun durch die Herleitung des t -Tests lösen. Beim Gauß-Test wurde verwendet dass

$$\sqrt{n} \frac{\bar{x}_n - \mu_0}{\sigma} \sim \mathcal{N}(0, 1),$$

für den Mittelwert $\bar{x}_n = n^{-1} \sum_{i=1}^n X_i$, falls $X_i \sim \mathcal{N}(\mu_0, \sigma^2)$ u.i.v. sind. Der Standardschätzer für eine unbekannte Varianz ist

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{x}_n)^2.$$

Wir wollen dessen Verteilung und die von der Prüfgröße

$$T = \sqrt{n} \frac{\bar{x}_n - \mu_0}{\hat{\sigma}}$$

bestimmen mit dem Ziel, Konfidenzintervalle und Tests angeben zu können. Hierbei ist $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$. Wir stellen zunächst fest, dass

$$\begin{aligned} \sum_{i=1}^n \frac{(X_i - \bar{x}_n)^2}{\sigma^2} &= \sum_{i=1}^n \left(\frac{X_i - \mu_0}{\sigma} - \frac{\bar{x}_n - \mu_0}{\sigma} \right)^2 \\ &= \sum_{i=1}^n (Z_i - \bar{z}_n)^2, \text{ wobei } Z_i \stackrel{u.i.v.}{\sim} \mathcal{N}(0, 1). \end{aligned}$$

Wir benutzen die Zerlegung

$$\sum_{i=1}^n Z_i^2 = \sum_{i=1}^n (Z_i - \bar{z}_n)^2 + n\bar{z}_n^2, \quad (4.8)$$

und zeigen dass rechts die Summe von zwei *unabhängigen* Zufallsvariablen steht. Für jedes $i \in \{1, \dots, n\}$ ist

$$\left(\begin{array}{cccc} -\frac{1}{n} & \dots & \overbrace{1 - \frac{1}{n}}^{\text{ite Spalte}} & \dots & -\frac{1}{n} \\ \frac{1}{n} & \dots & \dots & \dots & \frac{1}{n} \end{array} \right) \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} Z_1 - \bar{z}_n \\ \vdots \\ \bar{z}_n \end{pmatrix},$$

und da $\text{Cov}(Z_i - \bar{z}_n, \bar{z}_n) = 0$ folgt aus Satz 2.37 mit Satz 2.24 insbesondere die Unabhängigkeit von $Z_i - \bar{z}_n$ und \bar{z}_n für jedes i .¹ Die Verteilung von Z_1^2 , für $Z_1 \sim \mathcal{N}(0, 1)$, heißt χ_1^2 -Verteilung, die χ^2 -Verteilung mit einem Freiheitsgrad. Die Verteilung von $\sum_{i=1}^n Z_i^2$ ist die χ_n^2 -Verteilung, die χ^2 -Verteilung mit n Freiheitsgraden. Wir leiten die charakteristische Funktion von Z_1^2 her, welche die χ_1^2 -Verteilung eindeutig charakterisiert:

$$\begin{aligned} \varphi_{Z_1^2}(u) &= \mathbb{E}[e^{iuZ_1^2}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{iux^2 - x^2/2} dx \\ &= \sqrt{\frac{2}{\pi}} \int_0^{\infty} e^{-x^2(1/2 - iu)} dx \\ &= \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{1/2 - iu}} \int_0^{\infty} e^{-t^2} dt \\ &= \frac{1}{\sqrt{1 - 2iu}}. \end{aligned}$$

Das bekannte Integral im letzten Schritt folgt aus der Normiertheit der Dichte von $\mathcal{N}(0, 1)$. Die charakteristische Funktion der χ_n^2 -Verteilung ergibt sich direkt aus der Produktform von charakteristischen Funktionen bei Summen unabhängiger Zufallsvariablen:

$$\varphi_{\sum_{i=1}^n Z_i^2}(u) = (1 - 2iu)^{-n/2}.$$

Aus (4.8), mit der Unabhängigkeit der Summanden, ergibt sich direkt

$$\varphi_{\sum_{i=1}^n (Z_i - \bar{z}_n)^2}(u) = (1 - 2iu)^{-(n-1)/2},$$

also dass

$$\sum_{i=1}^n (Z_i - \bar{z}_n)^2 = \sum_{i=1}^n \frac{(X_i - \bar{x}_n)^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Auch wenn die charakteristische Funktion die Verteilung eindeutig bestimmt, wollen wir auch noch die Dichte der χ_n^2 -Verteilung herleiten. Bemerke, dass generell die Dichte einer Zufallsvariablen $V = Z^2$ sich aus der Dichte von Z wie folgt ergibt:

$$f_V(x) = \frac{f_Z(\sqrt{x})}{2\sqrt{x}} = \frac{f_Z(\sqrt{x}) + f_Z(-\sqrt{x})}{2\sqrt{x}}.$$

Mit der Symmetrie ergibt sich also für $Z \sim \mathcal{N}(0, 1)$:

$$f_V(x) = \frac{1}{\sqrt{2\pi}} e^{-x/2} \frac{1}{\sqrt{x}} \mathbf{1}_{[0, \infty)}(x).$$

Die Dichte der χ_n^2 -Verteilung zeigen wir per Induktion durch Anwendung der Faltungsformel. Es ist für

¹vgl. auch Satz 2.18, 4.

$V_n = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$:

$$f_{V_n}(x) = \frac{x^{n/2-1} e^{-x/2}}{2^{n/2} \Gamma(n/2)} \mathbf{1}_{[0,\infty)}(x),$$

wobei Γ die **Eulersche Gammafunktion**² bezeichnet. Der Induktionsschritt

$$\begin{aligned} f_{V_{n+1}}(x) &= \int_0^\infty f_{V_n}(z) f_{V_1}(x-z) dz \\ &= \int_0^\infty \frac{z^{n/2-1} e^{-z/2}}{2^{n/2} \Gamma(n/2)} \mathbf{1}_{[0,\infty)}(z) \frac{e^{-(x-z)/2} (x-z)^{-1/2}}{\sqrt{2\Gamma(1/2)}} \mathbf{1}_{[0,\infty)}(x-z) dz \\ &= \int_0^x \frac{z^{n/2-1} e^{-x/2} (x-z)^{-1/2}}{2^{(n+1)/2} \Gamma(n/2) \Gamma(1/2)} dz \mathbf{1}_{[0,\infty)}(x) \\ &\quad \overbrace{=}^{=B(n/2, 1/2)} \frac{e^{-x/2} x^{(n+1)/2-1} \int_0^1 u^{n/2-1} (1-u)^{-1/2} du}{\Gamma(n/2) \Gamma(1/2)} \mathbf{1}_{[0,\infty)}(x) \\ &= \frac{x^{(n+1)/2-1} e^{-x/2}}{2^{(n+1)/2} \Gamma((n+1)/2)} \mathbf{1}_{[0,\infty)}(x) \end{aligned}$$

beweist die Behauptung. Wir haben die Integraldarstellung der **Eulerschen Betafunktion**, $B(n/2, 1/2)$, und ihre Beziehung zur Gammafunktion genutzt.

