Stochastik - 2. Teil

Andrej Depperschmidt

Vorlesungsskript

Universität Freiburg

Sommersemester 2016

Version: 13. Juli 2016

1 Einleitung und Wiederholung

In diesem Kapitel wiederholen wir kurz einige grundlegende Begriffe und Resultate aus dem 1. Teil der Stochastikvorlesung. Das Kapitel wird im Laufe der Vorlesung nach und nach ergänzt, wenn wir auf Resultate verweisen wollen, die schon bekannt sind.

Im Folgenden ist Ω stets eine nichtleere Menge, genannt *Grundraum*. Diese Menge enthält alle möglichen Ausgänge der betrachteten Zufallsexperimente.

1.1 Mengenoperationen und Notation

In diesem Abschnitt wiederholen wir einige Begriffe und Notation aus der Mengenlehre. Im Folgenden ist Ω stets eine nichtleere Menge.

Die *Potenzmenge von* Ω ist als die Menge aller Teilmengen von Ω definiert und wird mit $\mathcal{P}(\Omega)$ bezeichnet. Also ist

$$\mathcal{P}(\Omega) := \{A : A \subset \Omega\}. \tag{1.1}$$

Mit $A \cap B$, $A \cup B$ bezeichnen wir wie üblich den Durchschnitt bzw. die Vereinigung der Mengen A und B. Für $A \subset \Omega$ bezeichnen wir mit $A^{\mathsf{c}} := \Omega \setminus A = \{\omega \in \Omega : \omega \notin A\}$ das Komplement von A in Ω. Die symmetrische Differenz der Mengen A und B ist definiert durch

$$A \triangle B := (A \setminus B) \cup (B \setminus A). \tag{1.2}$$

Oft ist es nützlich Vereinigungen von Mengen als Vereinigungen von disjunkten Mengen darzustellen. Wie das geht, zeigt das folgende Resultat.

Proposition 1.1 (Disjunkte Vereinigungen)**.** *Es sei $A_1, A_2, \ldots$ eine Folge von Teilmengen von Ω und $A = \bigcup_n A_n$. Dann sind die Mengen*

$$B_1 = A_1, B_2 = A_2 \setminus B_1, \ldots, B_n = A_n \setminus \bigcup_{k=1}^{n-1} B_k, \ldots$$

paarweise disjunkt und es gilt $A = \bigcup_n B_n$.

Beweis. Übung! □

Auch sehr nützlich (insbesondere beim Rechnen mit Wahrscheinlichkeiten) sind die *De Morganschen Regeln*: Für $(A_i)_{i\in I}$, $A_i \subset \Omega$ mit einer beliebigen Indexmenge I (abzählbar[1] oder überabzählbar) gilt

$$\Big(\bigcup_{i\in I} A_i\Big)^{\mathsf{c}} = \bigcap_{i\in I} A_i^{\mathsf{c}} \quad \text{und} \quad \Big(\bigcap_{i\in I} A_i\Big)^{\mathsf{c}} = \bigcup_{i\in I} A_i^{\mathsf{c}}. \tag{1.3}$$

1.2 Wahrscheinlichkeitsraum

Definition 1.2. Das Tripple $(\Omega, \mathcal{A}, \mathrm{P})$ heißt *Wahrscheinlichkeitsraum*, falls gilt:

1. $\mathcal{A}$ ist eine σ-Algebra, d.h. $\mathcal{A} \subset \mathcal{P}(\Omega)$ (= Potenzmenge von Ω) mit
   (i) $\Omega \in \mathcal{A}$,
   (ii) $A \in \mathcal{A} \;\Rightarrow\; A^{\mathsf{c}} \in \mathcal{A}$,
   (iii) $A_1, A_2, \ldots \in \mathcal{A} \;\Rightarrow\; \cup_i A_i \in \mathcal{A}$.
2. P ist ein *Wahrscheinlichkeitsmaß*, d.h. $\mathrm{P} : \mathcal{A} \to [0,1]$ mit
   (i) $\mathrm{P}(\Omega) = 1$,
   (ii) $A_1, A_2, \cdots \in \mathcal{A}$ paarweise disjunkt, dann $\mathrm{P}(\cup_i A_i) = \sum_i \mathrm{P}(A_i)$.

Die Nichtnegativität des Wahrscheinlichkeitsmaßes zusammen mit der Normiertheit (2.(i)) und mit der σ-Additivität (2.(ii)) sind die *Kolmogorovschen Axiome* für Wahrscheinlichkeitsmaße.

Das Wahrscheinlichkeitsmaß P ist eine Mengenfunktion, die jeder Menge $A \in \mathcal{A}$ eine Wahrscheinlichkeit, also einen Wert $\mathrm{P}(A) \in [0,1]$ zuordnet. Typischerweise reicht es P auf einer Teilmenge von $\mathcal{A}$ zu kennen um es vollständig zu beschreiben. Unter anderem dadurch sind Verteilungsfunktionen so wichtig.

Elemente von $\mathcal{A}$ heißen *Ereignisse*. Es gibt viele anderen Möglichkeiten (statt 1.(i)-(iii)) σ-Algebren zu definieren. Einige davon kann man sich sehr leicht mit Hilfe von elementaren Mengenoperationen überlegen.

Aus der Definition von Wahrschenlichkeitsräumen und insbesondere aus den Kolmogorovschen Axiomen ergeben sich leicht einige Rechenregeln für Wahrscheinlichkeitsmaße.

Satz 1.3. *Es gilt*

(i) $\mathrm{P}(\emptyset) = 0$,

(ii) $\mathrm{P}(A_1 \cup A_2) = \mathrm{P}(A_1) + \mathrm{P}(A_2) - \mathrm{P}(A_1 \cap A_2)$,

(iii) $\mathrm{P}(A_1 \cup A_2) \le \mathrm{P}(A_1) + \mathrm{P}(A_2)$,

(iv) $\mathrm{P}(A_1 \cup A_2) = \mathrm{P}(A_1) + \mathrm{P}(A_2)$ *wenn* A_1 *und* A_2 *disjunkt sind,*

[1] Ohne eine besondere Hervorhebung meinen wir mit "abzählbar" immer "endlich" oder "abzählbar unendlich"

(v) $\mathrm{P}(A^{\mathsf{c}}) = 1 - \mathrm{P}(A)$,

(vi) $\mathrm{P}(A_1) \leq \mathrm{P}(A_2)$ *falls* $A_1 \subset A_2$.

Beweis. Stochastik, Teil 1, bzw. Übung. □

Das nächste Resultat ist eine Verallgemeinerung von (ii) im obigen Satz.

Satz 1.4 (Einschluss-Ausschluss Formel)**.** *Für* $A_1, \ldots, A_n \in \mathcal{A}$ *gilt (mit* $[n] = \{1, \ldots, n\}$*)*

$$\mathrm{P}(\cup_{i=1}^{n} A_i) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{\{i_1,\ldots,i_k\} \subset [n]} \mathrm{P}(A_{i_1} \cap \cdots \cap A_{i_k}).$$

Beweis. Stochastik, Teil 1, bzw. Übung. □

1.3 Laplace Modelle

Zufallsexperimente mit endlich vielen möglichen Ausgängen, die alle dieselbe Wahrscheinlichkeit haben werden als *Laplace'sche Zufallsexperimente* bezeichnet. Die einfachsten Beispiele sind fairer Münzwurf und faires Würfeln. Man spricht dann von Gleichverteilung auf der Grundmenge.

Sei Ω endlich und sei $\mathcal{A} = \mathcal{P}(\Omega)$. Insbesondere enthält $\mathcal{A}$ alle *Elementarereignisse* die $\omega \in \Omega$ (wir identifizieren hier ω mit der Menge $\{\omega\}$).

Da alle Ausgänge dieselbe Wahrscheinlichkeit haben, muss natürlich

$$\mathrm{P}(\omega) = \frac{1}{|\Omega|} \quad \text{für alle } \omega \in \Omega,$$

gelten, wobei $|A|$ die Anzahl der Elemente der Menge A ist. Dadurch ist P natürlich vollständig charakterisiert, denn für alle für alle $A \in \mathcal{A}$ ist dann notwendigerweise

$$\mathrm{P}(A) = \frac{|A|}{|\Omega|}.$$

Die Wahrscheinlichkeit von A ist also ein Quotient aus der Anzahl günstiger Fälle (Elemente in A) und der Anzahl möglicher Fälle (Elemente in Ω). *Kombinatorik* ist ein wichtiges Hilfsmittel zur Berechnung von Wahrscheinlichkeiten in Laplace Modellen.

1.4 Allgemeine diskrete Wahrscheinlichkeitsräume

Sei Ω abzählbar und sei $\mathcal{A} = \mathcal{P}(\Omega)$. Durch die Angabe der Wahrscheinlichkeitsgewichte $\mathrm{P}(\omega) := \mathrm{P}(\{\omega\}) \in [0,1]$ für alle *Elementarereignisse* $\omega \in \Omega$ ist P vollständig beschrieben, denn es ist

$$\mathrm{P}(A) = \sum_{\omega \in A} \mathrm{P}(\omega), \quad A \in \mathcal{A}.$$

Es muss natürlich $\mathrm{P}(\Omega) = 1$ gelten!

1.5 Bedingte Wahrscheinlichkeiten

Definition 1.5. Seien $A, B \in \mathcal{A}$ und $\mathrm{P}(B) > 0$. Dann heißt

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

bedingte Wahrscheinlichkeit von A gegeben B.

Es ist leicht zu sehen, dass $\mathrm{P}(\,\cdot \mid B)$ ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{A})$ ist mit $\mathrm{P}(A \mid B) = 0$ für alle $A \in \mathcal{A}$ mit $A \cap B = \emptyset$.

Gilt $\mathrm{P}(A), \mathrm{P}(B) > 0$ dann folgt („einfache *Bayes-Formel*“)

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(B \cap A)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \frac{\mathrm{P}(B)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A \mid B)\,\mathrm{P}(B)}{\mathrm{P}(A)}.$$

Satz 1.6 (Satz von der totalen Wahrscheinlichkeit)**.** *Sei $(B_i)_{i \in I}$ eine disjunkte Zerlegung von Ω. Dann gilt für alle $A \in \mathcal{A}$*

$$\mathrm{P}(A) = \sum_{i \in I : \mathrm{P}(B_i) > 0} \mathrm{P}(A \mid B_i)\,\mathrm{P}(B_i).$$

Beweis. Stochastik, Teil 1. □

Satz 1.7 (Satz von Bayes)**.** *Sei $(B_i)_{i \in I}$ eine disjunkte Zerlegung von Ω und $A \in \mathcal{A}$ mit $\mathrm{P}(A) > 0$, dann gilt*

$$\mathrm{P}(B_i \mid A) = \frac{\mathrm{P}(B_i)\,\mathrm{P}(A \mid B_i)}{\sum_{j \in I} \mathrm{P}(A \mid B_j)\,\mathrm{P}(B_j)}.$$

Dabei setzen wir hier $\mathrm{P}(A \mid B_j) = 0$ (beliebiger anderer Wert wäre auch möglich) für B_j mit $\mathrm{P}(B_j) = 0$.

Beweis. Stochastik, Teil 1. □

Übung 1.8. *Beweisen oder widerlegen Sie die Gleichungen*

(a) $\mathrm{P}(B|A) + \mathrm{P}(B|A^c) = 1$ (b) $\mathrm{P}(B|A) + \mathrm{P}(B^c|A) = 1$ (c) $\mathrm{P}(B|A) + \mathrm{P}(B^c|A^c) = 1$.

Dabei haben die Ereignisse, auf die bedingt wird, jeweils positive Wahrscheinlichkeit.

1.6 Unabhängigkeit

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum.

Definition 1.9. Eine Familie von Ereignissen $(A_i)_{i\in I}$ heißt *unabhängig* wenn für *alle* endlichen $J \subset I$ gilt

$$P(\cap_{i\in J} A_i) = \prod_{i\in J} P(A_i).$$

Das „alle endlichen“ in der Definition oben wichtig! Man kann Beispiele mit $\{A_1, A_2, A_3\}$ konstruieren mit

$$\textit{paarweise Unabhängigkeit} \quad \begin{matrix} \not\Leftarrow \\ \not\Rightarrow \end{matrix} \quad \mathrm{P}(\cap_{i=1}^{3} A_i) = \prod_{i=1}^{3} \mathrm{P}(A_i).$$

Sind A und B unabhängig und gilt $P(A), P(B) > 0$, dann folgt

$$\mathrm{P}(A \mid B) = \mathrm{P}(A) \quad \text{und} \quad \mathrm{P}(B \mid A) = \mathrm{P}(B).$$

Übung 1.10. *(i) Das Ereignis A sei unabhängig von sich selbst. Zeigen Sie, dass dann $\mathrm{P}(A) \in \{0, 1\}$ gilt.*

(ii) Die Ereignisse A und B seien unabhängig und es gelte $A \subset B$. Zeigen Sie, dass dann $\mathrm{P}(A) = 0$ oder $\mathrm{P}(B) = 1$ gelten muss.

(iii) Es sei ein Ereignis A mit $\mathrm{P}(A) \in \{0, 1\}$ gegeben. Zeigen Sie, dass A und ein beliebiges Ereignis B unabhängig sind.

1.7 Zufallsvariablen

Definition 1.11 (Zufallsvariablen und Zufallsvektoren)**.** Eine Abbildung $X : \Omega \to \mathbb{R}$ mit

$$X^{-1}([a, b]) = \{\omega \in \Omega \colon X(\omega) \in [a, b]\} \in \mathcal{A}, \quad \text{für alle } a, b \in \mathbb{R}$$

heißt *Zufallsvariable.*

Ist $X = (X_1, \ldots, X_n)$ wobei $X_1, \ldots, X_n$ Zufallsvariablen sind, dann nennen wir X *Zufallsvektor.*

Bemerkung 1.12.

- Statt $[a, b]$ hätte man in Definition der Zufallsvariablen auch $(a, b]$, $(-\infty, a]$, alle offenen Mengen, alle abgeschlossenen Mengen und viele anderen Mengenklassen fordern können.
- Zufallsvariablen sind *messbare Abbildungen*, hier messbar bezüglich der Borel-σ-Algebra auf $\mathbb{R}$, Bezeichnung $\mathcal{B}(\mathbb{R})$, oder einfach $\mathcal{B}$. Jedes Urbild einer Borel-Menge ist ein Element der σ-Algebra $\mathcal{A}$. Mengen der Form $X^{-1}(B)$, $B \in \mathcal{B}$ kann auf $(\Omega, \mathcal{A}, \mathrm{P})$ also die Wahrscheinlichkeit $\mathrm{P}(X^{-1}(B)) = \mathrm{P}(X \in B)$ zugeordnet werden.

Definition 1.13. Ist X eine Zufallsvariable, so heißt $F : \mathbb{R} \to [0,1]$ definiert durch

$$F(x) \coloneqq \mathrm{P}(X \leq x), \quad x \in \mathbb{R}$$

die *Verteilungsfunktion* von X. Das auf $\mathcal{B}$ definierte Wahrscheinlichkeitsmaß

$$\mathrm{P}_X(B) \coloneqq \mathrm{P}(X \in B), \quad B \in \mathcal{B}$$

heißt *Verteilung* von X. Ist $\mu = \mathrm{P}_X$ so schreiben wir $X \sim \mu$ und sagen, dass X nach μ verteilt ist.

Ist $X = (X_1, \ldots, X_n)$ ein Zufallsvektor, dann heißt

$$\mathrm{P}(X_1 \in A_1, \ldots, X_n \in A_n)$$

die *gemeinsame Verteilung* von $X_1, \ldots, X_n$. Die Verteilung P_{X_i} der einzelnen Komponenten X_i heißt *Randverteilung*.

Bemerkung 1.14.

(i) Verteilungsfunktionen sind monoton, rechtststetig, und es gilt

$$\lim_{x \to -\infty} F(x) = 0, \quad \text{und} \quad \lim_{x \to \infty} F(x) = 1.$$

(ii) Verteilungsfunktion von X bestimmt die Verteilung von X eindeutig! ($\to$ Maßtheorie). Das Wahrscheinlichkeitsmaß P_X ist durch seine Werte auf den Mengen der Form $(-\infty, a]$, $a \in \mathbb{R}$, eindeutig festgelegt.

(iii) Ist die gemeinsame Verteilung von $X_1, \ldots, X_n$ bekannt, so bekommt die Randverteilungen durch

$$\mathrm{P}(X_i \in B) = \mathrm{P}(X_i \in B, X_k \in \mathbb{R}, k \neq i) \quad i \in \{1, \ldots, n\}.$$

Durch die gemeinsame Verteilung sind also die Randverteilung festgelegt. Umgekehrt ist es nicht der Fall. Ohne weitere einschränkende Voraussetzungen (wie z.B. Unabhängigkeit) bestimmen die Randverteilungen nicht die gemeinsame Verteilung.

Beispiel 1.15. (Einige wichtige diskrete Verteilungen)

(i) Ist $\mathrm{P}(X = c) = 1$ für ein $c \in \mathbb{R}$, dann ist $\mathrm{P}_X = \delta_c$ das Dirac-Maß in c.

(ii) Ist $X : \Omega \to \{0,1\}$, $p \in [0,1]$ und $\mathrm{P}(X = 1) = p$, $\mathrm{P}(X = 0) = 1 - p$, dann heißt P_X *Bernoulli-Verteilung mit Parameter* p und wird mit Ber_p bezeichnet. Formal ist

$$\mathrm{Ber}_p = (1-p)\delta_0 + p\delta_1.$$

(iii) Seien $n \in \mathbb{N}$ und $p \in [0,1]$. Ist $X : \Omega \to \{0, \ldots, n\}$ mit

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

dann ist $\mathrm{P}_X =: \mathrm{Bin}_{n,p}$ die *Binomialverteilung mit Parametern* n *und* p. Die Zufallsvariable X kann man als Anzahl von Erfolgen bei n unabhängigen Bernoulli Experimenten interpretieren.

(iv) Sei $p \in (0,1]$. Ist $X : \Omega \to \mathbb{N}_0$ mit

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

dann ist $\mathrm{P}_X =: \mathrm{Geo}_p$ die *geometrische Verteilung mit Parameter* p. Die Zufallsvariable X kann man als die Wartezeit bis zum ersten „Erfolg“ bei unabhängigen Bernoulli Zufallsexperimenten auffassen.

(v) Sei $\lambda \in [0, \infty)$. Ist $X : \Omega \to \mathbb{N}_0$ mit

$$\mathrm{P}(X = n) = e^{-\lambda}\frac{\lambda^n}{n!}, \quad n \in \mathbb{N}_0,$$

dann heißt $\mathrm{P}_X =: \mathrm{Poi}_\lambda$ die *Poisson-Verteilung* mit Parameter λ.

(vi) Seien $n, N, M \in \mathbb{N}$. Ist $X : \Omega \to \{0, \dots, n\}$ mit

$$\mathrm{P}(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

dann ist $\mathrm{P}_X =: \mathrm{Hyp}_{n;N,M}$ die *hypergeometrische Verteilung mit Parametern* n, N *und* M. Die Zufallsvariable X kann wie folgt interpretieren: In einer Urne seien N Kugeln enthalten wovon M markiert sind. Es werden n Kugeln ohne Zurücklegen gezogen. Dann gibt X die Anzahl der markierten Kugeln, die dabei gezogen werden.