Mit der Unabhängigkeit kennen wir also bereits die gemeinsame Dichte von $(\bar{z}_n, \sum_{i=1}^n (Z_i - \bar{z}_n)^2)$:

$$f_{\bar{z}_n, V_{n-1}}(z, v) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \frac{v^{n/2-3/2} e^{-v/2}}{2^{(n-1)/2} \Gamma((n-1)/2)} \mathbf{1}_{[0,\infty)}(v).$$

Daraus leiten wir die Verteilung von

$$T = \frac{\bar{z}_n}{\sqrt{\sum_{i=1}^n (Z_i - \bar{z}_n)^2 / (n-1)}} = \frac{(\bar{x}_n - \mu_0) / \sigma}{\sqrt{\sum_{i=1}^n \frac{(X_i - \bar{x}_n)^2}{(n-1)\sigma^2}}}$$

her. Mit der Transformation

$$z = \frac{t\sqrt{v}}{\sqrt{n-1}} \text{ aus } t = \frac{z}{\sqrt{v/(n-1)}}$$

bekommen wir mit der Abbildung $(z, v) \mapsto (t\sqrt{v}/\sqrt{n-1}, v)$ die gemeinsame Dichte von $(T, \sum_{i=1}^n (Z_i - \bar{z}_n)^2)$:

$$\begin{aligned} f_{T, \sum_{i=1}^n (Z_i - \bar{z}_n)^2}(t, v) &= f_{\bar{z}_n, \sum_{i=1}^n (Z_i - \bar{z}_n)^2}\left(\frac{t\sqrt{v}}{\sqrt{n-1}}, v\right) \det\begin{pmatrix} \frac{\sqrt{v}}{\sqrt{n-1}} & \frac{t}{2\sqrt{v(n-1)}} \\ 0 & 1 \end{pmatrix} \\ &= \sqrt{\frac{v}{n-1}} \frac{e^{-v/2-t^2v/(n-1)} v^{n/2-3/2}}{\sqrt{2\pi} 2^{(n-1)/2} \Gamma((n-1)/2)} \\ &= \frac{\exp(-v/2(1+t^2/(n-1))) v^{n/2-1}}{\sqrt{2\pi(n-1)} 2^{(n-1)/2} \Gamma((n-1)/2)}. \end{aligned}$$

Wir nutzen die Integraldarstellung der Gammafunktion, $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$. Durch Integration nach v erhalten wir die Marginaldichte von T , welche die Verteilung eindeutig bestimmt:

$$f_T(t) = \int_0^\infty \frac{\exp(-v/2(1+t^2/(n-1))) v^{n/2-1}}{\sqrt{2\pi(n-1)} 2^{(n-1)/2} \Gamma((n-1)/2)} dv$$

²wiki: Γ -Funktion

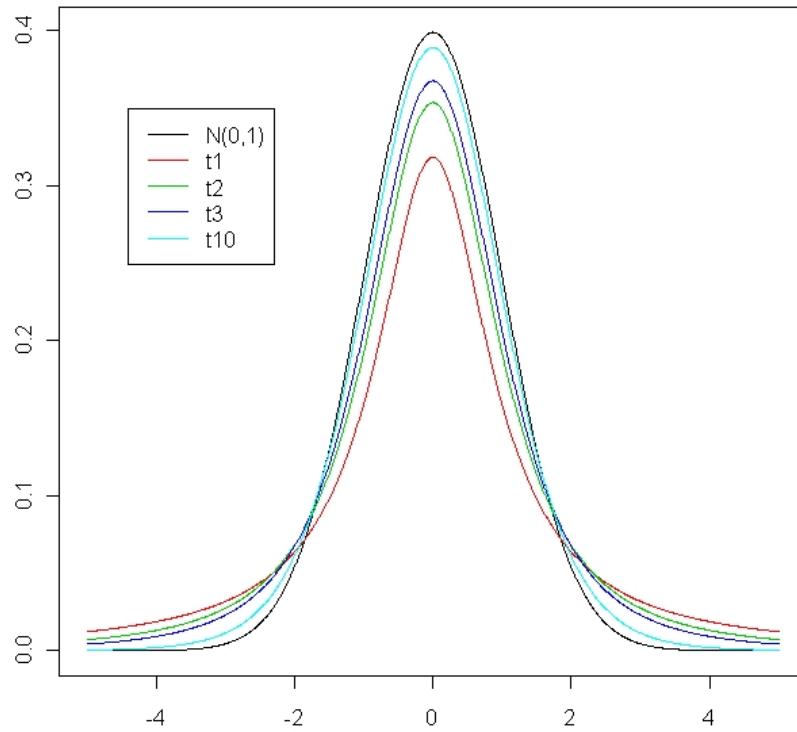


Abbildung 4.4: Dichten der t_n -Verteilungen für unterschiedliche n und der Standardnormalverteilung.

$$\begin{aligned}
 &= \frac{1}{\sqrt{2\pi(n-1)} 2^{(n-1)/2} \Gamma((n-1)/2)} \int_0^\infty e^{-w} w^{n/2-1} \left(\frac{2}{1+t^2/(n-1)} \right)^{n/2} dw \\
 &= \frac{1}{\sqrt{2\pi(n-1)} 2^{(n-1)/2} \Gamma((n-1)/2)} \left(\frac{1+t^2/(n-1)}{2} \right)^{-n/2} \Gamma(n/2) \\
 &= \frac{1}{\sqrt{\pi(n-1)} \Gamma((n-1)/2)} \left(1+t^2/(n-1) \right)^{-n/2} \Gamma(n/2).
 \end{aligned}$$

Definition 4.48. Die **t -Verteilung** (oder Student- t -Verteilung^a) mit $n \in \mathbb{N}$ Freiheitsgraden auf $(\mathbb{R}, \mathfrak{B}_{\mathbb{R}})$ ist eine absolutstetige Verteilung mit der Lebesgue-dichte

$$t_n(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}} \left(1 + \frac{x^2}{n} \right)^{-(n+1)/2}, \quad x \in \mathbb{R}.$$

^aDer Name erklärt sich durch eine lebenswerte Anekdote, siehe [wiki: Gosset](#).

Einstichproben- t -Test

Sind X_1, \dots, X_n u.i.v. $\mathcal{N}(\mu, \sigma^2)$. Der t -Test liefert eine Methode, Hypothesen $H_0 : \mu = \mu_0$ (oder einseitige Hypothesen) für den Erwartungswert zu testen. Die Varianz $\sigma^2 > 0$ ist unbekannt und wird durch $\hat{\sigma}^2$

geschätzt. Als Prüfgröße betrachtet man

$$T(x) = \frac{\sqrt{n(n-1)}(\bar{x}_n - \mu_0)}{\sqrt{\sum_i (x_i - \bar{x}_n)^2}}.$$

Die Zufallsvariable T hat unter der Hypothese H_0 eine t_{n-1} -Verteilung. Für eine größer werdende Anzahl an Freiheitsgraden n nähert sich die t_n -Verteilung an die Standardnormalverteilung an, siehe Abbildung 4.4. Im Vergleich zur Normalverteilung haben die t -Verteilungen langsamere abfallende Tails. Sie sind ebenfalls symmetrisch.

Folgende Tabelle enthält die entsprechenden Ablehnungsbereiche des t -Tests zum Niveau α für die einseitigen bzw. zweiseitigen Hypothesen:

H_0	Ablehnungsbereich
$\mu \leq \mu_0$	$\{x \in \mathcal{X} T(x) > t_{1-\alpha, n-1}\}$
$\mu \geq \mu_0$	$\{x \in \mathcal{X} T(x) < -t_{1-\alpha, n-1}\}$
$\mu = \mu_0$	$\{x \in \mathcal{X} T(x) > t_{1-\frac{\alpha}{2}, n-1}\}$

Auf verbundene Zweistichprobenprobleme wie in Beispiel 4.47 kann der Einstichproben- t -Test angewendet werden.

Zweistichproben- t -Test

Seien X_1, \dots, X_n u.i.v. $\mathcal{N}(\mu_1, \sigma^2)$ und Y_1, \dots, Y_m u.i.v. $\mathcal{N}(\mu_2, \sigma^2)$. Der Zweistichproben- t -Test basiert auf einem Vergleich der Mittelwerte. Die Prüfgröße ist

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{1}{m} + \frac{1}{n}} \sqrt{\frac{1}{m+n-2} (\sum_i (X_i - \bar{x}_n)^2 + \sum_j (Y_j - \bar{y}_m)^2)}}.$$

Unter der Hypothese $H_0 : \mu_1 = \mu_2$ kann man zeigen, dass diese t -verteilt ist mit $(m+n-2)$ Freiheitsgraden. Die Hypothese $H_0 : \mu_1 = \mu_2$ wird dann verworfen, falls

$$|T(\mathbf{X}, \mathbf{Y})| > t_{m+n-2, 1-\alpha/2}.$$

Während der Zweistichproben-Fall unter der Annahme gleicher unbekannter Varianzen sich leicht als Verallgemeinerung des Einstichproben- t -Tests ergibt, führt der heteroskedastische Ansatz, X_1, \dots, X_n u.i.v. $\mathcal{N}(\mu_1, \sigma_1^2)$ und Y_1, \dots, Y_m u.i.v. $\mathcal{N}(\mu_2, \sigma_2^2)$ mit $\sigma_1 \neq \sigma_2$ (und unbekannten Quotienten und $n \neq m$) auf das **Behrens-Fisher-Problem**. Es lässt sich in dem Fall kein exakter Niveau- α -Test bestimmen. Numerische Näherungsmethoden sind praktisch jedoch ausreichend gut und heute, zum Beispiel in R, verfügbar. Ein Lösungsansatz (der in R zur Verfügung steht) bietet der sogenannte Welch-Test. Dieser führt die Bestimmung des kritischen Wertes bei festem Niveau α auf die Lösung partieller Differentialgleichungen unendlicher Ordnung zurück. Durch Taylorentwicklungen wird eine approximative Lösung ermittelt. Eine Anwendung auf das Beispiel 4.46 ergibt dann einen p-Wert von etwa 1,1%, also etwas größer als mit obigem Gauß-Test.