1.8 (Absolut) Stetige Zufallsvariablen

Definition 1.16. Eine Zufallsvariable $X : \Omega \to \mathbb{R}$ heißt *absolut stetig*, falls es eine Funktion $f : \mathbb{R} \to [0, \infty)$, genannt *Dichte* von X, gibt mit

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1 \quad \text{und} \quad F(b) = \mathrm{P}(X \le b) = \int_{-\infty}^{b} f(x)\,\mathrm{d}x, \text{ für alle } b \in \mathbb{R}. \tag{1.4}$$

Eine Dichte bestimmt eindeutig die Verteilungsfunktion und damit auch die Verteilung einer Zufallsvariablen. Für allgemeine Mengen $A \in \mathcal{B}$ ist

$$\mathrm{P}(X \in A) = \int_A f(x)\,\mathrm{d}x.$$

Außerdem gilt

$$\mathrm{P}(X = a) = 0 \quad \text{für alle } a \in \mathbb{R}$$

und

$$\mathrm{P}(a \le X \le b) = \mathrm{P}(a < X < b) = \int_a^b f(x)\,\mathrm{d}x \quad \text{für alle } a \le b,\, a, b \in \mathbb{R}.$$

Beispiel 1.17. (Einige wichtige stetige Verteilungen)

(i) Seien $a, b \in \mathbb{R}$ mit $a < b$. Die *Gleicheverteilung auf* $[a, b]$ ist eine Verteilung auf $\mathbb{R}$ mit Dichte

$$f(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x). \tag{1.5}$$

Wir schreiben $\mathcal{U}_{[a,b]}$ für die Verteilung. Die zugehörige Verteilungsfunktion hat die Form

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y = \begin{cases} 0 & : x < a, \\ \frac{x-a}{b-a} & : a \le x \le b, \\ 1 & : x > b. \end{cases} \tag{1.6}$$

(ii) Sei $\lambda > 0$. Die *Exponentialverteilung mit Parameter* λ ist eine Verteilung auf $\mathbb{R}$ mit Dichte

$$f(x) = \mathbb{1}_{[0,\infty)}(x) \lambda e^{-\lambda x}.$$

Wir bezeichnen die Exponentialverteilung mit Exp_λ. Die zugehörige Verteilungsfunktion hat die Form

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y = \begin{cases} 0 & : x < 0, \\ 1 - e^{-\lambda x} & : x \ge 0. \end{cases} \tag{1.7}$$

(iii) Für $\mu \in \mathbb{R}$ und $\sigma > 0$ besitzt die *Normalverteilung mit Parametern* μ *und* σ die Dichte

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{-\frac{(x-\mu)^2}{2\sigma^2}\Big\}.$$

Wir bezeichnen die Normalverteilung mit $\mathcal{N}_{\mu,\sigma^2}$. Die Verteilungsfunktion kann „nur“ als Integral $F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y$ angegeben werden. Im Fall $\mu = 0$ und $\sigma^2 = 1$ wird sie üblicherweise mit $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\,\mathrm{d}y$ bezeichnet. Werte von Φ kann man in Tabellen nachschlagen (die insbesondere bei „älteren“ Büchern zu Stochastik und Statistik typischerweise im Anhang enthalten waren). In (den meisten) modernen Softwarepaketen ist die Verteilungsfunktion Φ implementiert.

Bemerkung 1.18 (Berechnung von Dichten durch Ableiten von Verteilungsfunktionen)**.** Durch (1.4) ist eine Dichte nicht eindeutig bestimmt. Man kann eine Dichte f z.B. in einem Punkt (und sogar in abzählbar vielen Punkten) verändern ohne das Integral von f und insbesondere die Verteilungsfunktion F zu verändern.

Dort wo die Verteilungsfunktion F differenzierbar ist kann man eine Dichte f durch Ableiten von F bestimmen. In den Punkten wo F nicht differenzierbar ist kann man f beliebig wählen ohne, dass die Beziehung (1.4) gestört wird.

Zum Beispiel ist die Verteilungsfunktion $F(x) = \frac{x-a}{b-a} \cdot \mathbb{1}_{(a,b)}(x) + \mathbb{1}_{[b,\infty)}(x)$ der Gleichverteilung auf $[a, b]$ in den Punkten a und b nicht differenzierbar. In diesen Punkten könnte man die Dichte f aus (1.5) beliebig setzen.

Analog zu eindimensionalen Dichten kann man auch Dichten für Zufallsvektoren definieren. Man muss dann nur mit Mehrfachintegralen rechnen.

Definition 1.19 (Multivariate Dichten)**.** Es sei $X = (X_1, \dots, X_n)$ ein $\mathbb{R}^n$-wertiger Zufallsvektor und sei $f : \mathbb{R}^n \to \mathbb{R}$ eine nichtnegative integrierbare Funktion mit

$$\int \cdots \int_{\mathbb{R}^n} f(a_1, \dots, a_n)\, \mathrm{d}a_1 \dots \mathrm{d}a_n = 1.$$

Gilt für alle $A \in \mathcal{B}(\mathbb{R}^n)$

$$\mathrm{P}(X \in A) = \int \cdots \int_A f(a_1, \dots, a_n)\, \mathrm{d}a_1 \dots \mathrm{d}a_n,$$

dann nennt man f die *Dichte von* X.

Beispiel 1.20 (Multivariate Normalverteilung)**.** Es sei $\mu \in \mathbb{R}^n$ und sei Σ eine positiv definite $n \times n$ Matrix. Dann ist f definiert durch

$$f(t) = \det(2\pi\Sigma)^{-1/2} \exp\Big\{-\frac{1}{2}(t-\mu)^{\mathsf{T}}\Sigma^{-1}(t-\mu)\Big\}, \quad t = (t_1, \dots, t_n) \in \mathbb{R}^n$$

die Dichte der multivariaten Normalverteilung mit Parametern μ und Σ.

1.9 Erwartungswert, Varianz und Kovarianz

Definition 1.21 (Erwartungswert diskreter Zufallsvariablen)**.** Der *Erwartungswert* einer diskreten Zufallsvariablen X mit Werten in S ist definiert durch

$$\mathrm{E}[X] = \sum_{a \in S} a\, \mathrm{P}(X = a),$$

sofern die Summe wohldefiniert ist ($\pm\infty$ sind als Werte zugelassen).

Lemma 1.22 (Transformation von Erwartungswerten)**.** *Sei X eine diskrete Zufallsvariable mit $\mathrm{P}(X \in S) = 1$ und $h : S \to \mathbb{R}$. Dann ist $h(X)$ eine diskrete Zufallsvariable und für den Erwartungswert gilt (sofern wohldefiniert)*

$$\mathrm{E}[h(X)] = \sum_{a \in S} h(a)\, \mathrm{P}(X = a).$$

Beweis. Stochastik, Teil 1. □

Das folgende Resultat behandelt eine alternative Darstellung des Erwartungswertes nichtnegativer ganzzahliger Zufallsvariablen, die oft nützlich ist. Wie Abbildung 1.1 zeigt, kann man den Erwartungswert als Flächeninhalt interpretieren. Definition 1.21 des Erwartungswertes und Satz 1.23 sind zwei Möglichkeiten den Flächeninhalt der grauen Fläche zu berechnen. Entsprechendes Resultat gilt auch für nichtnegative Zufallsvariablen mit Dichten (siehe Satz 1.26) und kann noch weiter verallgemeinert werden (siehe Übung 1.27).

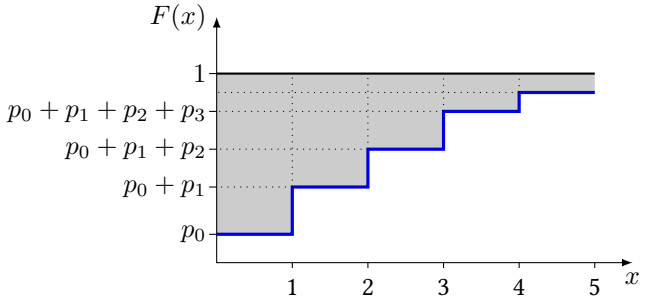

Abbildung 1.1: Erwartungswert nichtnegativer ganzzahliger Zufallsvariablen als Flächeninhalt. Blaue Kurve zeigt die Verteilungsfunktion F.

Satz 1.23. *Ist X eine $\mathbb{N}_0$-wertige Zufallsvariable, so gilt*

$$\mathrm{E}[X] = \sum_{n=0}^{\infty} \mathrm{P}(X > n). \tag{1.8}$$

Beweis. Wir setzen $p_k := \mathrm{P}(X = k)$, $k \in \mathbb{N}_0$. Es gilt

$$\mathrm{E}[X] = \sum_{k=1}^{\infty} k \cdot p_k = \sum_{k=1}^{\infty} \sum_{n=0}^{k-1} p_k = \sum_{n=0}^{\infty} \sum_{k=n+1}^{\infty} p_k = \sum_{n=0}^{\infty} \mathrm{P}(X > n).$$

□

Definition 1.24 (Erwartungswert stetiger Zufallsvariablen)**.** Der *Erwartungswert* einer stetiger Zufallsvariablen X mit Dichte f ist gegeben durch

$$\mathrm{E}[X] = \int_{\mathbb{R}} x f(x)\,\mathrm{d}x,$$

sofern das Integral wohldefiniert ist ($\pm\infty$ sind als Werte zugelassen).

Lemma 1.25 (Transformation von Erwartungswerten)**.** *Sei X eine stetige Zufallsvariable mit $\mathrm{P}(X \in S) = 1$ und $h : S \to \mathbb{R}$. Dann gilt für den Erwartungswert von $h(X)$ (sofern wohldefiniert)*

$$\mathrm{E}[h(X)] = \int_{\mathbb{R}} h(x) f(x)\,\mathrm{d}x.$$

Beweis. Stochastik, Teil 1. □

Das folgende Resultat ist das Analogon von Satz 1.23 für nichtnegative Zufallsvariablen mit Dichten.

Satz 1.26. *Ist X eine nichtnegative Zufallsvariable mit Dichte f und Verteilungsfunktion F, so gilt*

$$\mathrm{E}[X] = \int_0^\infty (1 - F(x))\,\mathrm{d}x. \tag{1.9}$$

Beweis. Es gilt

$$\begin{aligned}\mathrm{E}[X] &= \int_0^\infty x f(x)\,\mathrm{d}x = \int_0^\infty \int_0^x f(x)\,\mathrm{d}y\,\mathrm{d}x = \int_0^\infty \int_y^\infty f(x)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^\infty (1 - F(y))\,\mathrm{d}y.\end{aligned}$$

□

Übung 1.27. *Sei X eine reellwertige Zufallsvariable mit Dichte f und Verteilungsfunktion F, dann gilt*

$$\mathrm{E}[X] = \int_0^\infty (1 - F(x))\,\mathrm{d}x - \int_{-\infty}^0 F(x)\,\mathrm{d}x.$$

Lemma 1.28 (Rechnen mit Erwartungswerten)**.**

(i) Linearität: $\mathrm{E}[aX + bY] = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y]$

(ii) Positivität: Sei $X \geq 0$ dann gilt

(a) $\mathrm{E}[X] \geq 0$,

(b) $\mathrm{E}[X] = 0$ *genau dann, wenn* $\mathrm{P}(X = 0) = 1$.

(iii) Monotonie: Ist $X \leq Y$, dann gilt $\mathrm{E}[X] \leq \mathrm{E}[Y]$.

Beweis. Stochastik, Teil 1, bzw. Übung. □

Bemerkung 1.29. Wenn eine Zufallsvariable X nichtnegativ ist, so ist $\mathrm{E}[X]$ stets definiert, wobei $+\infty$ möglich ist. Nimmt die Zufallsvariable sowohl positive als auch negative Werte an, dann betrachten wir die Zerlegung $X = X^+ - X^-$ in Positivteil $X^+ = \max(X, 0)$ und Negativteil $X^- = -\min(X, 0) = (-X)^+$. Der Erwartungswert von X ist definiert sofern $\min(\mathrm{E}[X^+], \mathrm{E}[X^-]) < \infty$ und es gilt

$$\mathrm{E}[X] = \mathrm{E}[X^+] - \mathrm{E}[X^-].$$

Definition 1.30 (Varianz, Kovarianz, Korrelation)**.** Die *Varianz* einer Zufallsvariablen X ist definiert durch (Wohldefiniertheit vorausgesetzt)

$$\mathrm{Var}[X] \coloneqq \mathrm{E}[(X - \mathrm{E}[X])^2].$$

Die *Kovarianz* von Zufallsvariablen X und Y ist definiert durch

$$\mathrm{Cov}[X, Y] \coloneqq \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])],$$

insbesondere ist $\operatorname{Var}[X] = \operatorname{Cov}[X, X]$.

Sind X und Y Zufallsvariablen mit endlichen positiven und Varianzen dann ist ihr *Korrelationskoeffizient* definiert durch

$$\rho_{X,Y} \coloneqq \frac{\operatorname{Cov}[X, Y]}{\sqrt{\operatorname{Var}[X] \operatorname{Var}[Y]}}.$$

Mit der Linearität des Erwartungswertes ergeben sich folgende Formeln für die Varianz und Kovarianz

$$\operatorname{Var}[X] = \operatorname{E}[X^2] - (\operatorname{E}[X])^2 \quad \text{und} \quad \operatorname{Cov}[X, Y] = \operatorname{E}[XY] - \operatorname{E}[X] \operatorname{E}[Y].$$

Damit lassen sich Varianz und Kovarianz häufig leichter berechnen. Für den Korrelationskoeffizienten gilt stets $-1 \leq \rho_{X,Y} \leq 1$. Das folgt mit der Cauchy-Schwarz Ungleichung; siehe (2.1) und insbesondere (2.2). Ist $\operatorname{Cov}[X, Y] = 0$ und damit auch $\rho_{X,Y} = 0$, so sagt man, dass die Zufallsvariablen X und Y *unkorreliert* sind.

Lemma 1.31 (Eigenschaften der Varianz)**.** *Es gilt*

(i) $\operatorname{Var}[aX + b] = a^2 \operatorname{Var}[X]$,

(ii) $\operatorname{Var}[X] \geq 0$,

(iii) $\operatorname{Var}[X] = 0$ *gilt genau dann, wenn* $\operatorname{P}(X = \operatorname{E}[X]) = 1$,

(iv) $\operatorname{Var}[X_1 + \cdots + X_n] = \sum_{i=1}^n \operatorname{Var}[X_i] + 2 \sum_{i<j} \operatorname{Cov}[X_i, X_j]$,

(v) Sind die Zufallsvariablen $X_1, \ldots, X_n$ paarweise unkorreliert, dann gilt $\operatorname{Var}[X_1 + \cdots + X_n] = \sum_{i=1}^n \operatorname{Var}[X_i]$.

Beweis. Stochastik, Teil 1, bzw. Übung. □

Lemma 1.32 (Eigenschaften der Kovarianz)**.** *Es gilt*

(i) $\operatorname{Cov}[X, Y] = \operatorname{Cov}[Y, X]$,

(ii) $\operatorname{Cov}[a_1 X_1 + a_2 X_2, Y] = a_1 \operatorname{Cov}[X_1, Y] + a_2 \operatorname{Cov}[X_2, Y]$.

Beweis. Stochastik, Teil 1, bzw. Übung. □

1.10 Unabhängige Zufallsvariablen

Definition 1.33. Zufallsvariablen $X_1, \ldots, X_n$ heißen (stochastisch) *unabhängig* falls alle Ereignisse $\{X_i \in A_i\}$, $A_i \in \mathcal{B}$ die Produktform

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \cdots P(X_n \in A_n)$$

haben.

Bemerkung 1.34.

1. Ist die Familie $X_1, \ldots, X_n$ unabhängig, dann ist jede Teilfamilie unabhängig.
2. Ist die Familie $X_1, \ldots, X_n$ unabhängig, dann auch die Familie $h_1(X_1), \ldots, h_n(X_n)$ (für sinnvolle Funktionen $h_1, \ldots, h_n$).

Satz 1.35 (Erwartungswert und Kovarianz unabhängiger Zufallsvariablen)**.** *Seien X_1 und X_2 unabhängige Zufallsvariablen mit Wertebereichen S_1 und S_2 und seien h_1 und h_2 reellwertige Funktionen auf S_1 bzw. S_2. Wenn $h_1(X_1)$ und $h_2(X_2)$ endliche Erwartungswerte haben, dann gilt*

$$\mathrm{E}[h_1(X_1)h_2(X_2)] = \mathrm{E}[h_1(X_1)]\,\mathrm{E}[h_2(X_2)].$$

Mit $h_1(X_1) = X_1 - \mathrm{E}[X_1]$ und $h_2(X_2) = X_2 - \mathrm{E}[X_2]$ folgt $\mathrm{Cov}[X_1, X_2] = 0$. Insbesondere sind unabhängige Zufallsvariablen unkorreliert.

Beweis. Stochastik, Teil 1. □

Mit Lemma 1.31(v) folgt für unabhängige Zufallsvariablen $X_1, \ldots, X_n$

$$\mathrm{Var}[X_1 + \cdots + X_n] = \sum_{i=1}^{n} \mathrm{Var}[X_i].$$

Bei den folgenden Resultaten unterscheiden wir die Fälle diskreter und stetiger Zufallsvariablen und geben jeweils eine Charakterisierung der Unabhängigkeit von Zufallsvariablen an.

Satz 1.36 (Produktform der Wahrscheinlichkeitsgewichte im diskreten Fall)**.** *Seien $X_1, \ldots, X_n$ diskrete Zufallsvariablen mit Werten in $S_1, \ldots, S_n$ und seien $\mu_1, \ldots, \mu_n$ Verteilungen auf $S_1, \ldots, S_n$. Folgende Aussagen sind äquivalent:*

(i) Die Familie $X_1, \ldots, X_n$ ist unabhängig und es gilt $X_i \sim \mu_i$, $i = 1, \ldots, n$.

(ii) Es gilt

$$\mathrm{P}(X_1 = a_1, \ldots, X_n = a_n) = \mu(a_1) \cdots \mu(a_n)$$

für alle $a_1 \in S_1, \ldots, a_n \in S_n$.

Beweis. Stochastik, Teil 1. □

Satz 1.37 (Produktform multivariater Dichten)**.** *Seien $X_1, \dots, X_n$ reellwertige Zufallsvariablen und $f_1, \dots, f_n$ nichtnegative integrierbare Funktionen auf $\mathbb{R}$ mit Integral 1. Folgende Aussagen sind äquivalent:*

(i) Die Familie $X_1, \dots, X_n$ ist unabhängig und X_i hat die Dichte f_i, $i = 1, \dots, n$.

(ii) Die gemeinsame Dichte von $(X_1, \dots, X_n)$ ist

$$f(a_1, \dots, a_n) = f_1(a_1) \cdots f_n(a_n), \quad (a_1, \dots, a_n) \in \mathbb{R}^n.$$

Beweis. Stochastik, Teil 1. □

2 Gesetze der großen Zahlen

In diesem Kapitel diskutieren wir das schwache und das starke Gesetz der großen Zahlen. Diese Gesetze gehören zu den wichtigsten Grenzwertsätzen in der Wahrscheinlichkeitstheorie und machen unter Anderem die Interpretation von Wahrscheinlichkeiten als Frequenzen bei unabhängigen Versuchen mathematisch rigoros: Sei A ein Ereignis mit $\mathrm{P}(A) = p$. Wir wiederholen ein Versuch unabhängig bei dem wir A als „Erfolg“ interpretieren und wir setzen

$$X_n \coloneqq \mathbb{1}_{\{A \text{ tritt ein beim } n\text{-ten Versuch}\}}.$$

Dann konvergiert die durchschnittliche Anzahl der Erfolge (also die Frequenz von Ausgängen des Experiments in A) in Wahrscheinlichkeit und fast sicher gegen p.

2.1 Wichtige Ungleichungen

In diesem Abschnitt wiederholen wir einige wichtige Ungleichungen, die uns später dabei helfen werden Beziehungen zwischen unterschiedlichen Konvergenzarten zu verstehen.

Satz 2.1 (Cauchy-Schwarz Ungleichung)**.** *Für reellwertige Zufallsvariablen X und Y mit endlichen zweiten Momenten $\mathrm{E}[X^2]$ und $\mathrm{E}[Y^2]$ gilt*

$$(\mathrm{E}[XY])^2 \leq (\mathrm{E}[|XY|])^2 \leq \mathrm{E}[X^2]\,\mathrm{E}[Y^2]. \tag{2.1}$$

Beweis. Stochastik, Teil 1 (oder Analysis bzw. Lineare Algebra). □

Wendet man die Cauchy-Schwarz Ungleichung auf die Zufallsvariablen $X - \mathrm{E}[X]$ und $Y - \mathrm{E}[Y]$ an, so folgt sofort

$$(\mathrm{Cov}[X, Y])^2 \leq \mathrm{Var}[X]\,\mathrm{Var}[Y]. \tag{2.2}$$

Satz 2.2 (Allgemeine Markov Ungleichung)**.** *Ist X eine Zufallsvariable und $h : \mathbb{R} \to (0, \infty)$ eine monoton wachsende Funktion, dann gilt für jedes $x \in \mathbb{R}$*

$$\mathrm{P}(X \geq x) \leq \frac{\mathrm{E}[h(X)]}{h(x)}. \tag{2.3}$$

Beweis. Für alle $x \in \mathbb{R}$ gilt

$$\mathrm{E}[h(X)] \geq \mathrm{E}[h(X)\mathbb{1}_{\{X \geq x\}}] \geq h(x)\,\mathrm{E}[\mathbb{1}_{\{X \geq x\}}] = h(x)\,\mathrm{P}(X \geq x).$$

□

Im nächsten Resultat stellen wir zwei Versionen von (2.3) vor.

Korollar 2.3. *Für alle $x > 0$ gilt*

$$\mathrm{P}(|X| \geq x) \leq \frac{\mathrm{E}[|X|^r]}{x^r},\ r \geq 0 \qquad \text{(Markov Ungleichung)} \tag{2.4}$$

und

$$\mathrm{P}(|X - \mathrm{E}[X]| \geq x) \leq \frac{\mathrm{Var}[X]}{x^2} \qquad \text{(Chebyshev Ungleichung).} \tag{2.5}$$

Beweis. Für (2.4) wende die allgemeine Markov Ungleichung mit $h(x) = |x|^r$ an. Für (2.5) wende die Markov Ungleichung (2.4) auf die Zufallsvariable $X - \mathrm{E}[X]$ mit $r = 2$ an. □

Natürlich sind die Ungleichungen (2.3), (2.4) und (2.5) nur dann nützlich, wenn die betreffenden Momente auf der rechten Seite jeweils endlich sind und berechnet oder abgeschätzt werden können. Die Stärke der Ungleichungen liegt in ihrer Universalität. Dafür liefern sie typischerweise nur grobe Abschätzungen.

Beispiel 2.4. Für $X \sim \mathcal{N}_{0,1}$ gilt

$$\mathrm{P}(-2 \leq X \leq 2) = 1 - \mathrm{P}(|X| \geq 2) \geq 1 - \frac{\mathrm{Var}[X]}{4} = 1 - \frac{1}{4} = 0.75.$$

Tatsächlich ist aber $\mathrm{P}(-2 \leq X \leq 2) \approx 0.95$.

Nach Definition ist die Varianz einer Zufallsvariable nichtnegativ und mit der Darstellung $\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$ folgt $\mathrm{E}[X^2] \geq (\mathrm{E}[X])^2$. Die Ungleichung könnten wir auch mit der Cauchy-Schwarz Ungleichung (setze dort $Y \equiv 1$), aber auch mit der folgenden Jensen Ungleichung bekommen.

Satz 2.5 (Jensen-Ungleichung)**.** *Ist $h : I \to \mathbb{R}$ eine konvexe Funktion und ist $\mathrm{P}(X \in I) = 1$ und existieren die Erwartungswerte von X und $h(X)$, dann gilt*

$$h\big(\mathrm{E}[X]\big) \leq \mathrm{E}[h(X)]. \tag{2.6}$$

Insbesondere gilt

$$|\mathrm{E}[X]| \leq \mathrm{E}[|X|] \quad \textit{und} \quad (\mathrm{E}[X])^2 \leq \mathrm{E}[X^2].$$

Beweis. Tangenten konvexer Funktionen liegen unterhalb des Funktionsgraphen. Für jedes $x_0 \in I$ gibt es also ein $a(x_0)$ mit

$$h(x) \geq h(x_0) + (x - x_0)a(x_0),\ x \in I. \tag{2.7}$$

Wenden wir diese Ungleichung auf $x = X$ und $x_0 = \mathrm{E}[X]$ an, so folgt

$$h(X) \geq h(\mathrm{E}[X]) + (X - \mathrm{E}[X])a(\mathrm{E}[X]).$$

Die Behauptung folgt wenn wir auf beiden Seite dieser Ungleichung den Erwartungswert nehmen. □

Natürlich liefert die Jensen-Ungleichung auch eine Abschätzung für konkave Funktionen. Ist nämlich h konkav, so ist $-h$ konvex und nach Anwendung der Jensen-Ungleichung auf $-h$ folgt

$$h\big(\mathrm{E}[X]\big) \geq \mathrm{E}[h(X)]. \tag{2.8}$$

Damit ist für positive Zufallsvariablen (sofern die fraglichen Erwartungswerte existieren):

$$(\mathrm{E}[X])^{-1} \leq \mathrm{E}[X^{-1}] \quad \text{und} \quad \log \mathrm{E}[X] \geq \mathrm{E}[\log X].$$

2.2 Konvergenzarten

Definition 2.6 (Konvergenz in Wahrscheinlichkeit, fast sicher und im p-ten Mittel)**.**

(i) Die Folge (X_n) konvergiert *in Wahrscheinlichkeit* oder *stochastisch* gegen X, wir schreiben $X_n \xrightarrow{\mathrm{P}} X$, wenn für alle $\varepsilon > 0$ gilt

$$\mathrm{P}(|X_n - X| \geq \varepsilon) \xrightarrow{n\to\infty} 0. \tag{2.9}$$

(ii) Die Folge (X_n) konvergiert *fast sicher* gegen X, wir schreiben $X_n \xrightarrow{\text{f.s.}} X$, wenn es eine Menge $N \subset \Omega$, $N \in \mathcal{A}$ mit $\mathrm{P}(N) = 0$ gibt, sodass

$$X_n(\omega) \xrightarrow{n\to\infty} X(\omega), \quad \text{für alle } \omega \notin N. \tag{2.10}$$

(iii) Seien $X, X_1, X_2, \dots \in L^p$ für $p > 0$, d.h. $\mathrm{E}[|X|^p] < \infty$. Die Folge (X_n) konvergiert *in L^p* oder *im p-ten Mittel* gegen X, wir schreiben $X_n \xrightarrow{L^p} X$, wenn

$$\mathrm{E}[|X_n - X|^p] \xrightarrow{n\to\infty} 0. \tag{2.11}$$

Bemerkung 2.7 (unendliche Folgen von Zufallsvariablen und betreffende Ereignisse)**.** Immer dann wenn wir von einer unendlichen Folge von Zufallsvariablen sprechen, die möglicherweise auch noch unabhängig sein sollen setzen wir eigentlich voraus, dass es Wahrscheinlichkeitsräume $(\Omega, \mathcal{A}, \mathrm{P})$ gibt, auf denen wir solche Folgen definieren können, was von vornherein nicht klar ist. Resultate, die es garantieren werden in den weiterführenden Wahrscheinlichkeitstheorievorlesungen bewiesen.

Gehen wir von Existenz eines Wahrscheinlichkeitsraumes $(\Omega, \mathcal{A}, \mathrm{P})$ auf dem wir $X, X_1, X_2, \dots$ gemeinsam definieren können, dann ist eine andere Möglichkeit fast sichere Konvergenz von X_n gegen X zu definieren zu fordern

$$\mathrm{P}(\lim_{n\to\infty} X_n = X) = 1.$$

Das setzt wiederum voraus, dass wir Ereignissen der Form $\{\lim_{n\to\infty} X_n = X\}$ eine Wahrscheinlichkeit zuordnen können. Dafür müssen sie Elemente von $\mathcal{A}$ sein, was auch nicht von vornherein klar ist. Auch das wird in den weiterführenden Vorlesungen gezeigt.

Lemma 2.8 (Stochastische Konvergenz von Summen)**.** *Seien $X, X_1, X_2, \dots$ und $Y, Y_1, Y_2, \dots$ Zufallsvariablen mit $X_n \xrightarrow{\mathrm{P}} X$ und $Y_n \xrightarrow{\mathrm{P}} Y$, dann gilt auch $X_n + Y_n \xrightarrow{\mathrm{P}} X + Y$.*

Beweis. Für alle $\varepsilon > 0$ gilt

$$\begin{aligned}\mathrm{P}(|X_n + Y_n - (X+Y)| \geq \varepsilon) &\leq \mathrm{P}(|X_n - X| + |Y_n - Y| \geq \varepsilon) \\ &\leq \mathrm{P}(|X_n - X| \geq \varepsilon/2 \text{ oder } |Y_n - Y| \geq \varepsilon/2) \\ &\leq \mathrm{P}(|X_n - X| \geq \varepsilon/2) + \mathrm{P}(|Y_n - Y| \geq \varepsilon/2) \\ &\xrightarrow{n\to\infty} 0.\end{aligned}$$

□

Satz 2.9. *Konvergenz in L^p impliziert Konvergenz in Wahrscheinlichkeit.*

Beweis. Mit Markov-Ungleichung (2.4) gilt für jedes $\varepsilon > 0$

$$\mathrm{P}(|X_n - X| \geq \varepsilon) \leq \varepsilon^{-p}\,\mathrm{E}[|X_n - X|^p] \xrightarrow{n\to\infty} 0.$$

□

Satz 2.10. *Fast sichere Konvergenz impliziert Konvergenz in Wahrscheinlichkeit.*

Beweis. Stochastik, Teil 1. □

Die Umkehrungen in den Sätzen 2.9 und 2.10 gelten ohne weitere Voraussetzungen nicht. Eine teilweise Umkehrung der Aussage von Satz 2.10 liefert das folgende Resultat.

Satz 2.11. *Gilt $X_n \xrightarrow{\mathrm{P}} X$, so existiert eine Teilfolge $X_{n(j)}$, fast sicher gegen X konvergiert.*

Beweis. Stochastik, Teil 1. □

Satz 2.12 (Lemma von Borel-Cantelli)**.** *Sei $(A_n)_{n\in\mathbb{N}}$ eine Folge von Ereignissen und sei (u.o. steht für „unendlich oft“)*

$$A = \{A_n \text{ u.o.}\} = \{A_n \text{ tritt für unendlich viele } n \text{ ein}\}.$$

Dann gelten folgende Aussagen:

(i) Ist $\sum_{n=1}^{\infty} \mathrm{P}(A_n) < \infty$, so ist $\mathrm{P}(A) = 0$.

(ii) Ist $(A_n)_{n\in\mathbb{N}}$ unabhängig und gilt $\sum_{n=1}^{\infty} \mathrm{P}(A_n) = \infty$, so ist $\mathrm{P}(A) = 1$.

Beweis. Stochastik, Teil 1. □

Aussagen (i) und (ii) im obigen Satz werden üblicherweise als das erste, beziehungsweise zweite Borel-Cantelli Lemma bezeichnet. Die Voraussetzung der Unabhangigkeit in (ii) ist entscheidend (kann aber etwas auf verschiedene Weisen etwas abgeschwächt werden). Ist nämlich $A_n = A_1$ für alle $n \geq 2$ und $\mathrm{P}(A_1) = 1/2$, dann ist $\mathrm{P}(A) = \mathrm{P}(A_1) = 1/2$ und $\sum_{n=1}^{\infty} \mathrm{P}(A_n) = \infty$.

Für unabhängige Folgen von Ereignissen $(A_n)_{n\in\mathbb{N}}$, liefern die beiden Borel-Cantelli Lemmata ein „Null-Eins Gesetz“, es gilt nämlich $\mathrm{P}(A_n \text{ u.o.}) \in \{0, 1\}$ und

$$\mathrm{P}(A_n \text{ u.o.}) = 0 \iff \sum_{n=1}^{\infty} \mathrm{P}(A_n) < \infty,$$
$$\mathrm{P}(A_n \text{ u.o.}) = 1 \iff \sum_{n=1}^{\infty} \mathrm{P}(A_n) = \infty.$$

In der Wahrscheinlichketstheorie gibt es weitere „Null-Eins Gesetze“. Einige davon werden in den weiterführenden Vorlesungen behandelt.

Mit dem Borel-Cantelli Lemma lässt sich fast sichere Konvergenz von Folgen von Zufallsvariablen auf Konvergenz bestimmter Reihen zurückführen.

Satz 2.13 (Charakterisierung der fast sicheren Konvergenz)**.**

(i) Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsvariablen und X eine Zufallsvariable mit

$$\sum_{n=1}^{\infty} \mathrm{P}(|X_n - X| \geq \varepsilon) < \infty \quad \textit{für alle } \varepsilon > 0. \tag{2.12}$$

Dann gilt $X_n \xrightarrow{\text{f.s.}} X$.

(ii) Ist die Folge $(X_n)_{n\in\mathbb{N}}$ unabhängig und c eine Konstante, dann gilt $X_n \xrightarrow{\text{f.s.}} c$ genau dann, wenn

$$\sum_{n=1}^{\infty} \mathrm{P}(|X_n - c| \geq \varepsilon) < \infty \quad \textit{für alle } \varepsilon > 0. \tag{2.13}$$

Beweis. Stochastik, Teil 1. □

2.3 Gesetze der großen Zahlen

Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge reellwertiger Zufallsvariablen. Wir sagen, dass die Folge $(X_n)_{n\in\mathbb{N}}$ dem *schwachen Gesetz der großen Zahlen* genügt, wenn

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathrm{E}[X_i]) \xrightarrow{\mathrm{P}} 0, \quad \text{für } n \to \infty. \tag{2.14}$$

Wir sagen, dass die Folge $(X_n)_{n\in\mathbb{N}}$ dem *starken Gesetz der großen Zahlen* genügt, wenn

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathrm{E}[X_i]) \xrightarrow{\text{f.s.}} 0, \quad \text{für } n \to \infty. \tag{2.15}$$

Für *Folgen identisch verteilter paarweise unabhängiger* Zufallsvariablen sind die obigen Aussagen gleichbedeutend mit

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{P}} \mathrm{E}[X_1] \tag{2.16}$$

bzw.

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{f.s.}} \mathrm{E}[X_1]. \tag{2.17}$$

Die allgemeine Version des starken Gesetzes der großen Zahlen für Folgen von paarweise unabhängigen Zufallsvariablen ist das folgende Resultat.

Satz 2.14 (Starkes GGZ von Etemadi und Kolmogorov)**.** *Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge integrierbarer, paarweise unabhängiger und identisch verteilter reellwertiger Zufallsvariablen. Dann gilt*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{f.s.}} \mathrm{E}[X_1], \quad n \to \infty.$$

Der Beweis dieses Satzes ist anspruchsvoll und kann mit den uns zur Verfügung stehenden Mitteln nicht bewiesen werden. Der Beweis wird in den weiterführenden Vorlesungen gegeben.

Natürlich impliziert das starke GGZ das schwache GGZ. Wir werden mit uns zur Verfügung stehenden Mitteln Versionen von beiden GGZ mit stärkeren Voraussetzungen als in Satz 2.14 beweisen. Die stärkeren Voraussetzungen sind keinesfalls notwendig, erleichtern die Beweise aber erheblich.

Satz 2.15 ((ein) schwaches Gesetz der großen Zahlen)**.** *Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge unabhängiger (oder unkorrelierter) identisch verteilter Zufallsvariablen mit $\mathrm{Var}[X_1] < \infty$. Dann genügt die Folge $(X_n)_{n\in\mathbb{N}}$ dem schwachen Gesetz der großen Zahlen.*

Beweis. Es gilt $\mathrm{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_1] = \mathrm{E}[X_1]$ und mit der Chebyshev Ungleichung folgt für jedes $\varepsilon > 0$

$$\begin{aligned}\mathrm{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}[X_1]\Big| > \varepsilon\Big) &\le \frac{\mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} X_i]}{\varepsilon^2} \\ &= \frac{\sum_{i=1}^{n}\mathrm{Var}[X_i]}{n^2\varepsilon^2} = \frac{\mathrm{Var}[X_1]}{n\varepsilon^2} \xrightarrow{n\to\infty} 0.\end{aligned} \tag{2.18}$$

□

Eine Stärke des Beweises mit der Chebyshev Ungleichung liegt darin, dass man leicht die Unabhängigkeits- bzw. die Unkorreliertheitsvoraussetzung abschwächen kann. In (2.18) hätte man dann noch eine Doppelsumme mit Kovarianzen und mit dem Vorfaktor $2/n^2$. Die Voraussetzungen müssen dann so gewählt werden, dass also

$$\frac{1}{n^2}\sum_{i<j}\operatorname{Cov}[X_i, X_j] \xrightarrow{n\to\infty} 0$$

gilt.

Bei dem Beweis der folgenden Version des Starkes Gesetzes der großen Zahlen benutzen wir dieselbe Idee wie bei dem Beweis von Satz 2.15 zusammen mit der Charakterisierung der fast sicheren Konvergenz aus Satz 2.13.

Satz 2.16 ((ein) starkes Gesetz der großen Zahlen)**.** *Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge unabhängiger identisch verteilter Zufallsvariablen mit $\mathrm{E}[X_1^4] < \infty$. Dann genügt die Folge $(X_n)_{n\in\mathbb{N}}$ dem starken Gesetz der großen Zahlen.*

Beweis. Wie angekündigt ist der Ansatz derselbe wie im Beweis von Satz 2.15. Wir möchten für $\varepsilon > 0$

$$\mathrm{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}[X_1]\Big| > \varepsilon\Big)$$

abschätzen und zwar so, dass die Schranken in n summierbar sind. Anderenfalls können wir Satz 2.13 nicht anwenden.

Ohne Einschränkung der Allgemeinheit nehmen wir an $\mathrm{E}[X_1] = 0$ ansonsten gehen wir zu Zufallsvariablen $X_i - \mathrm{E}[X_i]$ über (wir zentrieren also die Zufallsvariablen).