5 Grenzwertsätze der Stochastik

5.1 Univariate zentrale Grenzwertsätze

Der zentrale Grenzwertsatz (ZGWS) beschreibt die asymptotische Normalverteilung des Mittelwertes um den Erwartungswert herum, wenn man geeignet reskaliert. Er ist ein grundlegendes Resultat der Wahrscheinlichkeitstheorie mit großer Bedeutung in der Statistik. In diesem Abschnitt beweisen wir diesen Satz mit dem Stetigkeitssatz von Lévy, wozu noch der Zusammenhang zwischen den Momenten einer Zufallsvariablen und der Taylorentwicklung der zugehörigen charakteristischen Funktion in Null wichtig ist. Weiter zeigen wir, wie sich mit der Δ -Methode die asymptotische Normalität auf Funktionen des Erwartungswertes erweitern lässt. Anschließend geben wir den zentralen Grenzwertsatz für Dreiecksschemata unter der Lindeberg-Bedingung an. Wir erinnern zunächst an das einfachste asymptotische Resultat der Stochastik, das **schwache Gesetz der großen Zahlen** mit der Tschebyschev-Ungleichung: Sind X_1, \dots, X_n unabhängig identisch verteilt mit $\mathbb{E}[X_i^2] < \infty$, so gilt für $n \rightarrow \infty$:

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\mathbb{P}} \mathbb{E}[X_1]$$

Mit $S_n = \sum_{i=1}^n X_i$, ist

$$\mathbb{E}[S_n] = \sum_{i=1}^n \mathbb{E}[X_i] = n\mathbb{E}[X_1].$$

und daher

$$\mathbb{E}\left[\frac{S_n}{n}\right] = \mathbb{E}[X_1], \quad \text{Var}\left(\frac{S_n}{n}\right) = \frac{\text{Var}(X_1)}{n}.$$

Also gilt mit der Tschebyschev-Ungleichung:

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mathbb{E}[X_1]\right| \geq \epsilon\right) \leq \frac{1}{\epsilon^2} \text{Var}\left(\frac{S_n}{n}\right) = \frac{1}{\epsilon^2} \cdot \frac{\text{Var}(X_1)}{n}.$$

Tatsächlich zeigt dieser Beweis mehr als das Resultat, nämlich $n^\alpha |n^{-1} \sum_{i=1}^n X_i - \mathbb{E}[X_1]| \xrightarrow{\mathbb{P}} 0$ für u.i.v. L_2 -Zufallsvariablen, für alle $\alpha < 1/2$. Beim zentralen Grenzwertsatz reskalieren wir gerade mit $n^{1/2}$, sozusagen der Geschwindigkeit für das obige Gesetz der großen Zahlen. Es sei daran erinnert, dass stärkere Resultate als obiges über Konvergenz von Mittelwerten gelten. So gilt bereits für u.i.v. L_1 -Zufallsvariablen das **Starke Gesetz der großen Zahlen von Khinchine** welches wir am Ende der Stochastik I bewiesen haben.

Korollar 5.1. Sei X eine reellwertige Zufallsvariable mit charakteristischer Funktion φ . Gilt $\mathbb{E}[|X|^n] < \infty$, so folgt

$$\varphi(t) = \sum_{k=0}^n \frac{(it)^k \mathbb{E}[X^k]}{k!} + o(|t|^n), \quad \text{für } t \rightarrow 0.$$

Insbesondere ist φ in 0 n -mal differenzierbar mit $\varphi^{(k)}(0) = i^k \mathbb{E}[X^k]$, $k = 1, \dots, n$.

Das Resultat folgt mit Satz 1.23 und wird in den Übungen gezeigt.

Zentraler Grenzwertsatz für u.i.v. Zufallsvariablen.

Satz 5.2 (ZGWS nach Lindeberg-Lévy). Seien $(X_n)_{n \geq 1}$ u.i.v. reellwertige Zufallsvariablen mit

$\mathbb{E}[X_1^2] < \infty$, und sei $\mu = \mathbb{E}[X_1]$, $\sigma^2 = \text{Var}(X_1)$. Angenommen, $\sigma^2 > 0$. Dann gilt für $S_n = X_1 + \dots + X_n$

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{oder äquivalent} \quad \sqrt{n} \left(\frac{S_n}{n} - \mu \right) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

■

Beweis. Wir geben zunächst eine Beweisskizze. Durch Übergang zu $(X_n - \mu)/\sigma$ können wir o.E.d.A. $\mu = 0$ und $\sigma = 1$ annehmen. Für die charakteristische Funktion φ von X_k gilt dann nach Korollar 5.1 für $n = 2$ dass $\varphi(s) = 1 - \frac{s^2}{2} + o(s^2)$, $s \rightarrow 0$. Für festes t erhalten wir mit $s = t/\sqrt{n}$

$$\varphi(t/\sqrt{n}) = 1 - \frac{t^2}{2n} + o(n^{-1}), \quad n \rightarrow \infty. \quad (5.1)$$

Damit folgern wir

$$\begin{aligned} \varphi_{S_n/\sqrt{n}}(t) &= \mathbb{E}[e^{iS_n(t/\sqrt{n})}] \\ &= (\varphi(t/\sqrt{n}))^n = \left(1 - \frac{t^2}{2n} + o(n^{-1})\right)^n \rightarrow e^{-t^2/2}, \end{aligned} \quad (5.2)$$

und der ZGWS folgt mit dem Stetigkeitssatz in Satz 3.34. ■

Die ‘‘Skizze’’ in obigem Argument liegt darin, dass der Restterm in der Entwicklung $1 - \frac{t^2}{2n} + o(n^{-1})$ von $\varphi(t)$ komplexwertig ist, und somit muss der letzte Grenzübergang in (5.2) für komplexes Argument richtig sein.

Statt diesen allgemein zu zeigen, kann man wie folgt argumentieren. Wir benutzen

Lemma 5.3. Sind $z_1, \dots, z_n, w_1, \dots, w_n \in \mathbb{C}$ mit $|z_j|, |w_j| \leq 1$, so gilt

$$|z_1 \cdot \dots \cdot z_n - w_1 \cdot \dots \cdot w_n| \leq \sum_{k=1}^n |z_k - w_k|.$$

Beweis.

$$\begin{aligned} z_1 \cdot \dots \cdot z_n - w_1 \cdot \dots \cdot w_n &= z_1 \cdot \dots \cdot z_n - w_1 \cdot z_2 \cdot \dots \cdot z_n + w_1 \cdot z_2 \cdot \dots \cdot z_n - w_1 \cdot \dots \cdot w_n \\ &= (z_1 - w_1) \cdot z_2 \cdot \dots \cdot z_n + w_1 \cdot (z_2 \cdot \dots \cdot z_n - w_2 \cdot \dots \cdot w_n), \end{aligned}$$

also

$$|z_1 \cdot \dots \cdot z_n - w_1 \cdot \dots \cdot w_n| \leq |z_1 - w_1| + |z_2 \cdot \dots \cdot z_n - w_2 \cdot \dots \cdot w_n|,$$

und die Behauptung folgt mit Induktion. ■

Wir steigen noch einmal in den Beweis des zentralen Grenzwertsatzes ein. Es ist

$$|\varphi_{S_n/\sqrt{n}}(t) - e^{-t^2/2}| \leq |(\varphi(t/\sqrt{n}))^n - (1 - \frac{t^2}{2n})^n| + |(1 - \frac{t^2}{2n})^n - e^{-t^2/2}|,$$

und der zweite Summand auf der rechten Seite konvergiert gegen 0 für $n \rightarrow \infty$ (hier nur reelles Argument). Für den ersten Term ist nach obigem Lemma und (5.1)

$$|(\varphi(t/\sqrt{n}))^n - (1 - \frac{t^2}{2n})^n| \leq n |\varphi(t/\sqrt{n}) - (1 - \frac{t^2}{2n})| = o(1).$$

Beispiel 5.4. Asymptotik für empirische Verteilungsfunktion.

Sind X_1, X_2, \dots u.i.v. mit Verteilungsfunktion F , und

$$\hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{(-\infty, x]}(X_k),$$

so gilt

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{d} \mathcal{N}(0, F(x)(1 - F(x))).$$

Delta-Methode

Satz 5.5 (Delta-Methode). Seien X_n, X reellwertige Zufallsvariablen und $\mu, a_n \in \mathbb{R}$ mit $a_n \rightarrow \infty$ derart, dass

$$a_n(X_n - \mu) \xrightarrow{d} X,$$

und ist $f : I \rightarrow \mathbb{R}$ eine messbare, in μ differenzierbare Funktion mit $\mathbb{P}(X_n \in I) = 1$, so folgt

$$a_n(f(X_n) - f(\mu)) \xrightarrow{d} f'(\mu)X.$$

■

Beweis. Nach dem Satz 3.22 von Skorochod existieren Zufallsvariablen Y_n, Y auf einem gemeinsamen Wahrscheinlichkeitsraum mit $Y_n \xrightarrow{d} X_n, Y \xrightarrow{d} X$ und

$$a_n(Y_n - \mu) \rightarrow Y, \quad \text{fast sicher.}$$

Da $a_n \rightarrow \infty$, gilt insbesondere $Y_n - \mu \rightarrow 0$ fast sicher. Da

$$f(\mu + h) - f(\mu) = f'(\mu)h + o(|h|), \quad |h| \rightarrow 0,$$

existiert für fast alle ω eine Folge $c_n = c_n(\omega) \rightarrow 0$, so dass

$$f(Y_n(\omega)) - f(\mu) = f'(\mu)(Y_n(\omega) - \mu) + c_n|Y_n(\omega) - \mu|$$

und daher für fast alle ω

$$a_n(f(Y_n(\omega)) - f(\mu) - f'(\mu)(Y_n(\omega) - \mu)) = c_n|a_n(Y_n(\omega) - \mu)| \rightarrow 0.$$

Somit folgt

$$a_n(f(Y_n) - f(\mu)) \rightarrow f'(\mu)Y, \quad \text{fast sicher,}$$

und insbesondere die schwache Konvergenz. ■

Die Delta-Methode wird insbesondere in folgender Situation angewendet: Gilt

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

so folgt unter den obigen Annahmen an f , dass

$$\sqrt{n}(f(X_n) - f(\mu)) \xrightarrow{d} \mathcal{N}(0, f'(\mu)^2 \sigma^2).$$

Beispiel 5.6. (Varianz-stabilisierende Transformationen)

1. Seien X_1, X_2, \dots unabhängig und $\text{Poi}(\lambda)$ -verteilt und sei $S_n = X_1 + \dots + X_n$. Dann ist

$$\sqrt{n}(S_n/n - \lambda) \xrightarrow{d} \mathcal{N}(0, \lambda).$$

Setze $f(x) = \sqrt{x}$, dann ist $f'(x) = 1/(2\sqrt{x})$, und

$$(\sqrt{S_n} - \sqrt{n\lambda}) \xrightarrow{d} \mathcal{N}(0, 1/4).$$

2. Seien X_1, X_2, \dots unabhängig und $Ber(p)$ -verteilt und sei $S_n = X_1 + \dots + X_n$. Dann gilt

$$\sqrt{n}(S_n/n - p) \xrightarrow{d} \mathcal{N}(0, p(1-p)).$$

Setze $f(x) = \arcsin(\sqrt{x})$, dann ist $f'(x) = 1/(2\sqrt{x}\sqrt{1-x})$, und

$$\sqrt{n}(\arcsin(\sqrt{S_n/n}) - \arcsin(\sqrt{p})) \xrightarrow{d} \mathcal{N}(0, 1/4).$$

Zentraler Grenzwertsatz für Dreiecksschemata

Unter einem (quadratintegrierbaren, zentrierten) **Dreiecksschema** von Zufallsvariablen verstehen wir reellwertige Zufallsvariablen $(X_{n,k})_{n \geq 1, 1 \leq k \leq r_n}$, $r_n \in \mathbb{N}$, $r_n \rightarrow \infty$ für $n \rightarrow \infty$, so dass

1. $\mathbb{E}[X_{n,k}] = 0$, $\mathbb{E}[X_{n,k}^2] = \sigma_{n,k}^2 < \infty$,
2. für alle n sind $X_{n,1}, \dots, X_{n,r_n}$ unabhängig.

Für $S_n = X_{n,1} + \dots + X_{n,r_n}$ ist dann

$$s_n^2 = \sigma_{n,1}^2 + \dots + \sigma_{n,r_n}^2.$$

Das Dreiecksschema genügt der **Lindeberg-Bedingung**, falls für alle $\varepsilon > 0$ gilt

$$\lim_{n \rightarrow \infty} \frac{1}{s_n^2} \sum_{k=1}^{r_n} \int_{\{|X_{n,k}| \geq \varepsilon s_n\}} X_{n,k}^2 d\mathbb{P} = 0. \quad (5.3)$$

Das Dreiecksschema genügt der **Lyapunov-Bedingung**, falls ein $\delta > 0$ existiert mit $\mathbb{E}[|X_{n,k}|^{2+\delta}] < \infty$, und

$$\lim_{n \rightarrow \infty} \frac{1}{s_n^{2+\delta}} \sum_{k=1}^{r_n} \mathbb{E}[|X_{n,k}|^{2+\delta}] = 0. \quad (5.4)$$

Lemma 5.7. Sei $(X_{n,k})_{n \geq 1, 1 \leq k \leq r_n}$, $r_n \in \mathbb{N}$ ein quadratintegrierbares, zentriertes Dreiecksschema.

1. Genügt es der Lyapunov-Bedingung (5.4), so auch der Lindeberg-Bedingung (5.3).
2. Genügt es der Lindeberg-Bedingung, so folgt

$$\max_{1 \leq k \leq r_n} \frac{\sigma_{n,k}^2}{s_n^2} \rightarrow 0, \quad n \rightarrow \infty.$$

Diese Eigenschaft nennt man die **Feller-Bedingung** (oder auch asymptotisch vernachlässigbar).