Mit der obigen Annahme ist zu zeigen ist: Für alle $\varepsilon > 0$ gilt

$$\sum_{n=1}^{\infty}\mathrm{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i\Big| > \varepsilon\Big) < \infty \tag{2.19}$$

Für eine Zufallsvariable X mit $\mathrm{E}[X^4] < \infty$ gilt mit der Markov-Ungleichung, siehe (2.4),

$$\mathrm{P}(|X| \geq \varepsilon) \leq \frac{\mathrm{E}[X^4]}{\varepsilon^4}.$$

Diese Ungleichung wenden wir auf die Zufallsvariable $\frac{1}{n}\sum_{i=1}^{n} X_i$ an, um (2.19) zu zeigen.

Es gilt

$$\mathrm{E}\Big[\Big(\sum_{i=1}^{n} X_i\Big)^4\Big] = \sum_{i_1,i_2,i_3,i_4=1}^{n} \mathrm{E}[X_{i_1}X_{i_2}X_{i_3}X_{i_4}].$$

Wegen $\mathrm{E}[X_i] = 0$ sehen wir mit Satz 1.35, dass nur Terme zur Summe beitragen in denen je zwei oder vier Indizes gleich sind. Damit gilt

$$\begin{aligned}\mathrm{E}\Big[\Big(\sum_{i=1}^{n} X_i\Big)^4\Big] &= \sum_{i=1}^{n} \mathrm{E}[X_i^4] + \sum_{i\neq j} \mathrm{E}[X_i^2]\,\mathrm{E}[X_j^2] \\ &\leq n\,\mathrm{E}[X_1^4] + n^2(\mathrm{E}[X_1^2])^2\end{aligned}$$

Es folgt

$$\mathrm{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i\Big| > \varepsilon\Big) \leq \frac{\mathrm{E}\Big[\Big(\sum_{i=1}^{n} X_i\Big)^4\Big]}{n^4\varepsilon^4} \leq \frac{n\,\mathrm{E}[X_1^4] + n^2(\mathrm{E}[X_1^2])^2}{n^4\varepsilon^4}$$

und die rechte Seite ist offenbar summierbar, was (2.19) zeigt. □

Zusammenfassend für diesen Abschnitt lässt sich sagen, dass mit den Chebyshev und Markov Ungleichungen mit relativ elementaren Mitteln schwache und starke Gesetze der großen Zahlen bewiesen können, aber nur unter ziemlich starken Momentenannahmen. Um das starke Gesetz unter den Minimalannahmen von Satz 2.14 zu beweisen bedarf es stärkerer Ungleichungen. In diesem Fall ist es die *Kolmogorov Ungleichung* auf die wir hier aber nicht eingehen werden.

3 Diskrete Zufallsvariablen

In diesem Kapitel wiederholen wir etwas ausführlicher einige wichtige diskrete Verteilungen und diskutieren deren Beziehungen untereinander.

3.1 Binomialverteilung und verwandte Verteilungen

Ist $(\Omega, \mathcal{A}, \mathrm{P})$ und $A \in \mathcal{A}$ ein Ereignis, so ist bekanntlich die Zufallsvariable $X = \mathbb{1}_A$ *Bernoulli verteilt mit Parameter* $p = \mathrm{P}(A)$; kurz $X \sim \mathrm{Ber}_p$. Die Verteilung von X können wir in der Form schreiben

$$\mathrm{P}(X = a) = p^a (1-p)^{1-a}, \quad a \in \{0, 1\}. \tag{3.1}$$

Für den Erwartungswert und die Varianz von X gilt

$$\begin{aligned} \mathrm{E}[X] &= 0 \cdot (1-p) + 1 \cdot p = p, \\ \mathrm{Var}[X] &= \mathrm{E}[X^2] - \mathrm{E}[X]^2 = p - p^2 = p(1-p). \end{aligned}$$

Nun wiederholen wir (unabhängig) das Zufallsexperiment n-mal bei dem A eintritt oder nicht. Wie üblich interpretieren wir ersteres als „Erfolg“ und letzteres als „Misserfolg“. Wir erhalten einen Vektor $(X_1, \dots, X_n)$ von unabhängigen identisch verteilten Bernoulli-Zufallsvariablen mit Parameter $p = \mathrm{P}(A)$. Mit (3.1) und Satz 1.36 sieht man leicht, dass die gemeinsame Verteilung des Vektors durch

$$\begin{aligned} \mathrm{P}((X_1, \dots, X_n) = a) &= \prod_{i=1}^{n} p^{a_i} (1-p)^{1-a_i} \\ &= p^{\sum_i a_i} (1-p)^{n - \sum_i a_i}, \quad a = (a_1, \dots, a_n) \in \{0,1\}^n \end{aligned} \tag{3.2}$$

gegeben ist. Im Fall $p = 1/2$ ist die Verteilung eine Gleichverteilung auf $\{0,1\}^n$ und jedes Element hat die Masse $1/2^n$.

Oft ist man nicht an der genauen Abfolge der „Erfolge“ und „Misserfolge“ interessiert, sondern nur an der jeweiligen Anzahl. Die Anzahl der „Erfolge“ ist gegeben durch $Z = \sum_{i=1}^{n} X_i$. Das ist eine Zufallsvariable mit Werten in $\{0, \dots, n\}$.

Um die Verteilung von Z zu bestimmen, erinnern wir an die folgende Transformationsformel.

Satz 3.1. *Sei X eine diskrete Zufallsvariable mit Wertebereich S und $h : S \to S'$. Für $Y = h(X)$ ist die Verteilung gegeben durch*

$$\mathrm{P}(Y = b) = \sum_{a \in h^{-1}(b)} \mathrm{P}(X = a), \quad b \in S'. \tag{3.3}$$

Beweis. Es gilt

$$\{\omega \in \Omega : Y(\omega) = b\} = \{\omega \in \Omega : h(X(\omega)) = b\} = \{\omega \in \Omega : X(\omega) = h^{-1}(b)\},$$

was die ausführlich Schreibweise von $\{Y = b\} = \{X \in h^{-1}(b)\}$ ist. Damit folgt die Aussage sofort. □

Es gilt $Z = h(X_1, \dots, X_n)$ für $h : \{0,1\}^n \to \{0, \dots, n\}$, $h(x_1, \dots, x_n) = \sum_i x_i$. Mit (3.3) und (3.2) erhalten wir

$$\begin{aligned}
\mathrm{P}(Z = k) &= \mathrm{P}(h(X_1, \dots, X_n) = k) \\
&= \sum_{a \in h^{-1}(k)} \mathrm{P}((X_1, \dots, X_n) = a) \\
&= \sum_{a \in h^{-1}(k)} p^{\sum_i a_i}(1-p)^{n-\sum_i a_i} = \sum_{a \in h^{-1}(k)} p^k (1-p)^{n-k} \\
&= |h^{-1}(k)| p^k (1-p)^{n-k}.
\end{aligned}$$

Hier ist $|h^{-1}(k)|$ die Anzahl von 01 Folgen der Länge n mit genau k Einsen. Sie ist gegeben durch $\binom{n}{k}$, also durch die Anzahl von k-elementigen Teilmengen von $\{0, \dots, n\}$. Wir erhalten

$$\mathrm{P}(Z = k) = b(k; n, p) \coloneqq \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, \dots, n\} \tag{3.4}$$

und erkennen darin die *Binomialverteilung mit Parametern n und p.* Also kann eine binomialverteilte Zufallsvariable als Summe von unabhängigen und identisch verteilten Bernoulli-Zufallsvariablen aufgefasst werden. Das macht die Berechnung des Erwartungswertes und der Varianz der Binomialverteilung denkbar einfach: Für $Z \sim \mathrm{Bin}_{n,p}$ und $X_1, \dots, X_n$ u.i.v. mit $X_i \sim \mathrm{Ber}_p$ gilt

$$\mathrm{E}[Z] = \mathrm{E}\Big[\sum_{i=1}^n X_i\Big] = \sum_{i=1}^n \mathrm{E}[X_i] = np$$

und

$$\mathrm{Var}[Z] = \mathrm{Var}\Big[\sum_{i=1}^n X_i\Big] = \sum_{i=1}^n \mathrm{Var}[X_i] = np(1-p).$$

Die direkte Berechnung mit Definition 1.21 ist auch elementar, erfordert aber zumindest mehr Schreibarbeit. Wir demonstrieren die direkte Methode. Es gilt

$$\begin{aligned}
\mathrm{E}[Z] &= \sum_{k=0}^n k \,\mathrm{P}(Z = k) = \sum_{k=1}^n k \binom{n}{k} p^k (1-p)^{n-k} = np \sum_{k=1}^n \binom{n-1}{k-1} p^{k-1} (1-p)^{n-1-(k-1)} \\
&= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} = np.
\end{aligned}$$

Die letzte Gleichung folgt sofort, da $\binom{n-1}{k}p^k(1-p)^{n-1-k}$ die Wahrscheinlichkeitsgewichte der $\mathrm{Bin}_{n-1,p}$-Verteilung sind die sich zu 1 aufsummieren.

Zur Berechnung der Varianz diskreter Zufallsvariablen X ist die folgende Formel oft hilfreich

$$\mathrm{Var}[X] = \mathrm{E}[X(X-1)] + \mathrm{E}[X] - \mathrm{E}[X]^2. \tag{3.5}$$

Für $Z \sim \mathrm{Bin}_{n,p}$ gilt

$$\begin{aligned}\mathrm{E}[Z(Z-1)] &= \sum_{k=0}^{n} k(k-1)\,\mathrm{P}(Z=k) = \sum_{k=2}^{n} k(k-1)\binom{n}{k}p^k(1-p)^{n-k} \\ &= n(n-1)p^2 \sum_{k=1}^{n}\binom{n-2}{k-2}p^{k-2}(1-p)^{n-2-(k-2)} \\ &= n(n-1)p^2.\end{aligned}$$

Mit (3.5) erhalten wir

$$\mathrm{Var}[Z] = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

Übung 3.2. *Sei Z_n binomialverteilt mit Parametern n und p. Was können Sie über das Verhalten von Z_n/n für $n \to \infty$ sagen?*

Wir definieren eine Verallgemeinerung der Binomialverteilung und geben später eine Interpretation an. Für $k_1, \ldots, k_r \in \mathbb{N}_0$ mit $k_1 + \cdots + k_r = n$ ist der *Multinomialkoeffizient* definiert durch

$$\binom{n}{k_1, \ldots, k_r} := \frac{n!}{k_1! \cdots k_r!}.$$

Für $n \in \mathbb{N}$ seien $p_1, \ldots, p_r \geq 0$ mit $p_1 + \cdots + p_r = 1$. Dann heißt der Zufallsvektor $X = (X_1, \ldots, X_r)$ *multinomialverteilt* mit Parametern $(n; p_1, \ldots, p_r)$, wenn der Wertebereich von X durch

$$S_{n,r} = \{(k_1, \ldots, k_r) \in \mathbb{N}_0^r : k_1 + \cdots + k_r = n\}$$

gegeben ist und es gilt

$$\mathrm{P}(X = (k_1, \ldots, k_r)) = \binom{n}{k_1, \ldots, k_r} p_1^{k_1} \cdots p_r^{k_r}.$$

Wir haben oben eine binomialverteilte Zufallsvariable als Anzahl von Erfolgen bei n unabhängigen und identischen Experimenten aufgefasst. Da es nur zwei Ausgänge gab bestimmt die Zufallsvariable natürlich auch die Anzahl der Misserfolge. Die *Multinomialverteilung* ist eine Verallgemeinerung der Binomialverteilung auf den Fall wenn mehr als zwei Ausgänge des Zufallsexperimentes möglich sind. Seien $A_1, \ldots, A_r$ disjunkte Mengen mit $\Omega = \cup_{i=1}^r A_i$ und $\mathrm{P}(A_i) = p_i$, $i = 1, \ldots, r$. Wir wiederholen ein Experiment unabhängig n-mal und bezeichnen mit X_i die Anzahl der Ausgänge in A_i, $i = 1, \ldots, r$. Dann ist $X = (X_1, \ldots, X_r)$ *multinomialverteilt* mit Parameter $(n; p_1, \ldots, p_r)$.

Übung 3.3. *Bestimmen Sie die Verteilung von X_i, bzw. von $X_i + X_j$ für $i \neq j$ wenn $X = (X_1, \ldots, X_r)$ multinomialverteilt mit Parameter $(n; p_1, \ldots, p_r)$ ist.*

3.2 Hypergeometrische Verteilung und die Beziehung zu Binomialverteilung

Oft lassen sich (endliche) diskrete Modelle als Urnenmodelle interpretieren. Betrachten wir eine Urne mit N Kugeln von denen $M \leq N$ markiert sind und ziehen n-mal jeweils mit Zurücklegen eine Kugel aus der Urne. Sei Z die Zufallsvariable die angibt wie oft eine markierte Kugel gezogen wurde. Dann ist $Z \sim \mathrm{Bin}_{n,p}$ mit $p = M/N$.

Ziehen wir die Kugeln dagegen ohne Zurücklegen und ist Z wieder die Anzahl der gezogenen markierten Kugeln, dann ist Z hypergeometrisch verteilt mit Parametern n, N und M, kurz $Z \sim \mathrm{Hyp}_{n;N,M}$, d.h. für $k \in \{0, \ldots, n\}$ ist

$$\begin{aligned} \mathrm{P}(Z = k) &= \mathrm{P}(\text{es werden } k \text{ markierte und } n-k \text{ unmarkierte Kugeln gezogen}) \\ &= h(k; n, N, M) := \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}. \end{aligned}$$

Dabei verwenden wir die Konvention $\binom{n}{k} = 0$ für $k < 0$ oder $k > n$.

Satz 3.4 (Erwartungswert und Varianz hypergeometrischer Verteilung)**.** *Ist $X \sim \mathrm{Hyp}_{n;N,M}$, dann gilt*

$$\mathrm{E}[X] = np \quad \textit{und} \quad \mathrm{Var}[X] = np(1-p)\Big(1 - \frac{n-1}{N-1}\Big), \tag{3.6}$$

wobei $p = M/N$.

Beweis. Die Berechnung ist eine Übung, kann aber z.B. in Kersting and Wakolbinger (2010) nachgelesen werden. □

Intuitiv ist klar, dass für große N und M und im Vergleich dazu kleiner Stichprobengröße n es nicht so sehr darauf ankommt, ob mit oder ohne Zurücklegen gezogen wird. Das nächste Resultat zeigt, dass die hypergeometrische Verteilung durch die Binomialverteilung approximiert werden kann.

Satz 3.5. *Für alle $j \in \mathbb{N}$ seien $M_j \leq N_j$ natürliche Zahlen mit $M_j, N_j \to \infty$ und $M_j/N_j \to p \in [0,1]$. Dann gilt für jedes $n \in \mathbb{N}$ (vgl. (3.4))*

$$\lim_{j\to\infty} h(k; n, N_j, M_j) = b(k; n, p),\ 0 \leq k \leq n.$$

Beweis. Übung! □

Beachten Sie, dass unter den Voraussetzungen des obigen Satzes auch der Erwartungswert und die Varianz der hypergeometrischen Verteilung gegen die der binomialverteilung konvergieren. Das sieht man leicht mit der in (3.6) angegebenen Darstellung des Erwartungswertes und der Varianz.

Das folgende Beispiel stammt aus Feller (1968), was ein sehr empfehlenswertes Buch mit einer Fülle von interessanten Beispielen ist.

Beispiel 3.6 (Ausflug in Statistik, Maximum Likelihood Schätzung einer Populationsgröße)**.** In einem See soll die unbekannte Populationsgröße von Fischen geschätzt werden. Es werden 1000 Fische gefangen, markiert und wieder in den See entlassen. Nach einer Woche (dann ist die Population wieder durchmischt und der Bestand ist gleichgroß) werden wieder 1000 Fische gefangen und es werden 100 markierte gezählt.

Frage: Was ist eine gute Schätzung der Populationsgröße N?

Wir modellieren das Problem wie folgt:

$$\begin{aligned} N &= \text{Populationsgröße,} \\ M &= 1000, \text{Anzahl der markierten Fische, die beim ersten mal gefangen wurden,} \\ n &= 1000, \text{Stichprobengröße, Anzahl der Fische beim zweiten Fang,} \\ k &= 100, \text{Anzahl der markierten Fische beim zweiten Fang.} \end{aligned}$$

Parameter N ist zwar unbekannt, aber *nicht zufällig*. Wir können mit dem unbekannten Parameter rechnen als würden wir es kennen. Weil es nicht zufällig ist, macht es keinen Sinn nach der Wahrscheinlichkeit für $N > 6000$ zu fragen. Was wir nur mit Sicherheit wissen ist, dass die obige Beobachtung nur dann möglich ist, wenn $N \geq M + n - k = 1900$ ist.

Wäre $N = 1900$ dann hätte die obige Beobachtung die Wahrscheinlichkeit

$$h(100; 1000, 1000, 1900) = \frac{\binom{1000}{100}\binom{1900-1000}{1000-100}}{\binom{1900}{1000}} = \frac{(1000!)^2}{100! \cdot 1900!}.$$

Mit der Stirling-Formel $n! \approx \sqrt{2\pi n} n^n e^{-n}$ kann man sehen, dass $h(100; 1000, 1000, 1900) \approx 10^{-430}$ ist, was also die Beobachtung bei $N = 1900$ extrem unwahrscheinlich macht.

Die Idee des *Maximum-Likelihood-Prinzips* ist es den Parameter N so zu wählen, dass die Beobachtung am wahrscheinlichsten (plausibelsten) ist. Wir wollen also

$$h(100; 1000, 1000, N) = \frac{\binom{1000}{100}\binom{N-1000}{900}}{\binom{N}{1000}}$$

maximieren. Dazu setzen wir $A_N := h(k; n, N, M)$. Dann ist

$$\frac{A_N}{A_{N-1}} = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \frac{\binom{N-1}{n}}{\binom{M}{k}\binom{N-1-M}{n-k}} = \frac{(N-M)(N-n)}{N(N-M-n+k)}$$

und man kann leicht zeigen, dass $A_N/A_{N-1} < 1$, wenn $Mn < kN$ ist, und dass $A_N/A_{N-1} > 1$, wenn $Mn > kN$ ist. Mit anderen Worten wächst A_N auf der Menge $N < Mn/k$ und fällt auf der Menge $N > Mn/k$. Plausibler Schätzer (der die Beobachtung am wahrscheinlichsten macht) ist also $\widehat{N} = \lfloor Mn/k \rfloor$. Mit den Zahlen von oben erhalten wir

$$\widehat{N} = \frac{1000 \cdot 1000}{100} = 10000.$$

3.3 Poissonverteilung und Poissonapproximation der Binomialverteilung

Es sei $\lambda \geq 0$. Eine $\mathbb{N}_0$-wertige Zufallsvariable X ist Poissonverteilt mit Parameter λ, kurz $X \sim \mathrm{Poi}_\lambda$, wenn

$$\mathrm{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots.$$

Satz 3.7. *Ist $X \sim \mathrm{Poi}_\lambda$, so gilt*

$$\mathrm{E}[X] = \lambda \quad \textit{und} \quad \mathrm{Var}[X] = \lambda.$$

Beweis. Es gilt

$$\begin{aligned}\mathrm{E}[X] &= \sum_{k=0}^{\infty} k\,\mathrm{P}(X = k) = \sum_{k=1}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} \\ &= \lambda\sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} = \lambda\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = \lambda.\end{aligned}$$

Die letzte Summe ist 1, da wir dort die Wahrscheinlichkeitsgewichte der Poi_λ-Verteilung aufsummieren.