Beweis. Zu 1.: Es ist

$$\int_{\{|X_{n,k}| \geq \varepsilon s_n\}} X_{n,k}^2 d\mathbb{P} \leq \int_{\{|X_{n,k}| \geq \varepsilon s_n\}} \frac{|X_{n,k}|^{2+\delta}}{(\varepsilon s_n)^\delta} d\mathbb{P} \leq \frac{1}{\varepsilon^\delta} \mathbb{E}[|X_{n,k}|^{2+\delta}].$$

Zu 2.: Für $1 \leq k \leq r_n$ und $\varepsilon > 0$ ist

$$\begin{aligned} \sigma_{n,k}^2 = \mathbb{E}[X_{n,k}^2] &= \int_{\{|X_{n,k}| \geq \varepsilon s_n\}} X_{n,k}^2 d\mathbb{P} + \int_{\{|X_{n,k}| < \varepsilon s_n\}} X_{n,k}^2 d\mathbb{P} \\ &= \int_{\{|X_{n,k}| \geq \varepsilon s_n\}} X_{n,k}^2 d\mathbb{P} + \varepsilon^2 s_n^2. \end{aligned}$$

Somit folgt

$$\limsup_{n \rightarrow \infty} \max_{1 \leq k \leq r_n} \frac{\sigma_{n,k}^2}{s_n^2} \leq \limsup_{n \rightarrow \infty} \left(\frac{1}{s_n^2} \max_{1 \leq k \leq r_n} \int_{\{|X_{n,k}| \geq \varepsilon s_n\}} X_{n,k}^2 d\mathbb{P} \right) + \varepsilon^2$$

$$\leq \limsup_{n \rightarrow \infty} \left(\frac{1}{s_n^2} \sum_{k=1}^{r_n} \int_{\{|X_{n,k}| \geq \varepsilon s_n\}} X_{n,k}^2 d\mathbb{P} \right) + \varepsilon^2 = \varepsilon^2.$$

Da $\varepsilon > 0$ beliebig war, folgt die Behauptung. ■

Bemerkung. Die Feller-Bedingung ist zwar intuitiv am einfachsten interpretierbar, aber nicht hinreichend für die asymptotische Normalität, wie wir im Kapitel über Poisson Konvergenz sehen werden. Ausschlaggebend für asymptotische Normalität ist, dass die Beiträge der einzelnen Summanden oberhalb der Standardabweichung der Summe asymptotisch vernachlässigbar sind, im Sinne der Lindeberg-Bedingung.

Satz 5.8 (ZGWS nach Lindeberg-Feller). Sei $(X_{n,k})_{n \geq 1, 1 \leq k \leq r_n}$, $r_n \in \mathbb{N}$ ein quadratintegrierbares, zentriertes Dreiecksschema. Genügt es der Lindeberg-Bedingung (5.3), so folgt

$$\frac{S_n}{s_n} \xrightarrow{d} \mathcal{N}(0, 1). \quad (5.5)$$
■

Für den Beweis, nachfolgende Beispiele und Diskussion siehe auch Billingsley (1994), Theorem 27.2 und nachfolgende Seiten.

Beweis. Mit dem Übergang von $X_{n,k}$ zu $X_{n,k}/s_n$ kann man o.E.d.A. den Fall $s_n^2 \rightarrow \sigma^2 = 1$ behandeln. Wir zeigen, dass falls

$$\sum_{k=1}^{r_n} \mathbb{E}[X_{n,k}^2] \rightarrow \sigma^2, \quad (5.6a)$$

$$\sum_{k=1}^{r_n} \mathbb{E}[X_{n,k}^2 \mathbf{1}_{\{|X_{n,k}| > \varepsilon\}}] \rightarrow 0, \quad (5.6b)$$

mit einer Konstanten $\sigma > 0$ und für alle $\varepsilon > 0$ gilt, $S_n \xrightarrow{d} N(0, \sigma^2)$ folgt.
Sei $\varphi_{n,k}(t) = \mathbb{E}[e^{iX_{n,k}t}]$. Wir haben zu zeigen dass

$$\prod_{k=1}^{r_n} \varphi_{n,k}(t) \rightarrow \exp(-t^2 \sigma^2 / 2), \quad n \rightarrow \infty. \quad (5.7)$$

Ähnlich wie Korollar 5.1 zeigt man die Abschätzung

$$\left| e^{itx} - \left(1 + itx - \frac{t^2 x^2}{2} \right) \right| \leq \min(|tx|^3, |tx|^2).$$

Für $\varepsilon < |t|$ folgt:

$$\begin{aligned} \left| \varphi_{n,k}(t) - \left(1 - \frac{t^2 \sigma_{n,k}^2}{2} \right) \right| &\leq \mathbb{E}[|tX_{n,k}|^3 \wedge |tX_{n,k}|^2] \\ &\leq \mathbb{E}[|tX_{n,k}|^3 \mathbf{1}_{\{|X_{n,k}| < \varepsilon\}}] + \mathbb{E}[|tX_{n,k}|^2 \mathbf{1}_{\{|X_{n,k}| \geq \varepsilon\}}] \\ &\leq \varepsilon t^3 \mathbb{E}[|tX_{n,k}|^2 \mathbf{1}_{\{|X_{n,k}| < \varepsilon\}}] + \mathbb{E}[|tX_{n,k}|^2 \mathbf{1}_{\{|X_{n,k}| \geq \varepsilon\}}]. \end{aligned}$$

Aus der Lindeberg-Bedingung (5.6b) sowie mit (5.6a) folgt dann

$$\limsup_{n \rightarrow \infty} \sum_{k=1}^{r_n} \left| \varphi_{n,k}(t) - \left(1 - \frac{t^2 \sigma_{n,k}^2}{2} \right) \right| \leq \varepsilon |t|^3 \sigma^2.$$

Damit konvergiert die Summe gegen Null und Lemma 5.3 liefert

$$\prod_{k=1}^{r_n} \varphi_{n,k}(t) - \prod_{k=1}^{r_n} \left(1 - \frac{t^2 \sigma_{n,k}^2}{2}\right) \rightarrow 0, n \rightarrow \infty.$$

Da $-\sum_{k=1}^{r_n} \frac{t^2 \sigma_{n,k}^2}{2} \rightarrow -\frac{\sigma^2 t^2}{2}$ mit (5.6a), folgt die Behauptung. ■

5.2 Statistische Anwendung: Asymptotik für Momentenschätzer

Wir beziehen uns auf die Momentenmethode aus Kapitel 4.2.2 für einen eindimensionalen, reellen Parameter $\theta \in \mathbb{R}$. Dieser erfüllt die Relation $h_j(\theta) = \mu_j = \mathbb{E}_\theta[X_1^j]$, und wir schätzen aus u.i.v. Zufallsvariablen X_1, \dots, X_n den Parameter über den Momentenschätzer

$$\hat{\theta}_n = h_j^{-1}(\hat{\mu}_j), \text{ mit } \hat{\mu}_j = n^{-1} \sum_{i=1}^n X_i^j,$$

unter der Annahme, dass h_j^{-1} und μ_j existieren. Wir schreiben dabei wieder $\mathbb{E}_\theta[\cdot]$, für den Erwartungswert bezüglich \mathbb{P}_θ , und analog für andere Momente. In der Regel betrachten wir $j = 1$, aber falls höhere Momente existieren könnten wir auch diese benutzen.

Proposition 5.9. Existiert μ_j , und für hinreichend großes n der Momentenschätzer $\hat{\theta}_n = h_j^{-1}(\hat{\mu}_j)$ und ist h_j stetig, so ist $\hat{\theta}_n$ (stark) konsistent, d.h. $\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta$ fast sicher.

Beweis. Nach dem starken Gesetz der großen Zahlen gilt

$$\hat{\mu}_j \rightarrow \mu_j, \text{ f.s.}$$

Die fast sichere Konvergenz ist bezüglich $\mathbb{P}_\theta^{\otimes \mathbb{N}}$, oder bezüglich einem Wahrscheinlichkeitsmaß auf einem Wahrscheinlichkeitsraum auf dem alle Zufallsvariablen (X_i) definiert werden. Ist h_j injektiv und stetig, so ist auch h_j^{-1} stetig, und die fast sichere Konvergenz überträgt sich:

$$\lim_{n \rightarrow \infty} h_j^{-1}(\hat{\mu}_j) = h_j^{-1}(\mu_j) = \theta, \text{ f.s.} \quad \blacksquare$$

Proposition 5.10. Es sei $\theta_0 \in \Theta$, und für hinreichend großes n existiere der Momentenschätzer $\hat{\theta}_n = h_j^{-1}(\hat{\mu}_j)$. Es seien $X_i^j \in L_2(\mathbb{P}_{\theta_0})$, also $\mathbb{E}_{\theta_0}[X_1^{2j}] < \infty$. Sofern h_j in einer Umgebung von θ_0 stetig differenzierbar ist, ist $\hat{\theta}_n$ unter $\mathbb{P}_{\theta_0}^{\otimes \mathbb{N}}$ asymptotisch normalverteilt und erfüllt den ZGWS

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \text{Var}_{\theta_0}(X_1^j)(h'_j(\theta_0))^{-2}\right).$$