Für die Berechnung der Varianz nutzen wir die Darstellung (3.5). Es gilt

$$\begin{aligned}\mathrm{E}[X(X-1)] &= \sum_{k=0}^{\infty} k(k-1)\,\mathrm{P}(X = k) = \sum_{k=2}^{\infty} k(k-1) e^{-\lambda}\frac{\lambda^k}{k!} \\ &= \lambda^2\sum_{k=2}^{\infty} e^{-\lambda}\frac{\lambda^{k-2}}{(k-2)!} = \lambda^2\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = \lambda^2.\end{aligned}$$

Mit (3.5) folgt

$$\mathrm{Var}[X] = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

□

Übung 3.8. *Es sei $X \sim \mathrm{Poi}_\lambda$. Berechnen Sie* $\mathrm{E}[e^{uX}]$.

Poissonverteilung ist in der Wahrscheinlichkeitstheorie eine sehr wichtige Verteilung, weil sie (bzw. „Variationen“ davon) oft als Approximation von Folgen von Verteilungen auftaucht oder zumindest ein Baustein von solchen Approximationen ist. Folgendes Resultat beleuchtet den Zusammenhang zwischen den Binomial- und Poissonverteilungen. Grob, lässt es sich wie folgt interpretieren: Hat ein Ereignis eine sehr kleine Wahrscheinlichkeit und interpretieren wir das Eintreten des Ereignisses als „Erfolg“, dann ist bei einer großen Anzahl von Versuchen die Anzahl der Erfolge in etwa Poissonverteilt.

Satz 3.9 (Poissonapproximation der Binomialverteilung). *Es sei $\lambda > 0$ und sei X_n, $n = 1, 2, \ldots$ eine Folge mit $X_n \sim \mathrm{Bin}_{n,p_n}$. Gilt*

$$\mathrm{E}[X_n] = np_n \xrightarrow{n\to\infty} \lambda,$$

dann folgt

$$\mathrm{P}(X_n = k) \xrightarrow{n\to\infty} e^{-\lambda}\frac{\lambda^k}{k!}, \quad \textit{für alle } k = 0, 1, 2, \ldots.$$

Beweis. Im Prinzip müssen nur die Wahrscheinlichkeitsgewichte der Binomialverteilung geschickt umgeschrieben werden. Es gilt

$$\begin{aligned}\binom{n}{k}p_n^k(1-p_n)^{n-k} &= \frac{n!}{(n-k)!}\cdot\frac{1}{k!}\cdot p_n^k\cdot(1-p_n)^{n-k}\\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k}\cdot\frac{1}{k!}\cdot(np_n)^k\cdot\big(1-\frac{np_n}{n}\big)^n\cdot(1-p_n)^{-k}\\ &\xrightarrow{n\to\infty} 1\cdot\frac{1}{k!}\cdot\lambda^k\cdot e^{-\lambda}\cdot 1.\end{aligned}$$

□

3.4 Wartezeiten

In diesem Abschnitt behandeln wir einige weitere Verteilungen, die mit Folgen von unabhängigen Bernoulli verteilten Zufallsvariablen in einem engen Zusammenhang stehen.

Wir betrachten also eine Folge $X_1, X_2, \ldots$ eine von unabhängigen und identisch Verteilten Zufallsvariablen mit $X_i \sim \mathrm{Ber}_p$ für ein $p \in (0, 1)$. Wir könnten natürlich die Fälle $p \in \{0, 1\}$ mitbehandeln, schließen sie aber aus um keine trivialen Fallunterscheidungen machen zu müssen.

3.4.1 Geometrische Verteilung

Sei Z die Wartezeit bis zum ersten „Erfolg“ bzw. der ersten „Eins“. Formal ist

$$Z \coloneqq \inf\{n \in \mathbb{N} : X_n = 1\}$$

mit der Konvention $\inf\{\emptyset\} = +\infty$. Dann ist Z eine $\mathbb{N}$-wertige Zufallsvariable mit (vgl. (3.2))

$$\mathrm{P}(Z = k) = \mathrm{P}(X_1 = X_2 = \cdots = X_{k-1} = 0, X_k = 1) = p\cdot(1-p)^{k-1}, \quad k \in \mathbb{N}. \tag{3.7}$$

Die Zufallsvariable Z ist *geometrisch verteilt mit Parameter p*, wir schreiben $Z \sim \mathrm{Geo}_p$. Es gilt

$$\mathrm{P}(Z > k) = \sum_{i=k+1}^{\infty}\mathrm{P}(Z = i) = p\cdot\sum_{i=k+1}^{\infty}(1-p)^{i-1} = p\cdot\frac{(1-p)^k}{p} = (1-p)^k \tag{3.8}$$

und durch diese Form der *oberen Verteilungsfunktion* erklärt sich der Name der geometrischen Verteilung. Alternativ kann man (3.8) auch leicht einsehen, wenn man sich klarmacht, dass

$$\{Z > k\} = \{X_1 = X_2 = \cdots = X_k = 0\}$$

gilt. Natürlich ist durch (3.8) die Verteilungsfunktion von Z und somit auch die Verteilung eindeutig bestimmt: Es gilt

$$F_Z(x) = \mathrm{P}(Z \leq x) = 1 - \mathrm{P}(Z > x) = \begin{cases} 1 - (1-p)^{\lfloor x \rfloor} & : x \geq 0, \\ 0 & : x < 0. \end{cases} \tag{3.9}$$

Bemerkung 3.10 (Variante der geometrischen Verteilung)**.** Sei $\widetilde{Z}$ die Anzahl der Fehlversuche bis eine „Eins“ auftritt. Es ist also $\widetilde{Z} = Z - 1$ und es gilt

$$\mathrm{P}(\widetilde{Z} = k) = \mathrm{P}(X_1 = X_2 = \cdots = X_k = 0, X_{k+1} = 1) = p \cdot (1-p)^k, \quad k \in \mathbb{N}_0. \tag{3.10}$$

Auch diese Verteilung wird oft als geometrische Verteilung mit Parameter p bezeichnet. Wir meinen hier immer die in (3.7) angegebene Verteilung wenn wir von geometrischer Verteilung sprechen. Im Zweifel sollte man zusammen mit Verteilung den Erwartungswert angeben. In etwa so: „Sei X geometrisch verteilt mit Parameter p und Erwartungswert $1/p$“.

Satz 3.11. *Ist $Z \sim \mathrm{Geo}_p$, so gilt*

$$\mathrm{E}[Z] = \frac{1}{p} \quad \textit{und} \quad \mathrm{Var}[Z] = \frac{1}{p}\Big(\frac{1}{p} - 1\Big).$$

Beweis. Wir geben zwei Varianten der Berechnung des Erwartungswertes an.

Variante 1: Es gilt

$$\begin{aligned} \mathrm{E}[Z] &= \sum_{k=1}^{\infty} kp \cdot (1-p)^{k-1} = p \cdot \sum_{k=1}^{\infty} k(1-p)^{k-1} \\ &= p \cdot \frac{\mathrm{d}}{\mathrm{d}p}\Big(-\sum_{k=0}^{\infty}(1-p)^k\Big) = p \cdot \frac{\mathrm{d}}{\mathrm{d}p}\Big(-\frac{1}{p}\Big) = \frac{1}{p}. \end{aligned}$$

Variante 2: Wir verwenden hierfür Satz 1.23. Es gilt

$$\mathrm{E}[Z] = \sum_{n=0}^{\infty} \mathrm{P}(Z > n) = \sum_{n=0}^{\infty} (1-p)^n = \frac{1}{1-(1-p)} = \frac{1}{p}.$$

Die Berechnung der Varianz ist eine Übung! Genauso wie bei der Berrechnung der Varianz von Binomial und Poisson verteilten Zufallsvariablen ist auch hier die Formel (3.5) hilfreich. □

Geometrische Verteilung hat eine unter diskreten Verteilungen einzigartige Eigenschaft, die als *Gedächtnislosigkeit* bezeichnet wird. Nehmen wir beispielsweise an, wir wollen beim (fairen) Würfeln eine „sechs“ würfeln. Wenn wir nach 100-maligen würfeln keine „sechs“ gesehen

haben (die Wahrscheinlichkeit dafür ist zwar winzig $(5/6)^{100}$, aber positiv) würden wir dann erwarten, dass die „sechs“ bald kommen sollte? Die Antwort ist natürlich: „Nein!“. Mit den Gesetzen der großen Zahlen kann man hier natürlich nicht argumentieren. Die besagen ja nur, dass *asymptotisch* die Frequenz der „sechs“ $\frac{1}{6}$ beträgt.

Satz 3.12 (Gedächtnislosigkeit der geometrischen Verteilung)**.** *Für eine diskrete Zufallsvariable Z sind folgende Aussagen äquivalent:*

(i) Z ist geometrisch verteilt;

(ii) Die Verteilung von Z ist eine gedächtnislose Verteilung *auf $\mathbb{N}$, d.h. es gilt*

$$\mathrm{P}(Z > k+m \mid Z > m) = \mathrm{P}(Z > k), \quad k, m \in \mathbb{N}. \tag{3.11}$$

Beweis. Der Beweis von „(i) $\Rightarrow$ (ii)“ ist einfaches Nachrechnen. Ist $Z \sim \mathrm{Geo}_p$, so gilt

$$\begin{aligned}\mathrm{P}(Z > k+m \mid Z > m) &= \frac{\mathrm{P}(Z > k+m, Z > k)}{\mathrm{P}(Z > m)} \\ &= \frac{\mathrm{P}(Z > k+m)}{\mathrm{P}(Z > m)} = \frac{(1-p)^{k+m}}{(1-p)^m} = (1-p)^k = \mathrm{P}(Z > k).\end{aligned}$$

Für den Beweis von „(ii) $\Rightarrow$ (i)“ bemerken wir zuerst, dass wir mit der obigen Rechnung die folgende Gleichung bekommen

$$\mathrm{P}(Z > k+m) = \mathrm{P}(Z > k)\,\mathrm{P}(Z > m), \quad k, m \in \mathbb{N}. \tag{3.12}$$

Zu zeigen ist $\mathrm{P}(Z > k) = (1-p)^k$, $k \in \mathbb{N}$ für ein p. Dann ist nämlich die Verteilungsfunktion und damit auch die Verteilung eindeutig bestimmt.

Wir setzen $p = \mathrm{P}(Z = 1)$, dann ist $\mathrm{P}(Z > 1) = 1 - p$ und mit (3.12) erhalten wir induktiv

$$\mathrm{P}(Z > 2) = \mathrm{P}(Z > 1)\,\mathrm{P}(Z > 1) = (1-p)^2,$$

und allgemein

$$\mathrm{P}(Z > k) = \mathrm{P}(Z > k-1)\,\mathrm{P}(Z > 1) = (1-p)^k, \quad k \in \mathbb{N}.$$

□

3.4.2 Negative Binomialverteilung

Für $r \in \mathbb{N}$ bezeichnen wir mit T_r die Wartezeit bis zum r-ten „Erfolg“ (bis zur r-ten „Eins“) in einer unabhängigen Folge $X_1, X_2, \ldots$ von Ber_p-verteilten Zufallsvariablen. Im folgenden Beispiel

$$0, 0, 1, 1, 0, 0, 0, 0, 1, 0, \ldots$$

ist $T_1 = 3, T_2 = 4, T_3 = 9$. Formal definieren wir rekursiv

$$T_1 = \inf\{n \geq 1 : X_n = 1\},$$
$$T_{r+1} = \inf\{n > T_r : X_n = 1\}, \quad r \geq 1.$$

Im vorherigen Unterabschnitt haben wir gesehen, dass T_1 geometrisch verteilt ist. Hier beschäftigen wir uns mit der Verteilung von T_r für allgemeines $r \in \mathbb{N}$. Es ist klar, dass sich die Verteilung von T_r auf $\{r, r+1, \dots\}$ konzentrieren sollte, denn für r Erfolge braucht man mindestens r Versuche. Die Wahrscheinlichkeitsgewichte der Verteilung von T_r sind im folgenden Resultat angegeben.

Satz 3.13. *Für $r \in \mathbb{N}$ gilt*

$$\mathrm{P}(T_r = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, \dots. \tag{3.13}$$

Beweis. Sei $r \in \mathbb{N}$ und sei $k \geq r$. Damit $\{T_r = k\}$ eintritt, muss der k-te Versuch ein „Erfolg" sein (Ereignis A) und bei den ersten $k-1$ Versuchen müssen genau $r-1$ „Erfolge" eingetreten sein (Ereignis B). Es gilt also

$$\begin{aligned}\mathrm{P}(T_r = k) &= \mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B) \\ &= p \cdot \binom{k-1}{r-1} p^{r-1} (1-p)^{(k-1)-(r-1)} = \binom{k-1}{r-1} p^r (1-p)^{k-r}.\end{aligned}$$

□

Definition 3.14. Ein Zufallsvariable mit Verteilung auf $\{r, r+1, \dots\}$ und Wahrscheinlichkeitsgewichten wie in (3.13), heißt *negativ binomialverteilt* mit Parametern r und p. Wir bezeichnen die Verteilung mit $\mathrm{NBin}_{r,p}$.

Natürlich ist die geometrische Verteilung ein Spezialfall der negativen Binomialverteilung mit $\mathrm{Geo}_p = \mathrm{NBin}_{1,p}$. Allgemeiner gilt der folgende Satz, der intuitiv sollte nach Definition der T_r's klar sein sollte.

Satz 3.15. *Sind $Z_1, Z_2, \dots$ unabhängige Zufallsvariablen mit $Z_i \sim \mathrm{Geo}_p$ dann gilt*

$$Z_1 + \cdots + Z_r \sim \mathrm{NBin}_{r,p}\,.$$

Für $T_r \sim \mathrm{NBin}_{r,p}$ gilt insbesondere

$$\mathrm{E}[T_r] = \frac{r}{p} \quad \textit{und} \quad \mathrm{Var}[T_r] = \frac{r}{p}\Big(\frac{1}{p} - 1\Big).$$

Beweis. Übung! □

Bemerkung 3.16. Der Name „negative Binomialverteilung“ wird verständlich, wenn man

$$\sum_{k=r}^{\infty} \binom{k-1}{r-1} p^r (1-p)^{k-r} = 1 \tag{3.14}$$

nachrechnet, was nach Satz 3.13 klar sein sollte.

Dafür benutzen wir die Formel für die Binomische Reihe

$$(1+x)^\alpha = \sum_{\ell=0}^{\infty} \binom{\alpha}{\ell} x^\ell, \quad |x| < 1, \alpha \in \mathbb{R}, \tag{3.15}$$

wobei (der verallgemeinerte) Binomialkoeffizient wie folgt definiert ist

$$\binom{\alpha}{\ell} = \frac{\alpha(\alpha-1)\cdots(\alpha-\ell+1)}{\ell!}, \quad \alpha \in \mathbb{R}, \ell \in \mathbb{N}_0.$$

Nehmen wir in (3.15) $\alpha = -r$ (das ist der Grund für den Namen der Verteilung) und $x = -p$, dann gilt

$$\begin{aligned}
(1-p)^{-r} &= \sum_{\ell=0}^{\infty} \binom{-r}{\ell} (-p)^\ell \\
&= \sum_{\ell=0}^{\infty} \frac{-r(-r-1)\cdots(-r-\ell+1)}{\ell!} (-1)^\ell p^\ell \\
&= \sum_{\ell=0}^{\infty} \frac{r(r+1)\cdots(r+\ell-1)}{\ell!} p^\ell \\
&= \sum_{\ell=0}^{\infty} \binom{r+\ell-1}{\ell} p^\ell.
\end{aligned}$$

Mit der Indexverschiebung $k = \ell + r$ erhalten wir

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

was (3.14) zeigt.

4 Transformationen und Faltung von Verteilungen

In diesem Kapitel behandeln wir Transformationen von Verteilungen und von gemeinsamen Verteilungen von Zufallsvariablen. In Definition 1.13 haben wir schon von gemeinsamen Verteilungen von Zufallsvariablen gesprochen. Hier schauen wir uns diskrete Zufallsvariablen und Zufallsvariablen mit Dichten etwas genauer an. Wir lernen Techniken kennen mit denen man Verteilungen von Transformationen von Zufallsvariablen und Zufallsvektoren berechnen kann. Nebenbei stellen wir neue Verteilungen vor und diskutieren Zusammenhänge zwischen einigen Verteilungen.

4.1 Einfache Transformationen von Verteilungen

Ist X eine diskrete Zufallsvariable oder ein Zufallsvektor mit Werten in S und $h : S \to R$ eine Abbildung, so ist $Y = h(X)$ eine R-wertige Zufallsvariable und mit Satz 3.1 können wir die Verteilung von Y bestimmen: Für alle $r \in R$ ist

$$\mathrm{P}(Y = r) = \mathrm{P}(h(X) = r) = \mathrm{P}(X \in h^{-1}(r)) = \sum_{s \in h^{-1}(r)} \mathrm{P}(X = s).$$

Ist X eine reellwertige Zufallsvariable mit Verteilungsfunktion F_X und $h : \mathbb{R} \to \mathbb{R}$ eine Funktion, dann gilt für die Verteilungsfunktion F_Y von $Y = h(X)$

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(h(X) \leq y) = \mathrm{P}(X \in h^{-1}((-\infty, y])). \tag{4.1}$$

Um die Verteilungsfunktion von Y angeben zu können muss man Wahrscheinlichkeiten von Urbildern $h^{-1}((-\infty, y])$ angeben können. Je nachdem wie kompliziert die Funktion h ist, kann es schwierig sein. Besonders einfach ist es wenn die Funktion h streng monoton ist.

Satz 4.1 (Transformation von Verteilungsfunktionen und Dichten)**.** *Es sei X eine Zufallsvariable mit Verteilungsfunktion F_X und sei $h : \mathbb{R} \to (a, b)$, $-\infty \leq a < b \leq \infty$ eine Funktion. Dann gelten folgende Aussagen.*

(i) Ist h streng monoton wachsend, so gilt für die Verteilungsfunktion F_Y von $Y = h(X)$

$$F_Y(y) = \begin{cases} 0 & : y \leq a, \\ F(h^{-1}(y)) & : y \in (a, b), \\ 1 & : y \geq b. \end{cases} \tag{4.2}$$

(ii) Besitzt X die Dichte f_X und ist h überall differenzierbar mit nicht verschwindender Ableitung, so ist die Dichte von Y gegeben durch

$$f_Y(y) = \begin{cases} \frac{f_X(h^{-1}(y))}{|h'(h^{-1}(y))|} & : y \in (a,b), \\ 0 & : y \notin (a,b). \end{cases} \tag{4.3}$$

Beweis. Mit dem Ansatz aus (4.1) erhalten wir für $y \in (a,b)$

$$F_Y(y) = \mathrm{P}(h(X) \le y) = \mathrm{P}(X \le h^{-1}(y)) = F_X(h^{-1}(y)),$$

womit man leicht (i) erhält.

Die in (4.3) angegebene Darstellung der Dichte f_Y erhält man durch Differenzieren von (4.2), wobei man eine Fallunterscheidung für wachsende h und fallende h machen muss. Im letzteren Fall ist $-h$ wachsend. □

Beispiel 4.2 (Lineare Transformationen)**.** Ist X eine Zufallsvariable mit Dichte f_X und ist $Y = aX + b$ mit $a \neq 0$, dann ist die Dichte von Y gegeben durch

$$f_Y(y) = \frac{1}{|a|} f_X\Big(\frac{y-b}{a}\Big), \quad y \in \mathbb{R}.$$

Übung 4.3. *Es sei $X \sim \mathcal{N}_{\mu,\sigma^2}$ für $\mu \in \mathbb{R}$ und $\sigma^2 > 0$. Dann ist $(X-\mu)/\sigma \sim \mathcal{N}_{0,1}$.*

Mit Hilfe des Satzes 4.1 kann man auch für stückweise streng monotone Transformationen Dichten bestimmen. Wir demonstrieren das im folgenden Beispiel.

Beispiel 4.4. Ist X eine nichtnegative Zufallsvariable mit Dichte f_X und ist $Y = X^a$ für ein $a \neq 0$, so ist die Dichte von Y gegeben durch

$$f_Y(y) = \frac{1}{|a|} y^{(1-a)/a} f_X(y^{1/a}).$$

Nimmt X auch negative Werte an und gibt es Intervalle auf denen h streng monoton wachsend und streng monoton fallend, so kann man die Dichten der transformierten Zufallsvariablen mit Fallunterscheidungen und Satz 4.1 berechnen. Es ist aber oft leichter (und man muss sich weniger Formeln merken!), die entsprechende Verteilungsfunktion zu berechnen und diese dann abzuleiten.

Ist z.B. $h(x) = |x|^a$ für ein $a > 0$, so gilt für die Verteilungsfunktion F_Y von $Y = h(X)$, $F_Y(y) = 0$ für $y \le 0$ und für $y \ge 0$

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \le y) = \mathrm{P}(|X|^a \le y) = \mathrm{P}(|X| \le y^{1/a}) \\ &= \mathrm{P}(-y^{1/a} \le X \le y^{1/a}) = F_X(y^{1/a}) - F_X(-y^{1/a}). \end{aligned}$$

Es folgt

$$f_Y(y) = F_Y'(y) = \frac{1}{a} y^{(1-a)/a} (f_X(y^{1/a}) + f_X(-y^{1/a})).$$

Die Formel vereinfacht sich weiter, wenn die Dichte von X eine gerade (um 0 symmetrische) Funktion ist. Dann ist

$$f_Y(y) = \frac{2}{a} y^{(1-a)/a} f_X(y^{1/a}), \quad y \geq 0.$$

Ist $X \sim \mathcal{N}_{0,\sigma^2}$ für ein $\sigma > 0$, so ist

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{-\frac{x^2}{2\sigma^2}\Big\}.$$

Die Dichte von X^2 ist dann gegeben durch

$$f_{X^2}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} y^{-1/2} \exp\Big\{-\frac{y}{2\sigma^2}\Big\}. \tag{4.4}$$

Definition 4.5 (Gamma-Verteilung)**.** Für $\theta, r > 0$ ist heißt die Verteilung auf $[0, \infty)$ mit der Dichte

$$f(x) = \frac{1}{\Gamma(r)} \theta^r x^{r-1} e^{-\theta x}, \quad x \geq 0 \tag{4.5}$$

die *Gamma-Verteilung* mit Parametern θ und r; Bezeichnung $X \sim \Gamma_{\theta,r}$. Dabei ist Γ die Gamma-Funktion.

Bemerkung 4.6. Beachten Sie, dass in der Literatur oft $\theta' = 1/\theta$ statt θ als Parameter der Gamma-Verteilung verwendet wird.

Übung 4.7. *Es sei $X \sim \Gamma_{\theta,r}$. Zeigen Sie*

$$\mathrm{E}[X] = \frac{r}{\theta} \quad \textit{und} \quad \mathrm{Var}[X] = \frac{r}{\theta^2}.$$

Bemerkung 4.8 (Zusammenhang von Gamma- und Normalverteilung)**.** Es ist $\Gamma(1/2) = \sqrt{\pi}$. Damit können wir die Dichte f_{X^2} aus (4.4) wie folgt umschreiben

$$f_{X^2}(y) = \frac{1}{\Gamma(1/2)} \Big(\frac{1}{2\sigma^2}\Big)^{1/2} y^{1/2-1} \exp\Big\{-\frac{1}{2\sigma^2} y\Big\}.$$

Also ist für $X \sim \mathcal{N}_{0,\sigma^2}$ die transformierte Zufallsvariable X^2 Gamma-verteilt mit Parametern $r = 1/2$ und $\theta = 1/(2\sigma^2)$.

4.2 Gemeinsame Verteilung von Zufallsvariablen und Faltung

In diesem Abschnitt beschäftigen wir uns mit Verteilungen von Summen von Zufallsvariablen. Bei unabhängigen Zufallsvariablen spricht man von einer *Faltung von Verteilungen.* Wir beginnen mit dem einfachen und anschaulichen Fall von diskreten Zufallsvariablen. Seien also X und

Y diskrete Zufallsvariablen mit Werten in S bzw. R definiert auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathrm{P})$. Die gemeinsame Verteilung von (X, Y) ist eindeutig bestimmt durch

$$\mathrm{P}(X = s, Y = r), \quad s, r \in S \times R.$$

Für die Randverteilungen, also für die Verteilungen von X und Y gilt jeweils

$$\mathrm{P}(X = s) = \sum_{r \in R} \mathrm{P}(X = s, Y = r), \quad s \in S,$$

und

$$\mathrm{P}(Y = r) = \sum_{s \in S} \mathrm{P}(X = s, Y = r), \quad r \in R.$$

Schreiben wir $\{X = s\}$ als disjunkte Vereinigung $\{X = s\} = \cup_{r \in R}\{X = s, Y = r\}$, so sehen wir, dass hinter diesen Formeln natürlich die σ-Additivität des Wahrscheinlichkeitsmaßes P steckt; vgl. Definition 1.2.2.(ii).

Sind X und Y unabhängig, dann ist die gemeinsame Verteilung durch die Randverteilungen eindeutig bestimmt, denn dann ist

$$\mathrm{P}(X = s, Y = r) = \mathrm{P}(X = s)\,\mathrm{P}(Y = r) \quad s, r \in S \times R.$$

Wie das folgende einfache Beispiel zeigt, bestimmen umgekehrt die Randverteilungen im Allgemeinen aber nicht die gemeinsame Verteilung.

Beispiel 4.9. Wir betrachten ein Würfelexperiment mit zwei fairen Würfeln. Seien X_1 und X_2 jeweils die Augenzahl des ersten und des zweiten Würfels. Dann sind X_1 und X_2 unabhängige Zufallsvariablen mit Werten in $S = \{1, \ldots, 6\}$. Sei $Y = X_1 + X_2 \in R = \{2, \ldots, 12\}$. Die folgende Tabelle zeigt die gemeinsame Verteilung von X_1 und Y, sowie die Randverteilungen.

$S \setminus R$	2	3	4	5	6	7	8	9	10	11	12	$\mathrm{P}(X_1 = s)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	0	$\frac{1}{6}$
2	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	$\frac{1}{6}$
3	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	$\frac{1}{6}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	$\frac{1}{6}$
5	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	$\frac{1}{6}$
6	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
$\mathrm{P}(Y = r)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

Natürlich sind X_1 und Y nicht unabhängig, denn es ist z.B.

$$\mathrm{P}(X_1 = 1, Y = 12) = 0 \neq \frac{1}{6} \cdot \frac{1}{36} = \mathrm{P}(X_1 = 1) \cdot \mathrm{P}(Y = 12).$$

Definition 4.10 (Faltung von Verteilungen)**.** Es seien X und Y unabhängige Zufallsvariablen mit Verteilungen μ_X und μ_Y. Die Verteilung von $X + Y$ heißt *Faltung von μ_X und μ_Y* und wird mit $\mu_X * \mu_Y$ bezeichnet, d.h.

$$\mu_{X+Y} = \mu_X * \mu_Y. \tag{4.6}$$

Aus der Kommutativität und Assoziativität der Addition folgt für unabhängige Zufallsvariablen X, Y, Z mit Verteilungen μ_X, μ_Y bzw. μ_Z sofort

$$\mu_X * \mu_Y = \mu_Y * \mu_X \quad \text{und} \quad \mu_X * (\mu_Y * \mu_Z) = (\mu_X * \mu_Y) * \mu_Z.$$

In den folgenden zwei Sätzen demonstrieren wir wie man Faltungen von Verteilungen von diskreten und stetigen Zufallsvariablen berechnen kann.

Satz 4.11 (Faltung diskreter Verteilungen)**.** *Seien X und Y unabhängige diskrete Zufallsvariablen mit Werten in $\mathbb{N}_0$. Dann $X + Y$ eine $\mathbb{N}_0$-wertige Zufallsvariable und es gilt*

$$\mathrm{P}(X + Y = k) = \sum_{j=0}^{k} \mathrm{P}(X = k - j)\,\mathrm{P}(Y = j) = \sum_{i=0}^{k} \mathrm{P}(X = i)\,\mathrm{P}(Y = k - i). \tag{4.7}$$

Beweis. Es gilt

$$\begin{aligned}\mathrm{P}(X + Y = k) &= \sum_{j=0}^{\infty} \mathrm{P}(X + Y = k, Y = j) = \sum_{j=0}^{\infty} \mathrm{P}(X + j = k, Y = j)\\ &= \sum_{j=0}^{k} \mathrm{P}(X = k - j)\,\mathrm{P}(Y = j),\end{aligned}$$

wobei wir im letzten Schritt die Unabhängigkeit und $\mathrm{P}(X = \ell) = 0$ für $\ell < 0$ ausgenutzt haben. □

Satz 4.12 (Faltung von Zufallsvariablen mit Dichten)**.** *Seien X und Y unabhängige Zufallsvariablen mit Dichten f_X, bzw. f_Y. Dann ist die Dichte von $X + Y$ gegeben durch*

$$f_X * f_Y(z) \coloneqq f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y)\,\mathrm{d}y. \tag{4.8}$$

Beweis. Nach Voraussetzung ist die gemeinsame Dichte von X und Y gegeben durch $f(x, y) = f_X(x) f_Y(y)$. Für die Verteilungsfunktion von $X + Y$ gilt

$$\begin{aligned}\mathrm{P}(X + Y \le z) &= \int_{\{x+y\le z\}} f(x, y)\,\mathrm{d}(x, y) = \int_{\{x+y\le z\}} f_X(x) f_Y(y)\,\mathrm{d}(x, y)\\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_X(x)\,\mathrm{d}x f_Y(y)\,\mathrm{d}y.\end{aligned}$$

Ableiten nach z liefert die Behauptung. □

Wir nennen eine Klasse von Verteilungen *stabil* wenn sie abgeschlossen bezüglich Faltung ist. Wir haben bereits in den Übungen gesehen, dass die Poissonverteilung stabil ist. Bei Verteilungen, die von mehreren Parametern abhängen, wie z.B. Binomial- oder Gamma-Verteilung kommt es vor, dass die Verteilung bezüglich des einen Parameters stabil ist, aber nicht bezüglich des anderen. Das folgende Resultat zeigt die Stabilität der Normalverteilung (in beiden Parametern). Weitere Beispiele stabiler Klassen behandeln wir in Übungen und Beispielen am Ende dieses Abschnittes.

Satz 4.13 (Abgeschlossenheit der Normalverteilung bezüglich Faltung)**.** *Seien X_1 und X_2 unabhängige Zufallsvariablen mit $X_i \sim \mathcal{N}_{\mu_i,\sigma_i^2}$, für $\mu_i \in \mathbb{R}$, $\sigma_i^2 > 0$, $i = 1, 2$. Dann gilt*

$$X_1 + X_2 \sim \mathcal{N}_{\mu_1+\mu_2,\sigma_1^2+\sigma_2^2}.$$

Beweis. Es gilt (vgl. Beispiel 4.2 und Übung 4.3) $X_i - \mu_i \sim \mathcal{N}_{0,\sigma_i^2}$. Für $Z = X_1 + X_2 - \mu_1 - \mu_2$ erhalten wir mit der Faltungsformel (4.8)

$$f_Z(z) = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp\Big\{-\frac{(z-y)^2}{2\sigma_1^2} - \frac{y^2}{2\sigma_2^2}\Big\}\,\mathrm{d}y. \tag{4.9}$$

Zu zeigen ist nun,

$$f_Z(z) = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\Big\{-\frac{z^2}{2(\sigma_1^2+\sigma_2^2)}\Big\}. \tag{4.10}$$

Es gilt

$$\begin{aligned}
\frac{(z-y)^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} &= \frac{z^2\sigma_2^2 + y^2(\sigma_1^2+\sigma_2^2) - 2yz\sigma_2^2}{\sigma_1^2\sigma_2^2} \\
&= \frac{z^2\Big(\sigma_2^2 - \frac{\sigma_2^4}{\sigma_1^2+\sigma_2^2}\Big) + (\sigma_1^2+\sigma_2^2)\Big(y - \frac{z\sigma_2^2}{(\sigma_1^2+\sigma_2^2)}\Big)^2}{\sigma_1^2\sigma_2^2} \\
&= \frac{z^2}{\sigma_1^2+\sigma_2^2} + \frac{\Big(y - \frac{z\sigma_2^2}{(\sigma_1^2+\sigma_2^2)}\Big)^2}{\sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2)}.
\end{aligned}$$

Einsetzen in (4.9) liefert nach elementaren Umformungen

$$\begin{aligned}
f_Z(z) = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} &\exp\Big\{-\frac{z^2}{2(\sigma_1^2+\sigma_2^2)}\Big\} \\
&\cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2)}} \exp\Big\{-\frac{\Big(y - \frac{z\sigma_2^2}{(\sigma_1^2+\sigma_2^2)}\Big)^2}{2\sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2)}\Big\}\,\mathrm{d}y.
\end{aligned}$$

Bei genauem Hinsehen erkennt man in dem Integrand in der letzten Zeile die Dichte einer Normalverteilung. Also ist das Integral gleich 1, was (4.10) zeigt und den Beweis abschließt. □

Korollar 4.14. *Seien $X_1, \dots, X_n$ unabhängig identisch verteilt mit $X_i \sim \mathcal{N}_{0,\sigma^2}$ für ein $\sigma^2 > 0$. Dann gilt*

$$\frac{1}{\sqrt{n\sigma^2}}(X_1 + \cdots + X_n) \sim \mathcal{N}_{0,1}.$$

Beweis. Übung! □

Bemerkung 4.15. Korollar 4.14 kann als ein Spezialfall des zentralen Grenzwertsatzes angesehen werden. Es besagt, dass für unabhängige und identisch verteilte Zufallsvariablen $X_1, X_2, \dots$ (deren Verteilung bestimmte Voraussetzungen erfüllt) die Verteilungsfunktion von $\frac{1}{\sqrt{n \operatorname{Var}[X_1]}}(X_1 + \cdots + X_n)$ für $n \to \infty$ gegen die Verteilungsfunktion der $\mathcal{N}_{0,1}$-Verteilung konvergiert. Im Korollar 4.14 besteht Gleichheit für jedes $n \in \mathbb{N}$.

Übung 4.16 (Stabilität der Cauchy Verteilung)**.** *Die Dichte der Cauchy Verteilung mit Parameter $u > 0$ ist gegeben durch*

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

Es seien $X_1, \dots, X_n$ unabhängig und identisch verteilte Zufallsvariablen mit Dichte c_u. Zeigen Sie, dass die Zufallsvariable $(X_1 + \cdots + X_n)/n$ dieselbe Dichte hat.

Hinweis: *Sie können ohne Beweis verwenden, dass $c_u * c_v = c_{u+v}$ für alle $u, v > 0$ gilt.*

Bemerkung 4.17. Mit Übung 4.16 sieht man, dass für Folgen von Cauchy verteilten Zufallsvariablen die Gesetze der großen Zahlen nicht gelten können. Die Voraussetzungen der Sätze 2.14 und 2.15 sind nicht erfüllt, weil die Cauchy Verteilung keinen Erwartungswert besitzt, denn es ist

$$\int_0^\infty x c_u(x)\, \mathrm{d}x = \infty \quad \text{und} \quad \int_{-\infty}^0 x c_u(x)\, \mathrm{d}x = -\infty.$$

Beispiel 4.18 (Stabilität der Binomialverteilung)**.** Sind X_1 und X_2 unabhängig mit $X_i \sim \mathrm{Bin}_{n_i,p}$, dann ist $X_1 + X_2 \sim \mathrm{Bin}_{n_1+n_2,p}$. Das kann man zwar mit der Faltungsformelformel (4.7) nachrechnen, es sollte aber wegen der Interpretation der Binomialverteilung klar sein.

Beispiel 4.19 (Exponentialverteilung und Gammaverteilung)**.** Exponentialverteilung ist eines von vielen Beispielen für eine nicht stabile Klasse von Verteilungen.

Man kann nachrechnen, dass Faltung von r Exponentialverteilungen mit demselben Parameter $\lambda > 0$ die Gammaverteilung $\Gamma_{\lambda,r}$ ist. Insbesondere gilt $\Gamma_{\lambda,r_1} * \Gamma_{\lambda,r_2} = \Gamma_{\lambda,r_1+r_2}$ für $r_1, r_2 \in \mathbb{N}$.

Durch die Beispiele und Übungen in diesem Abschnitt bekommt man hoffentlich einen guten Eindruck, dass das Rechnen mit Faltungen nicht unbedingt leicht ist. Außer Verteilungsfunktionen und Dichten gibt es einige andere Klassen von Funktionen, die die Verteilungen eindeutig bestimmen und viel besser fürs Arbeiten mit Summen von Zufallsvariablen geeignet sind. Das sind z.B. *(wahrscheinlichkeits)erzeugende Funktionen*, *momentenerzeugende Funktionen* und *charakteristische Funktionen.*

Als Abschluss des Kapitels geben wir hier (ohne Beweis) eine Verallgemeinerung von Satz 4.1 auf mehrdimensionalen Fall an.

Satz 4.20 (Mehrdimensionaler Transformationssatz). *Sei $X = (X_1, \dots, X_k)$ ein k-dimensionaler Zufallsvektor mit Dichte f_X und sei $Y = g(X)$, wobei $g : \mathbb{R}^k \to \mathbb{R}^k$ eine Funktion für die es disjunkte Mengen $A_1, \dots, A_m \subset \mathbb{R}^k$ gibt, so dass $\mathbb{R}^k \setminus (A_1 \cup \dots \cup A_m)$ eine Lebesgue-Nullmenge[1] ist und auf jedem A_j ist die Funktion g bijektiv mit nicht verschwindender Jacobi-Matrix, d.h. $\det(\partial g(x)/\partial x) \neq 0$ auf A_j, $j = 1, \dots, m$.*

Dann ist die Dichte von Y gegeben durch

$$f_Y(y) = \sum_{j=1}^{m} |\det(\partial h_j(x)/\partial x)| \cdot f_X(h_j(x)), \tag{4.11}$$

wobei h_j die Inverse von g auf A_j, $j = 1, \dots, m$ ist.