Beweis. Nach dem ZGWS nach Lindeberg-Lévy, Satz 5.2, gilt unter den Voraussetzungen der Proposition unter $\mathbb{P}_{\theta_0}^{\otimes \mathbb{N}}$:

$$\sqrt{n}\left(\frac{1}{n} \sum_{i=1}^n X_i^j - \mu_j\right) \xrightarrow{d} \mathcal{N}\left(0, \text{Var}_{\theta_0}(X_1^j)\right),$$

wobei $\text{Var}_{\theta_0}(X_1^j) = \mathbb{E}_{\theta_0}[X_1^{2j}] - (\mathbb{E}_{\theta_0}[X_1^j])^2$. Im Falle der Existenz der Umkehrfunktion, ist diese differenzierbar in $h_j(\theta_0)$, falls h_j in einer Umgebung von θ_0 differenzierbar ist. Die Behauptung folgt daher unmittelbar mit der Delta-Methode. Unter \mathbb{P}_{θ_0} , ist $\mu_j = \mathbb{E}_{\theta_0}[X_1^j]$, und daher $h_j^{-1}(\mu_j) = h_j^{-1}(h_j(\theta_0)) = \theta_0$. Für die Ableitung der Umkehrfunktion gilt

$$(h_j^{-1})'(\mathbb{E}_{\theta_0}[X_1^j]) = \frac{1}{h'_j(\theta_0)},$$

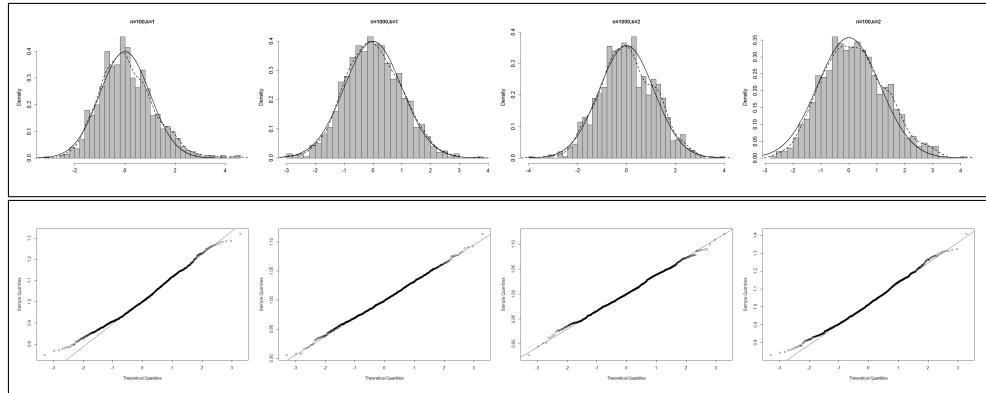


Abbildung 5.1: Monte-Carlo-Verteilungen (aus 1000 Iterationen) der Momentenschätzer für $k = 1, 2$, für Stichprobenumfänge $n = 100$ bzw. $n = 1000$. Histogramme mit theoretischer asymptotischer Dichte und Kern-Dichte-Schätzer mit Silverman's Bandweitenregel (gestrichelt) und QQ-Plots der empirischen Verteilungen gegen die Normalverteilung.

vorausgesetzt dass der Nenner nicht Null ist, womit man die asymptotische Varianz im ZGWS erhält. ■

Beim Erwartungswert und der Varianz der Grenzverteilung spricht man auch vom **asymptotischen Erwartungswert** und **asymptotischer Varianz**. Die Begriffe sind jedoch nicht ganz klar, da nicht notwendigerweise Konvergenz der Momente gilt. Dafür wird noch gleichgradige Integrierbarkeit benötigt.

Beispiel 5.11. Im Beispiel zur Exponentialverteilung aus Kapitel 4.2.2, gilt

$$(h_k^{-1})'(x) = -(k!/x)^{1/k} (kx)^{-1}, \quad h'_k(\lambda) = -kk! \lambda^{-k-1},$$

und

$$\text{Var}_{\lambda_0}(X_i^k) = ((2k)! - (k!)^2)/\lambda_0^{2k}.$$

Alle Momentenschätzer $\hat{\lambda}_{k,n}$, für alle $k \in \mathbb{N}$, sind daher asymptotisch normalverteilt mit Rate $n^{-1/2}$ und asymptotischer Varianz $\lambda_0^{2k-2}((2k)!/(k!)^2 - 1)$. Da $\hat{\lambda}_{1,n}$ die gleichmäßig kleinste asymptotische Varianz besitzt, wird dieser Schätzer im Allgemeinen vorgezogen. Ergebnisse einer Monte-Carlo-Simulation der Schätzer für $k = 1, 2$, und zwei unterschiedliche Stichprobengrößen sind in Abbildung 5.1 veranschaulicht.

Als wichtige **statistische Implikation** ermöglicht der ZGWS **asymptotische Konfidenzintervalle** für den Parameter. Ist die Varianz ebenfalls stetig in θ_0 , so kann die asymptotische Varianz stets durch Plug-in konsistent geschätzt werden. Das Lemma von Slutsky ergibt dann einen ZGWS mit einer Standardnormalverteilung als Grenzverteilung, aus welchem sich asymptotische Konfidenzintervalle ableiten lassen. Alternativ kann eine Varianz-stabilisierende Transformationen benutzt werden.

5.3 Multivariater zentraler Grenzwertsatz

Wir nutzen die charakteristische Funktion der multivariaten Normalverteilung und die Dimensionsreduktion bei charakteristischen Funktionen. Danach beweisen wir den multivariaten zentralen Grenzwertsatz, der aus der eindimensionalen Version und dem Satz von Cramér-Wold folgt.

Satz 5.12 (Multivariater ZGWS). Sind $\mathbf{X}_1, \mathbf{X}_2, \dots$ u.i.v. Zufallsvektoren im \mathbb{R}^d , $\mathbf{X}_n = (X_{n1}, \dots, X_{nd})^T$,

mit $\mathbb{E}[X_{ni}^2] < \infty$, und ist $\mu = \mathbb{E}[\mathbf{X}_1]$ sowie $\Sigma = \text{Cov}(\mathbf{X}_1)$, so gilt für $\mathbf{S}_n = \mathbf{X}_1 + \dots + \mathbf{X}_n$, dass

$$\sqrt{n}\left(\frac{\mathbf{S}_n}{n} - \mu\right) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

■

Beweis. Wir nutzen das Kriterium aus dem Satz 3.38 von Cramér-Wold. Ist $\theta \in \mathbb{R}^d$, so sind $\theta^T \mathbf{X}_1, \theta^T \mathbf{X}_2, \dots$ u.i.v. Zufallsvariablen mit Erwartungswert $\theta^T \mu$ und Varianz $\theta^T \Sigma \theta$. Daher folgt nach dem ZGWS, Satz 5.2, und mit Satz 2.19, dass

$$\sqrt{n}\theta^T\left(\frac{\mathbf{S}_n}{n} - \mu\right) = \sqrt{n}\left(\frac{\theta^T \mathbf{X}_1 + \dots + \theta^T \mathbf{X}_n}{n} - \theta^T \mu\right) \xrightarrow{d} \mathcal{N}(0, \theta^T \Sigma \theta). \quad \blacksquare$$

Beispiel 5.13. Seien $\mathbf{X}_n \sim \text{Mult}(1; p_1, \dots, p_d)$ unabhängig und multinomialverteilt, also $p_j \geq 0$, $p_1 + \dots + p_d = 1$, und $\mathbb{P}(\mathbf{X}_1 = e_j) = p_j$, wobei e_j der j -te Einheitsvektor im \mathbb{R}^d ist. Dann ist $\mathbb{E}[\mathbf{X}_1] = (p_1, \dots, p_d)^T$, und $\text{Cov}(\mathbf{X}_1)_{j,k} = -p_j p_k$, $j \neq k$, sowie $\text{Cov}(\mathbf{X}_1)_{j,j} = p_j(1 - p_j)$. Weiter ist

$$\mathbf{S}_n = \mathbf{X}_1 + \dots + \mathbf{X}_n,$$

und nach dem multivariaten ZGWS folgt

$$\sqrt{n}\left(\frac{\mathbf{S}_n}{n} - (p_1, \dots, p_d)^T\right) \xrightarrow{d} \mathcal{N}(0, \text{Cov}(\mathbf{X}_1)).$$

Proposition 5.14 (Multivariate Delta-Methode). *Es seien $(X_n)_{n \in \mathbb{N}}$ eine Folge von Zufallsvektoren im \mathbb{R}^k , $\sigma_n > 0$, $\sigma_n \rightarrow 0$, $\theta_0 \in \mathbb{R}^k$ sowie $\Sigma \in \mathbb{R}^{k \times k}$ positiv semidefinit symmetrisch und es gelte*

$$\sigma_n^{-1}(X_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Ist $f : \mathbb{R}^k \rightarrow \mathbb{R}^m$ in einer Umgebung von θ_0 stetig differenzierbar, so folgt

$$\sigma_n^{-1}(f(X_n) - f(\theta_0)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial f}{\partial \theta}(\theta_0) \Sigma \left(\frac{\partial f}{\partial \theta}(\theta_0)\right)^T\right),$$

wobei $\frac{\partial f}{\partial \theta} \in \mathbb{R}^{m \times k}$, und $\mathcal{N}(0, 0)$ gegebenenfalls als Punktmaß δ_0 in der Null zu verstehen ist.