Beispiel 4.21. Sei $X = (X_1, X_2)$ ein Zufallsvektor mit gemeinsamer Dichte f_X und sei $g : \mathbb{R}^2 \to \mathbb{R}^2$ definiert durch $g(x) = (x_1, x_1 + x_2)$. Die Inverse von g ist $h((x_1, y)) = (x_1, y - x_1)$, $x_1, y \in \mathbb{R}$ und es gilt $\det(\partial g(x)/\partial x) \equiv 1 \equiv \det(\partial h((x_1, y))/\partial(x_1, y))$. Mit Satz 4.20 gilt

$$f_{g(X)}(x_1, y) = f_X(x_1, y - x_1). \tag{4.12}$$

Wenn X_1 und X_2 unabhängig sind mit Dichten f_{X_1} und f_{X_2}, dann folgt

$$f_{g(X)}(x_1, y) = f_{X_1}(x_1) f_{X_2}(y - x_1). \tag{4.13}$$

Die Rand-Dichte von $Y = X_1 + X_2$ ist im allgemeinen Fall gegeben durch

$$f_Y(y) = \int_{-\infty}^{\infty} f_X(x_1, y - x_1)\, \mathrm{d}x_1, \tag{4.14}$$

und wenn X_1 und X_2 unabhängig sind, dann folgt

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(x_1) f_{X_2}(y - x_1)\, \mathrm{d}x_1 \tag{4.15}$$

und wir erkennen die Faltungsformel (4.8) wieder.

[1] Sie können hier an endliche oder abzählbar unendliche Punktemengen oder allgemeiner an unterdimensionale Mengen denken, also z.B. Geraden in $\mathbb{R}^2$, Flächen in $\mathbb{R}^3$ und abzählbare Vereinigungen davon.

5 Bedingte Verteilungen

Wir haben schon in Abschnitt 1.5 bedingte Wahrscheinlichkeiten von Ereignissen wiederholt. In diesem Kapitel behandeln wir bedingte Verteilungen. Am besten stellt man sich mehrere Zufallsvariablen (oder ein mehrstufiges Experiment) auf einem gemeinsamen Wahrscheinlichkeitsraum vor bei denen wir nach und nach die Informationen über die einzelnen Realisierungen bekommen und Rückschlüsse auf nachfolgenden Realisierungen machen wollen.

Wir behandeln wieder getrennt voneinander die Fälle von diskreten Zufallsvariablen und von Zufallsvariablen mit Dichten. Es sind natürlich auch Mischfälle möglich. Man denke z.B. an ein zweistufiges Experiment, bei dem zuerst unabhängige Zufallsvariablen $U \sim \mathcal{U}_{[0,1]}$ und $N \sim \mathrm{Poi}_\lambda$ realisiert werden und dann eine Zufallsvariable $X \sim \mathrm{Bin}_{N,U}$.

5.1 Bedingte Verteilungen diskreter Zufallsvariablen

Definition 5.1. Seien X_1 und X_2 diskrete Zufallsvariablen mit Wertebereichen S_1 bzw. S_2. Für $a_1 \in S_1$ und $A_2 \subset S_2$ ist die bedingte Wahrscheinlichkeit von $\{X_2 \in A_2\}$ gegeben $\{X_1 = a_1\}$ definiert durch

$$\mathrm{P}(X_2 \in A_2 \mid X_1 = a_1) = \frac{\mathrm{P}(X_1 = a_1, X_2 \in A_2)}{\mathrm{P}(X_1 = a_1)}, \tag{5.1}$$

wobei wir im Fall $\mathrm{P}(X_1 = a_1) = 0$ die rechte Seite gleich 0 setzen. Die durch (5.1) definierte Verteilung $\mathrm{P}(X_2 \in \cdot \mid X_1 = a_1)$ heißt *bedingte Verteilung von X_2 gegeben* $\{X_1 = a_1\}$.

Im Fall $\mathrm{P}(X_1 = a_1) > 0$ hat die bedingte Verteilung $\mathrm{P}(X_2 \in \cdot \mid X_1 = a_1)$ die Eigenschaften „gewöhnlicher Verteilungen“. Wir haben schon an einigen Stellen betont, dass Randverteilungen von Zufallsvariablen nicht die gemeinsame Verteilung bestimmen. Randverteilungen zusammen mit bedingten Verteilungen bestimmen aber die gemeinsame Verteilung. Für diskrete Zufallsvariablen X_1 und X_2 wie in Definition 5.1 folgt aus (5.1)

$$\mathrm{P}(X_1 = a_1, X_2 = a_2) = \mathrm{P}(X_1 = a_1)\,\mathrm{P}(X_2 = a_2 \mid X_1 = a_1), \quad a_1 \in S_1, a_2 \in S_2.$$

Man kann hier die rechte Seite als Produkt von Wahrscheinlichkeiten in einem zweistufigen Experiment interpretieren: Erst wird X_1 realisiert und dann, bedingt auf X_1 wird X_2 realisiert. Wir werden uns mit solchen Wahrscheinlichkeiten im Zusammenhang mit Markovketten beschäftigen.

Beispiel 5.2 (Verteilung von Wartezeiten: Nachfolger bedingt auf Vorgänger)**.** Betrachten wir die Wartezeiten aus Abschnitt 3.4.2. Ist eine Folge von unabhängigen Bernoulli verteilten Zufallsvariablen mit Parameter p gegeben, dann ist die Wartezeit T_r bis zum r-ten Erfolg negativ binomial verteilt mit Parametern r und p. Insbesondere ist die Verteilung von T_r auf $\{r, r+1, \dots\}$ konzentriert. Wissen wir aber, dass beispielsweise $T_{r-1} = r+10$ ist dann muss natürlich $T_r \geq r+11$ gelten. Die bedingte Verteilung von T_r gegeben $T_{r-1} = r+10$ ist also auf $\{r+11, r+12, \dots\}$ konzentriert.

Genauer gilt für $\ell \in \mathbb{N}$, $\ell \geq r-1$ und $k = \ell+1, \ell+2, \dots$

$$\begin{aligned}\mathrm{P}(T_r = k \mid T_{r-1} = \ell) &= \frac{\mathrm{P}(T_r = k, T_{r-1} = \ell)}{\mathrm{P}(T_{r-1} = \ell)} = \frac{\mathrm{P}(T_r - T_{r-1} = k-\ell, T_{r-1} = \ell)}{\mathrm{P}(T_{r-1} = \ell)} \\ &= \frac{\mathrm{P}(T_r - T_{r-1} = k-\ell)\,\mathrm{P}(T_{r-1} = \ell)}{\mathrm{P}(T_{r-1} = \ell)} \\ &= \mathrm{P}(T_r - T_{r-1} = k-\ell) = p(1-p)^{k-\ell-1}.\end{aligned}$$

Für die Gleichheit in der zweiten Zeile haben wir die Unabhängigkeit von $T_r - T_{r-1}$ und T_r ausgenutzt.

Bei der bedingten Verteilung von T_r gegeben $\{T_{r-1} = \ell\}$ handelt es sich also um eine um ℓ verschobene geometrische Verteilung.

Beispiel 5.3 (Verteilung von Wartezeiten: Vorgänger bedingt auf Nachfolger)**.** Wir qbetrachten eine ähnliche Situation wie im vorherigen Beispiel mit $r = 2$. Gesucht ist die Verteilung des ersten Erfolgs, bedingt auf die Realisierung des zweiten Erfolgs. Für $k \in \mathbb{N}$, $k \geq 2$ und $\ell = 1, \dots, k-1$ glt

$$\begin{aligned}\mathrm{P}(T_1 = \ell \mid T_2 = k) &= \frac{\mathrm{P}(T_1 = \ell, T_2 = k)}{\mathrm{P}(T_2 = k)} = \frac{\mathrm{P}(T_2 - T_1 = k-\ell, T_1 = \ell)}{\mathrm{P}(T_2 = k)} \\ &= \frac{\mathrm{P}(T_2 - T_1 = k-\ell) \cdot \mathrm{P}(T_1 = \ell)}{\mathrm{P}(T_2 = k)} \\ &= \frac{p(1-p)^{k-\ell-1} \cdot p(1-p)^{\ell-1}}{\binom{k-1}{2-1} p^2 (1-p)^{k-2}} = \frac{1}{k-1}.\end{aligned}$$

Also ist die bedingte Verteilung von T_1 gegeben $\{T_2 = k\}$ gegeben durch die Gleichverteilung auf $\{1, \dots, k-1\}$.

Man kann das Beispiel auch wie folgt interpretieren: Sind X_1 und X_2 unabhängige geometrisch verteilte Zufallsvariablen mit Parameter p, dann ist die bedingte Verteilung von X_1 gegeben $\{X_1 + X_2 = k\}$ gegeben durch die Gleichverteilung auf $\{1, \dots, k-1\}$.

Beispiel 5.4 (Poissonverteilung bedingt auf Summe)**.** Seien X_1 und X_2 unabhängig Poisson verteilt mit Parametern $\lambda_1 > 0$ und $\lambda_2 > 0$. In einer Übung haben wir gesehen, dass

$$\mathrm{P}(X_1 = k \mid X_1 + X_2 = n) = \binom{n}{k} p^k (1-p)^{n-k}$$

mit $p = \lambda_1/(\lambda_1 + \lambda_2)$ gilt. Also ist die bedingte Verteilung von X_1 bedingt auf $\{X_1 + X_2 = n\}$ gegeben durch die Binomialverteilung mit Parametern n und p.

5.2 Bedingte Dichten

Definition 5.5. Seien X_1 und X_2 Zufallsvariablen mit Dichten f_1 bzw. f_2 und der gemeinsamen Dichte $f : \mathbb{R} \to \mathbb{R}$. Die bedingte Dichte von X_2 gegeben $\{X_1 = a_1\}$ ist definiert durch

$$f_{X_2|X_1=a_1}(a_2) \coloneqq \frac{f(a_1, a_2)}{f_1(a_1)}, \tag{5.2}$$

wobei wir die rechte Seite gleich 0 setzen, wenn $f_1(a_1) = 0$.

Auch hier sieht man, dass die gemeinsame Dichte durch die Dichten der Randverteilungen zusammen mit bedingten Dichten bestimmt ist.

Wir geben ein Beispiel an das mit Beispiel 5.3 verwandt ist. Die „Verwandtschaft“ erklärt sich dadurch, dass man Exponentialverteilung durch Folgen von geometrischen Verteilungen approximieren kann. Viele Eigenschaften der geometrischen Verteilung im diskreten lassen sich auch für die Exponentialverteilung im stetigen Fall zeigen; vgl. Übung 5.7.

Beispiel 5.6. Es seien X_1 und X_2 unabhängige exponential verteilte Zufallsvariablen mit Parameter λ. Die gemeinsame Dichte f von $(X_1, X_1 + X_2)$ ist nach Beispiel 4.21 Gleichung (4.13) gegeben durch

$$f(x_1, y) = f_{X_1}(x_1) f_{X_2}(y - x_1) = \lambda^2 e^{-\lambda x_1 - \lambda(y-x_1)} = \lambda^2 e^{-\lambda y}, \quad 0 < x < y.$$

Die Bedingte Dichte von X_1 gegeben $X_1 + X_2 = y$ ist

$$f_{X_1|X_1+X_2=y}(x_1) = \frac{\lambda^2 e^{-\lambda y}}{\lambda^2 \int_0^y e^{-\lambda x_1} e^{-\lambda(y-x_1)} \,\mathrm{d}x_1} = \frac{e^{-\lambda y}}{e^{-\lambda y} \int_0^y 1 \,\mathrm{d}x_1} = \frac{1}{y}. \tag{5.3}$$

Also ist X_1 gegeben $X_1 + X_2 = y$ uniform verteilt auf $[0, y]$; vergleichen Sie das mit dem entsprechenden Resultat für geometrisch verteilte Zufallsvariablen in Beispiel 5.2.

Übung 5.7. *Es sei X eine exponential verteilte Zufallsvariable mit Parameter $\lambda > 0$.*

(i) Zeigen Sie

$$\mathrm{P}(X > s + t | X > t) = \mathrm{P}(X > s), \quad \textit{für alle } t, s > 0. \tag{5.4}$$

Diese Eigenschaft heißt Gedächtnislosigkeit.

(ii) Zeigen Sie: Exponentialverteilung ist die einzige absolut stetige Verteilung mit der Eigenschaft (5.4).

Hinweis zu (ii): *Überlegen Sie sich, dass $U(t) \coloneqq \mathrm{P}(X > t)$ die Funktionalgleichung $U(s + t) = U(s)U(t),\ s, t > 0$ mit $U(0) = 1$ löst. Wie sehen die Lösungen davon aus?*

6 Verteilungskonvergenz und zentraler Grenzwertsatz

Wir haben schon einige Konvergenzarten im Abschnitt 2.2 kennengelernt bzw. wiederholt. In diesem Kapitel diskutieren wir mit *Konvergenz in Verteilung* eine weitere Konvergenzart. Diese ist uns schon bei der Poissonapproximation der Binomialverteilung (Satz 3.9) begegnet, auch wenn wir dort noch nicht von Konvergenz in Verteilung gesprochen haben.

Das Hauptziel des Kapitels ist der Beweis einer Variante des zentralen Grenzwertsatzes. Die Aussage ist grob: Für eine große Klasse von Folgen von Zufallsvariablen (wir werden unabhängige und identisch verteilte Zufallsvariablen betrachten) lässt sich die (zentrierte und umskalierte) Summe der n ersten Folgenglieder durch eine normalverteilte Zufallsvariable approximieren.

6.1 Konvergenz in Verteilung

Seien X_n und X diskrete Zufallsvariablen auf $\mathbb{R}$ mit Verteilungen $P_n = \mathrm{P}_{X_n}$ und $P = \mathrm{P}_X$ und Verteilungsfunktionen F_n bzw. F. Nehmen wir an, dass

$$\lim_{n\to\infty} P_n(\{x\}) = P(\{x\}) \quad x \in \mathbb{R} \tag{6.1}$$

gilt. Natürlich muss $P\{x\}) = 0$ außerhalb einer diskreten Menge gelten. Dann kann man zeigen, dass auch

$$\lim_{n\to\infty} F_n(x) = F(x) \quad x \in \mathbb{R} \tag{6.2}$$

gilt. Ein Beispiel für so eine Situation ist die Poissonapproximation der Binomialverteilung (Satz 3.9). Dort hatten wir $\mu_n = \mathrm{Bin}_{n,np_n}$ und $\mu = \mathrm{Poi}_\lambda$, wobei $\lambda = \lim_{n\to\infty} np_n$.

Man würde gern von Verteilungskonvergenz sprechen wenn (6.1) oder (6.2) gilt. Dieser Forderungen sind aber etwas zu restriktiv wie das folgende Beispiel zeigt.

Beispiel 6.1. Seien $X_n = 1/n$, $\widehat{X}_n = -1/n$ deterministische Folge von Zufallsvariablen und sei $X = 0$. Die zugehörigen Verteilungen sind gegeben durch die Dirac-Maße $\delta_{1/n}$, $\delta_{-1/n}$ bzw. δ_0. Die Verteilungsfunktionen haben die Form $F_n(x) = \mathbb{1}_{[1/n,\infty)}(x)$, $\widehat{F}_n(x) = \mathbb{1}_{[-1/n,\infty)}$ und $F_0(x) = \mathbb{1}_{[0,\infty)}(x)$. Wegen $\lim_{n\to\infty} X_n = \lim_{n\to\infty} \widehat{X}_n = X$ sollte man zurecht erwarten, dass auch die Verteilungen von X_n und $\widehat{X}_n$ gegen die Verteilung von X konvergieren.

Für alle $x \neq 0$ gilt

$$\lim_{n\to\infty} \delta_{1/n}(\{x\}) = \lim_{n\to\infty} \delta_{-1/n}(\{x\}) = 0 = \delta_0(\{x\})$$

und

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} \widehat{F}_n(x) = F(x) = \begin{cases} 0 & : x < 0, \\ 1 & : x > 0. \end{cases} \tag{6.3}$$

Für $x = 0$ gilt aber

$$\lim_{n\to\infty} \delta_{1/n}(\{0\}) = \lim_{n\to\infty} \delta_{-1/n}(\{0\}) = 0 \neq 1 = \delta_0(\{x\})$$

und

$$\lim_{n\to\infty} F_n(0) = 0 \neq 1 = \lim_{n\to\infty} \widehat{F}_n(0) = F(0).$$

Im Punkt $x = 0$, also dem einzigen Unstetigkeitspunkt von F, sind (6.1) und (6.2) nicht erfüllt. Die Grenzwerte für X_n und $\widehat{X}_n$ sind sogar unterschiedlich.

Für eine Funktion $F : \mathbb{R} \to \mathbb{R}$ bezeichnen wir im Folgenden mit $\mathcal{C}(F)$ die Menge der Stetigkeitspunkte von F.

Definition 6.2 (Konvergenz in Verteilung)**.** Es seien $X_1, X_2, \ldots$ Zufallsvariablen auf $\mathbb{R}$ mit Verteilungen $P_1, P_2, \ldots$ und Verteilungsfunktionen $F_1, F_2, \ldots$. Die Folge der Zufallsvariablen X_n *konvergiert in Verteilung* gegen eine Zufallsvariable X mit Verteilungsfunktion F und Verteilung P, wenn

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \text{für alle} \quad x \in \mathcal{C}(F). \tag{6.4}$$

Man sagt dann, dass die Folge der Verteilungen P_n gegen die Grenzverteilung P *schwach konvergiert*; wir schreiben $P_n \Rightarrow P$. Entsprechend schreiben wir $X_n \Rightarrow X$, wenn X_n in Verteilung gegen X konvergiert. (Andere gebräuchliche Abkürzungen sind $X_n \xrightarrow{d} X$, d steht hier für *distribution*, also Verteilung, oder $X_n \xrightarrow{\mathcal{L}} X$, $\mathcal{L}$ steht hier für *law*, also Verteilungsgesetz).

Es gibt viele andere Äquivalente Charakterisierungen und Definitionen von Verteilungskonvergenz, auf die wir hier aber nicht eingehen können. Zumindest einige davon werden in den weiterführenden Wahrscheinlichkeitstheorie-Vorlesungen behandelt.

Mit Definition 6.2 konvergiert in Beispiel 6.1 sowohl die Folge X_n als auch die Folge $\widehat{X}_n$ in Verteilung gegen $X = 0$. Die Forderung (6.4) haben wir in (6.3) nachgerechnet.

Wir haben schon an einer anderen Stelle auf die Gemeinsamkeiten von Exponentialverteilung und geometrischer Verteilung hingewiesen und auch angedeutet, dass es mit Approximation der Exponentialverteilung durch Folgen von geometrischen Verteilungen zu tun hat. Im folgenden Satz präzisieren wir die Aussage.

Satz 6.3 (Exponentialapproximation)**.** *Es sei $X_1, X_2, \ldots$ eine Folge von Zufallsvariablen mit $X_n \sim \mathrm{Geo}_{p_n}$, $p_n \in (0,1)$ und $p_n \to 0$, d.h. $\mathrm{E}[X_n] = 1/p_n \to \infty$. Für $X \sim \mathrm{Exp}_1$ gilt*

$$\mathrm{P}(p_n X_n \leq x) \xrightarrow{n\to\infty} \mathrm{P}(X \leq x), \quad x > 0. \tag{6.5}$$

Insbesondere gilt $p_n X_n = \frac{X_n}{\mathrm{E}[X_n]} \Rightarrow X$ für $n \to \infty$.

Beweis. Wir haben nur (6.5) für Konvergenz in Verteilung zu zeigen, weil für $x < 0$ beide Seiten von (6.5) gleich 0 sind.