Beweis. Über das Lemma von Slutsky folgt, dass $X_n - \theta_0 = \sigma_n \frac{X_n - \theta_0}{\sigma_n} \xrightarrow{d} 0$ und somit $X_n \xrightarrow{\mathbb{P}} \theta_0$ für $n \rightarrow \infty$. Eine Taylorentwicklung ergibt

$$f(X_n) = f(\theta_0) + \frac{\partial f}{\partial \theta}(\theta_0)(X_n - \theta_0) + R_n,$$

mit $|R_n| / |X_n - \theta_0| \rightarrow 0$ für $X_n \rightarrow \theta_0$ bezüglich fast sicherer und damit auch stochastischer Konvergenz. Wiederum mit dem Lemma von Slutsky folgt

$$\frac{|R_n|}{\sigma_n} = \frac{|X_n - \theta_0|}{\sigma_n} \frac{|R_n|}{|X_n - \theta_0|} \xrightarrow{d} 0,$$

also auch bezüglich stochastischer Konvergenz. Eine Anwendung von Satz 3.21 ergibt daher, dass

$$\sigma_n^{-1}(f(X_n) - f(\theta_0)) = \sigma_n^{-1} \frac{\partial f}{\partial \theta}(\theta_0)(X_n - \theta_0) + \sigma_n^{-1} R_n \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial f}{\partial \theta}(\theta_0) \Sigma \left(\frac{\partial f}{\partial \theta}(\theta_0)\right)^T\right),$$

denn es gilt $\sigma_n^{-1} \frac{\partial f}{\partial \theta}(\theta_0)(X_n - \theta_0) \xrightarrow{d} \frac{\partial f}{\partial \theta}(\theta_0)Z$, mit $Z \sim \mathcal{N}(0, \Sigma)$, und

$$\frac{\partial f}{\partial \theta}(\theta_0)Z \sim \mathcal{N}\left(0, \frac{\partial f}{\partial \theta}(\theta_0) \Sigma \left(\frac{\partial f}{\partial \theta}(\theta_0)\right)^T\right). \quad \blacksquare$$

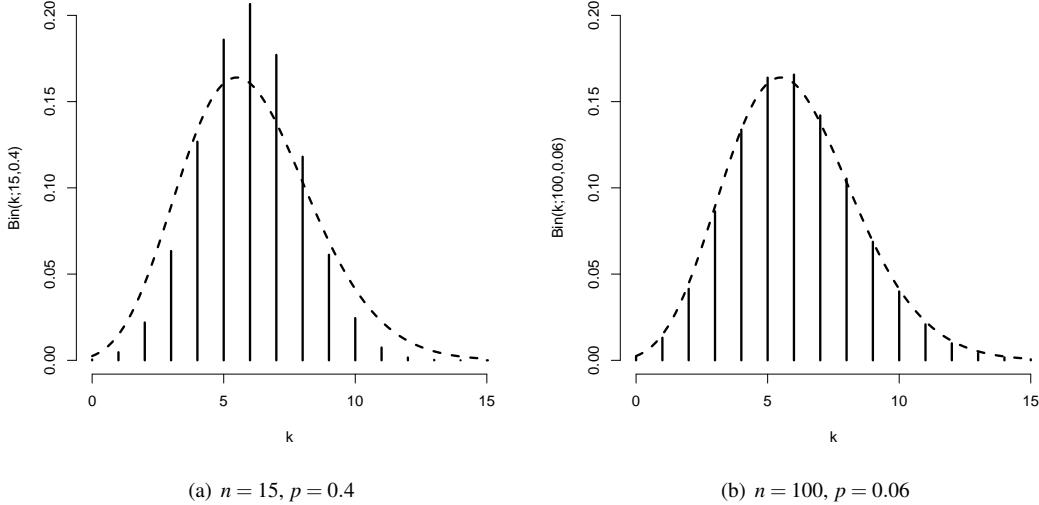


Abbildung 5.2: Wahrscheinlichkeiten $\text{Bin}(k; n, p)$ der Binomialverteilung zusammen mit der Grenzfunktion $x \mapsto e^{-\lambda} \lambda^x / x!$ der Poisson-Approximation (gestrichelt) für $\lambda = 6$.

5.4 Poisson-Konvergenz

Wir zeigen in diesem Kapitel, dass die Summe einer wachsenden Anzahl von Bernoulli-verteilten Zufallsvariablen asymptotisch Poisson-verteilt ist, wenn die Erfolgswahrscheinlichkeiten geeignet gegen Null konvergieren. Diese Tatsache wird auch als das **Gesetz der seltenen Ereignisse** bezeichnet.

Zunächst erinnern wir an die Poisson-Approximation der Binomialverteilung, anschließend arbeiten wir ein allgemeineres Resultat für Dreiecksschemata aus.

Satz 5.15 (Poisson-Approximation der Binomialverteilung). *Sei $(p_n) \in (0, 1)^{\mathbb{N}}$, sodass*

$$\lim_{n \rightarrow \infty} p_n = 0, \quad \lim_{n \rightarrow \infty} (n \cdot p_n) = \lambda \in (0, \infty).$$

Ist $X_n \sim \text{Bin}(n, p_n)$, so gilt die Konvergenz in Verteilung:

$$X_n \xrightarrow{d} \text{Poi}(\lambda).$$

■

Beweis. Für n hinreichend groß ist $n \geq k$, da k fest ist. Also gilt

$$\mathbb{P}(X_n = k) = \binom{n}{k} \cdot p_n^k \cdot (1 - p_n)^{n-k} = \frac{(n)_k}{n^k} \cdot (np_n)^k \cdot \frac{1}{k!} \cdot \left(1 - \frac{np_n}{n}\right)^n \cdot (1 - p_n)^{-k}.$$

Für die einzelnen Faktoren gilt

$$\lim_{n \rightarrow \infty} \frac{(n)_k}{n^k} = 1, \quad \lim_{n \rightarrow \infty} (1 - p_n)^k = (1 - 0)^k = 1, \quad \lim_{n \rightarrow \infty} (np_n)^k = \lambda^k.$$

Weiterhin gilt

$$\lim_{n \rightarrow \infty} x_n = x \Rightarrow \lim_{n \rightarrow \infty} \left(1 - \frac{x_n}{n}\right)^n = e^{-x},$$

also

$$\lim_{n \rightarrow \infty} \left(1 - \frac{np_n}{n}\right)^n = e^{-\lambda}.$$

Es folgt also

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n = k) = 1 \cdot \lambda^k \cdot \frac{1}{k!} \cdot e^{-\lambda} \cdot 1 = e^{-\lambda} \cdot \frac{\lambda^k}{k!}.$$

Die Konvergenz der Zähldichten impliziert punktweise Konvergenz der Verteilungsfunktionen und damit die Aussage. \blacksquare

Satz 5.16. Sei $(X_{n,k})_{n \geq 1, 1 \leq k \leq r_n}$ ein Dreiecksschema mit folgenden Eigenschaften:

1. Für alle n sind $X_{n,1}, \dots, X_{n,r_n}$ unabhängig.
2. $\mathbb{P}(X_{n,k} = 1) = p_{n,k} = 1 - \mathbb{P}(X_{n,k} = 0)$.
3. $\sum_{k=1}^{r_n} p_{n,k} \rightarrow \lambda \in (0, \infty)$, $n \rightarrow \infty$.
4. $\max_{k=1, \dots, r_n} p_{n,k} \rightarrow 0$, $n \rightarrow \infty$.

Dann gilt

$$S_n = X_{n,1} + \dots + X_{n,r_n} \xrightarrow{d} Poi(\lambda).$$

\blacksquare

Bemerkung. Wir betrachten die Situation des Satzes.

a. $\mathbb{E}[X_{n,k}] = p_{n,k}$, $\text{Var}(X_{n,k}) = p_{n,k}(1 - p_{n,k})$, sowie

$$\text{Var}(S_n) = p_{n,1}(1 - p_{n,1}) + \dots + p_{n,r_n}(1 - p_{n,r_n}).$$

Daher folgt

$$\mathbb{E}[S_n] \geq \text{Var}(S_n) \geq \min_{k=1, \dots, r_n} (1 - p_{n,k}) \sum_{j=1}^{r_n} p_{n,j} \rightarrow \lambda,$$

also $\text{Var}(S_n) \rightarrow \lambda$. Insbesondere folgt

$$\frac{1}{\text{Var}(S_n)} \max_{k=1, \dots, r_n} \text{Var}(X_{n,k}) \rightarrow 0,$$

also die Feller-Bedingung.

b. Da nach Annahme mit $\varepsilon < 1/(2\lambda^{1/2})$ für große n gilt dass $\varepsilon(\text{Var}(S_n))^{1/2} < 1/2$ und $p_{n,k} < 1/2$ für $1 \leq k \leq r_n$, ist dann

$$\{|X_{n,k} - p_{n,k}| \geq \varepsilon(\text{Var}(S_n))^{1/2}\} \supset \{|X_{n,k} - p_{n,k}| \geq 1/2\} = \{X_{n,k} = 1\}.$$

Daraus folgert man, dass die Lindeberg-Bedingung *nicht* erfüllt ist.

Beweis. [von Satz 5.16] Es ist $\varphi_{X_{n,k}}(t) = 1 - p_{n,k} + p_{n,k}e^{it} = 1 + p_{n,k}(e^{it} - 1)$, daher ist

$$\varphi_{S_n}(t) = \prod_{k=1}^{r_n} (1 + p_{n,k}(e^{it} - 1)).$$

Wir zeigen, dass

$$\varphi_{S_n}(t) \rightarrow \exp(\lambda(e^{it} - 1)), \quad n \rightarrow \infty. \quad (5.8)$$

Es ist

$$\begin{aligned} |\varphi_{S_n}(t) - \exp(\lambda(e^{it} - 1))| &\leq \left| \prod_{k=1}^{r_n} (1 + p_{n,k}(e^{it} - 1)) - \prod_{k=1}^{r_n} \exp(p_{n,k}(e^{it} - 1)) \right| \\ &\quad + \left| \prod_{k=1}^{r_n} \exp(p_{n,k}(e^{it} - 1)) - \exp(\lambda(e^{it} - 1)) \right|. \end{aligned}$$

Der zweite Summand konvergiert gegen 0 nach Annahme 3. des Satzes. Für den ersten Summand gilt $|1 + p_{n,k}(e^{it} - 1)| \leq |1 - p_{n,k}| + p_{n,k}|e^{it}| \leq 1$. Da $|\exp(z)| = \exp(\Re(z))$, $z \in \mathbb{C}$, ist

$$|\exp(p_{n,k}(e^{it} - 1))| = \exp(p_{n,k}(\cos(t) - 1)) \leq 1.$$

Daher folgt mit Lemma 5.3:

$$\begin{aligned} & \left| \prod_{k=1}^{r_n} (1 + p_{n,k}(e^{it} - 1)) - \prod_{k=1}^{r_n} \exp(p_{n,k}(e^{it} - 1)) \right| \\ & \leq \sum_{k=1}^{r_n} \left| \exp(p_{n,k}(e^{it} - 1)) - (1 + p_{n,k}(e^{it} - 1)) \right|. \end{aligned}$$

Da für $z \in \mathbb{C}$, $|z| \leq 1$, gilt

$$|\exp(z) - 1 - z| \leq \sum_{n=2}^{\infty} \frac{|z|^n}{n!} \leq |z|^2 \sum_{n=2}^{\infty} \frac{1}{n!} = C|z|^2,$$

und da für $p_{n,k} \leq 1/2$ gilt $|p_{n,k}(e^{it} - 1)| \leq 1$, folgt für $n \rightarrow \infty$, dass

$$\begin{aligned} & \left| \prod_{k=1}^{r_n} (1 + p_{n,k}(e^{it} - 1)) - \prod_{k=1}^{r_n} \exp(p_{n,k}(e^{it} - 1)) \right| \\ & \leq C \sum_{k=1}^{r_n} |p_{n,k}(e^{it} - 1)|^2 \leq 4C \max_{1 \leq k \leq r_n} p_{n,k} \sum_{j=1}^{r_n} p_{n,j} \rightarrow 0. \end{aligned}$$

■

Korollar 5.17. Sei $(X_{n,k})_{n \geq 1, 1 \leq k \leq r_n}$ ein Dreiecksschema von Zufallsvariablen mit Werten in \mathbb{N}_0 . Setze $p_{n,k} = \mathbb{P}(X_{n,k} = 1)$ und $q_{n,k} = \mathbb{P}(X_{n,k} \geq 2)$. Angenommen,

1. für alle n sind $X_{n,1}, \dots, X_{n,r_n}$ unabhängig,
2. $\sum_{k=1}^{r_n} p_{n,k} \rightarrow \lambda \in (0, \infty)$, $n \rightarrow \infty$,
3. $\max_{k=1, \dots, r_n} p_{n,k} \rightarrow 0$, $n \rightarrow \infty$,
4. $\sum_{k=1}^{r_n} q_{n,k} \rightarrow 0$, $n \rightarrow \infty$.

Dann gilt

$$S_n = X_{n,1} + \dots + X_{n,r_n} \xrightarrow{d} \text{Poi}(\lambda).$$

Beweis. Setze $X'_{n,k} = X_{n,k} 1_{\{X_{n,k} \leq 1\}}$. Die $(X'_{n,k})$ erfüllen die Bedingungen des Satzes, also gilt $S'_n = X'_{n,1} + \dots + X'_{n,r_n} \xrightarrow{d} \text{Poi}(\lambda)$. Weiter ist

$$\mathbb{P}(S_n \neq S'_n) = \mathbb{P}\left(\bigcup_{k=1}^{r_n} \{X_{n,k} \geq 2\}\right) \leq \sum_{k=1}^{r_n} q_{n,k} \rightarrow 0, \quad n \rightarrow \infty,$$

insbesondere $S_n - S'_n \xrightarrow{\mathbb{P}} 0$, und die Behauptung folgt mit Satz 3.21. ■

Literaturverzeichnis

- Bauer, H. (2002). *Wahrscheinlichkeitstheorie, 5th edition.* de Gruyter.
- Billingsley, P. (1994). *Probability and Measure.* Wiley.
- Breiman, L. (2007). *Probability.* Siam.
- Durrett, R. (2010). *Probability. Theory and examples. 4th ed.,* Cambridge: Cambridge University Press.
- Georgii, H.-O. (2007). *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik.* de Gruyter.
- Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Volume 1.* Wiley.
- Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Volume 2.* Wiley.
- Jacod, J. and Protter, P. (2000). *Probability Essentials.* Springer.
- Klenke, A. (2008). *Wahrscheinlichkeitstheorie.* Berlin: Springer.
- Krengel, U. (2002). *Einführung in die Wahrscheinlichkeitstheorie und Statistik. 6th Edition.* Braunschweig: Vieweg.
- Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer.
- Werner, D. (2011). *Funktionalanalysis.* Berlin: Springer.
- Van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.