Sei also $x > 0$. Dann gilt (vgl. (3.9))

$$\mathrm{P}(p_n X_n \le x) = \mathrm{P}(X_n \le x/p_n) = 1 - (1-p_n)^{\lfloor x/p_n \rfloor}.$$

Für $n \to \infty$ gilt nach Voraussetzung $p_n \to 0$ und somit $(1-p_n)^{1/p_n} \to e^{-1}$. Mit $p_n = 1/n$ dürfte die Formel aus Analysis I bekannt sein. Für allgemeine $p_n \to 0$ sieht man sie wie folgt

$$\begin{aligned}(1-p_n)^{1/p_n} &= \exp\Big(\frac{1}{p_n}\log(1-p_n)\Big) = \exp\Big(\frac{1}{p_n}(-p_n - \frac{p_n^2}{2} + o(p_n^3))\Big) \\ &= \exp\Big(-1 - \frac{p_n}{2} + o(p_n^2)\Big) \xrightarrow{n\to\infty} e^{-1}.\end{aligned}$$

Damit erhalten wir (vgl. (1.7))

$$\mathrm{P}(p_n X_n \le x) \xrightarrow{n\to\infty} 1 - e^{-x} = \mathrm{P}(X \le x).$$

□

Übung 6.4. *Seien $X_1, \dots, X_n$ unabhängige und exponential verteilte Zufallsvariablen mit Parameter $\lambda > 0$ und sei $Z_n := \max_{1\le i\le n} X_i$. Zeigen Sie, dass die Folge $Z_n - \frac{\ln n}{\lambda}$ in Verteilung gegen eine doppelexponentialverteilte Zufallsvariable Z mit Parameter λ konvergiert. Dabei ist die Verteilungsfunktion einer solchen doppelexponentialverteilten Zufallsvariablen gegeben durch $F(x) = e^{-e^{-\lambda x}}$, $x \in \mathbb{R}$.*

Satz 6.5 (Konvergenz in Wahrscheinlichkeit imliziert Konvergenz in Verteilung)**.** *Sind $X, X_1, X_2, \dots$ Zufallsvariablen mit $X_n \xrightarrow{\mathrm{P}} X$, dann gilt $X_n \Rightarrow X$.*

Beweis. Wir bezeichnen mit $F, F_1, F_2, \dots$ die Verteilungsfunktionen von $X, X_1, X_2, \dots$.

Sei x ein Stetigkeitspunkt von F. Zu zeigen ist

$$F_n(x) \xrightarrow{n\to\infty} F(x).$$

Wegen der Stetigkeit von F in x und Monotonie gibt es zu jedem $\varepsilon > 0$ ein $\delta > 0$ mit

$$F(x) - \varepsilon \le F(x-\delta) \le F(x+\delta) \le F(x) + \varepsilon.$$

Die folgenden allgemeine Inklusion gilt für alle $n \ge 1$

$$\begin{aligned}\{X \le x-\delta\} &= \{X \le x-\delta, X_n \le x\} \cup \{X \le x-\delta, X_n > x\} \\ &\subset \{X_n \le x\} \cup \{|X_n - X| > \delta\}.\end{aligned}$$

Wir erhalten die Abschätzung

$$F(x) \le F(x-\delta) + \varepsilon \le F_n(x) + \mathrm{P}(|X_n - X| > \delta) + \varepsilon.$$

Analog gilt für alle $n \geq 1$

$$\{X_n \leq x\} \subset \{X \leq x + \delta\} \cup \{|X_n - X| > \delta\}$$

und damit

$$F_n(x) \leq F(x + \delta) + \mathrm{P}(|X_n - X| > \delta) \leq F(x) + \varepsilon + \mathrm{P}(|X_n - X| > \delta).$$

Insgesamt folgt

$$|F_n(x) - F(x)| \leq \varepsilon + \mathrm{P}(|X_n - X| > \delta).$$

Für $n \to \infty$ konvergiert die Wahrscheinlichkeit auf der rechten Seite gegen 0 und weil $\varepsilon > 0$ beliebig war folgt $F_n(x) \to F(x)$ für $n \to \infty$. □

Die Umkehrung der Aussage im letzten Satz gilt im Allgemeinen nicht wie das folgende einfache Beispiel zeigt. Es gibt eine wichtige Ausnahme, die wir in Satz 6.7 behandeln.

Beispiel 6.6 (Konvergenz in Verteilung impliziert *nicht* Konvergenz in Wahrscheinlichkeit)**.** Seien $X, X_1, X_2, \ldots$ unabhängige Bernoulli verteilte Zufallsvariablen mit Parameter $p \in (0, 1)$. Alle diese Zufallsvariablen haben dieselbe Verteilungsfunktion und daher gilt trivialerweise $X_n \Rightarrow X$.

Es gilt aber für $\varepsilon = 1/2$

$$\begin{aligned}\mathrm{P}(|X_n - X| > \varepsilon) &= \mathrm{P}(X_n \neq X) = \mathrm{P}(X_n = 1, X = 0) + \mathrm{P}(X_n = 0, X = 1) \\ &= 2p(1 - p) > 0.\end{aligned}$$

Also kann X_n nicht in Wahrscheinlichkeit gegen X konvergieren.

Satz 6.7 (Konvergenz in Wahrsch. und in Verteilung äquivalent wenn Grenzwert konstant)**.** *Seien $X_1, X_2, \ldots$ Zufallsvariablen und sei $a \in \mathbb{R}$. Dann gilt $X_n \Rightarrow a$ genau dann, wenn $X_n \xrightarrow{\mathrm{P}} a$.*

Beweis. Wegen Satz 6.5 ist natürlich nur eine Richtung zu zeigen.

Die Verteilung von $X \equiv a$ ist δ_a und die zugehörige Verteilungsfunktion ist $F(x) = \mathbb{1}_{[a,\infty)}(x)$. Einziger Unstetigkeitspunkt ist a.

Für jedes $\varepsilon > 0$ ist nach Voraussetzung

$$\lim_{n\to\infty} \mathrm{P}(X_n \leq a - \varepsilon) = F(a - \varepsilon) = 0$$

und

$$\lim_{n\to\infty} \mathrm{P}(X_n > a + \varepsilon) = 1 - \lim_{n\to\infty} \mathrm{P}(X_n \leq a + \varepsilon) = 1 - F(a + \varepsilon) = 0.$$

Mit $\{|X_n - a| > \varepsilon\} \subset \{X_n \leq a - \varepsilon\} \cup \{X_n > a + \varepsilon\}$ erhalten wir

$$\lim_{n\to\infty} \mathrm{P}(|X_n - a| > \varepsilon) = 0.$$

Das gilt für alle $\varepsilon > 0$ und es folgt $X_n \xrightarrow{\mathrm{P}} a$. □

Das folgende Resultat hat einige wichtige Anwendungen in Statistik und kann auf vielerlei Weisen verallgemeinert werden (Stichwort: *Satz von stetigen Abbildungen, enlg. continuous mapping theorem*).

Satz 6.8 (Satz von Slutsky). *Seien $X, X_1, X_2, \ldots$ und $Y_1, Y_2, \ldots$ Zufallsvariablen mit $X_n \Rightarrow X$ und $Y_n \xrightarrow{\mathrm{P}} c$ (und damit auch $Y_n \Rightarrow c$), wobei c eine Konstante ist. Dann gilt*

(i) $X_n + Y_n \Rightarrow X + c$;

(ii) $Y_n X_n \Rightarrow cX$;

(iii) $X_n / Y_n \Rightarrow X/c$, *sofern* $c \neq 0$.

Beweis. Wir beweisen nur die erste Aussage. Beweis von (ii) und (iii) ist eine Übung.

Im Folgenden bezeichnen wir mit Durch F_Z die Verteilungsfunktion der Zufallsvariablen Z. Sei t eine Stetigkeitstelle der von F_{X+c}. Dann ist $t - c$ eine Stetigkeitstelle von F_X, denn es ist

$$F_{X+c}(t) = \mathrm{P}(X + c \leq t) = \mathrm{P}(X \leq t - c) = F_X(t - c).$$

Zu zeigen ist also

$$F_{X_n+Y_n}(t) \xrightarrow{n\to\infty} F_X(t - c). \tag{6.6}$$

Sei $\varepsilon > 0$. Dann gilt

$$\begin{aligned} F_{X_n+Y_n}(t) &= \mathrm{P}(X_n + Y_n \leq t) \\ &\leq \mathrm{P}(X_n + Y_n \leq t, |Y_n - c| < \varepsilon) + \mathrm{P}(|Y_n - c| \geq \varepsilon) \\ &\leq \mathrm{P}(X_n \leq t - c + \varepsilon) + \mathrm{P}(|Y_n - c| \geq \varepsilon) \end{aligned}$$

und analog erhalten wir

$$F_{X_n+Y_n}(t) \geq \mathrm{P}(X_n \leq t - c - \varepsilon) - \mathrm{P}(|Y_n - c| \geq \varepsilon).$$

Sind $t - c$, $t - c + \varepsilon$ und $t - c - \varepsilon$ Stetigkeitspunkte von F_X so gilt

$$F_X(t - c - \varepsilon) \leq \liminf_{n\to\infty} F_{X_n+Y_n}(t) \leq \limsup_{n\to\infty} F_{X_n+Y_n}(t) \leq F_X(t - c - \varepsilon).$$

Mit $\varepsilon \downarrow 0$ entlang von Punkten $t - c + \varepsilon$ und $t - c - \varepsilon$ in denen F_X stetig ist folgt (6.6). □

Übung 6.9. *Beweisen Sie die Aussagen (ii) und (iii) in Satz 6.8.*

6.2 Zentraler Grenzwertsatz

Neben den Gesetzen der großen Zahlen gehört der zentrale Grenzwertsatz zu den wichtigsten und grundlegendsten Resultaten in der Wahrscheinlichkeitstheorie. Auch ein großer Teil der Statistik baut auf dem zentralen Grenzwertsatz auf. Unter Anderem erklärt sich durch den zentralen Grenzwertsatz die zentrale Rolle der Normalverteilung in Wahrscheinlichkeitstheorie und Statistik.

Es gibt viele Varianten von zentralen Grenzwertsätzen und je nach Voraussetzungen gibt es auch einige Arten diese zu beweisen. Wir beschränken uns hier auf Folgen von unabhängigen und identisch verteilten Zufallsvariablen und geben einen Beweis an, der auf Lindeberg zurückgeht; Eichelsbacher and Löwe (2014) ist ein gut lesbarer Übersichtsartikel über die Methode, in dem auch Verallgemeinerungen behandelt werden.

Satz 6.10 (Zentraler Grenzwertsatz)**.** *Es sei $X_1, X_2, \dots$ eine Folge von unabhängigen identisch Verteilten Zufallsvariablen mit $\mathrm{E}[X_1] = \mu \in \mathbb{R}$ und $\mathrm{Var}[X_1] = \sigma^2 \in (0, \infty)$. Dann gilt*

$$\frac{1}{\sigma\sqrt{n}}\Big(\sum_{i=1}^{n} X_i - n\mu\Big) = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) \Rightarrow Z, \tag{6.7}$$

wobei $Z \sim \mathcal{N}_{0,1}$.

Die erste Gleichheit in (6.7) ist natürlich klar. Zu beweisen ist die Verteilungskonvergenz, also die punktweise Konvergenz der Verteilungsfunktion von $\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu)$ gegen die Verteilungsfunktion der Normalverteilung Φ, gegeben durch

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}\,\mathrm{d}y.$$

Natürlich ist Φ auf ganz $\mathbb{R}$ stetig und deswegen werden Konvergenz in jedem Punkt $x \in \mathbb{R}$ nachweisen müssen.

Bemerkung 6.11 (Zentraler Grenzwertsatz vs. Gesetze der Großen Zahlen)**.** Unter den Voraussetzungen des zentralen Grenzwertsatzes gilt das sowohl das starke als auch das schwache Gesetz der großen Zahlen. Wir haben unter diesen Voraussetzungen nur das schwache bewiesen und beschränken uns in hier nur auf diesen Fall. Danach wissen wir, dass $n^{-1}\sum_{i=1}^{n} X_i \xrightarrow{\mathrm{P}} \mu$ gilt. Also verhält sich (mit großer Wahrscheinlichkeit) die Summe $\sum_{i=1}^{n} X_i$ asymptotisch wie $n\mu$ und es gilt insbesondere

$$\frac{1}{n}\Big(\sum_{i=1}^{n} X_i - n\mu\Big) \xrightarrow{\mathrm{P}} 0. \tag{6.8}$$

Der Grenzwert hier ist also die triviale Zufallsvariable $X \equiv 0$. Natürlich bleibt es bei demselben Grenzwert, wenn wir auf der linken Seite den Faktor $1/n$ durch $1/n^\alpha$ ersetzen mit $\alpha > 1$.

Es stellt sich die Frage, ob für ein $\alpha \in (0, 1)$ der Grenzwert nicht trivial, also eine „echte“ (nicht deterministische) Zufallsvariable ist. Unter den Voraussetzungen des zentralen Grenzwertsatzes ist dies der Fall für $\alpha = 1/2$. Insbesondere beschreibt der zentrale Grenzwertsatz asymptotisch die Größenordnung der Fluktuation von $\sum_{i=1}^{n} X_i$ um $n\mu$.

Einen Spezialfall des zentralen Grenzwertsatzes, der auch hier für den Beweis wichtig sein wird, haben wir bereits im Korollar 4.14 gesehen. Mit diesem Korollar folgt unter anderem Folgendes: Seien $Z_1, Z_2, \dots$ unabhängige und identisch verteilte Zufallsvariablen mit $Z_i \sim \mathcal{N}_{0,1}$, dann gilt

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \sim \mathcal{N}_{0,1}, \quad n = 1, 2, \dots \tag{6.9}$$

Hier gilt also die Konvergenzaussage (6.7) trivialerweise.

Bevor wir mit den Vorbereitungen des Beweises von Satz 6.10 beginnen geben wir noch einen anderen Spezialfall den man aus Satz 6.10 herleiten kann, wenn man dort $X_i \sim \mathrm{Ber}_p$ wählt.

Korollar 6.12 (Satz von de Moivre-Laplace)**.** *Für $n \in \mathbb{N}$ sei S_n eine binomial verteilte Zufallsvariable mit Parametern $p \in (0,1)$ und n. Dann gilt*

$$\frac{S_n - np}{\sqrt{np(1-p)}} \Rightarrow Z, \tag{6.10}$$

wobei $Z \sim \mathcal{N}_{0,1}$.

Dies ist die historisch erste Version der zentralen Grenzwertsatzes. Der (direkte) Beweis davon basiert auf Rechnungen mit und Abschätzungen von Verteilungen mit Hilfe der Stirling-Approximation von Binomialkoeffizienten. Wie wir im Abschnitt 4.2 gesehen haben ist das Rechnen mit Verteilungen von Summen von unabhängigen Zufallsvariablen (also mit Faltungen von Verteilungen) im Allgemeinen schwierig. Also werden wir nicht direkt mit Verteilungsfunktionen von $\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$ rechnen.

Wir bereiten den Beweis von Satz 6.10 mit einigen technischen Resultaten vor.

Lemma 6.13 (Abschätzung des Restglieds in der Taylor-Formel)**.** *Sei f eine 3-mal stetig differenzierbare Funktion auf $\mathbb{R}$ mit*

$$M_f^{(i)} := \sup_{x \in \mathbb{R}} |f^{(i)}(x)| < \infty, \quad i = 0, 1, 2, 3.$$

Dann gibt er eine Konstante K_f, sodass für $h \in \mathbb{R}$

$$g(h) = \sup_{x \in \mathbb{R}} \left| f(x+h) - f(x) - h f^{(1)}(x) - \frac{1}{2} h^2 f^{(2)}(x) \right| \leq K_f \min\{h^2, |h|^3\}. \tag{6.11}$$

Beweis. Nach der Taylor-Formel mit Lagrangescher Form des Restgliedes gilt

$$f(x+h) - f(x) - h f^{(1)}(x) - \frac{1}{2} h^2 f^{(2)}(x) = \frac{h^3}{6} f^{(3)}(\xi),$$

wobei ξ zwischen x und $x+h$ liegt. Es folgt für alle $h \in \mathbb{R}$

$$\left| f(x+h) - f(x) - h f^{(1)}(x) - \frac{1}{2} h^2 f^{(2)}(x) \right| \leq \frac{M_f^{(3)}}{6} |h|^3. \tag{6.12}$$

Mit Dreiecksungleichung gilt aber auch

$$\Big|f(x+h)-f(x)-hf^{(1)}(x)-\frac{1}{2}h^2f^{(2)}(x)\Big| \le M_f^{(0)}+M_f^{(0)}+hM_f^{(1)}+\frac{1}{2}h^2M_f^{(2)}$$
$$= h^2\Big(\frac{2M_f^{(0)}}{h^2}+\frac{M_f^{(1)}}{|h|}+\frac{1}{2}M_f^{(2)}\Big). \qquad (6.13)$$

Wir wählen nun b_f so, dass $\frac{2M_f^{(0)}}{h^2}+\frac{M_f^{(1)}}{|h|}\le 1$ für $|h|\ge b_f$. Also kann für $|h|\ge b_f$ die rechte Seite von (6.13) durch $h^2(\frac{1}{2}M_f^{(2)}+1)$ nach oben abgeschätzt werden. Für $|h|<b_f$ erhalten wir mit (6.12)

$$\Big|f(x+h)-f(x)-hf^{(1)}(x)-\frac{1}{2}h^2f^{(2)}(x)\Big| \le \frac{M_f^{(3)}}{6}|h|^3 \le b_f|h|^2\frac{M_f^{(3)}}{6}.$$

Setzen wir

$$K_f := \max\Big\{\frac{M_f^{(3)}}{6}, b_f\frac{M_f^{(3)}}{6}, \frac{1}{2}M_f^{(2)}+1\Big\},$$

so gilt

$$\Big|f(x+h)-f(x)-hf^{(1)}(x)-\frac{1}{2}h^2f^{(2)}(x)\Big| \le K_fh^2$$

und

$$\Big|f(x+h)-f(x)-hf^{(1)}(x)-\frac{1}{2}h^2f^{(2)}(x)\Big| \le K_f|h|^3.$$

Es folgt

$$\Big|f(x+h)-f(x)-hf^{(1)}(x)-\frac{1}{2}h^2f^{(2)}(x)\Big| \le K_f\min\{h^2,|h|^3\}.$$

Damit folgt (6.11), weil die letzte Abschätzung nicht von x abhängt. □

Korollar 6.14. *Unter den Voraussetzungen und mit Bezeichnungen von Lemma 6.13 gilt*

$$\Big|f(x+h_1)-f(x+h_2)-f'(x)(h_1-h_2)-\frac{1}{2}f''(x)(h_1^2-h_2^2)\Big|$$
$$\le g(h_1)+g(h_2) \le K_f(\min\{h_1^2,|h_1|^3\}+\min\{h_2^2,|h_2|^3\}).$$

Beweis. Der Beweis ist eine einfache Anwendung der Dreiecksungleichung. □

Lemma 6.15. *Unter den Voraussetzungen und mit Bezeichnungen von Satz 6.10 gilt*

$$\mathrm{E}\Big[f\Big(\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu)\Big)\Big] \xrightarrow{n\to\infty} \mathrm{E}[f(Z)]$$

für alle 3-mal stetig differenzierbaren Funktionen f mit beschränkten Ableitungen.

Beweis. Um uns im Folgenden etwas Schreibarbeit zu ersparen nehmen wir ohne Einschränkung der Allgemeinheit $\mu = 0$ und $\sigma^2 = 1$ an. Ansonsten betrachten wir $X_1^*, X_2^*, \dots$ mit $X_i^* = (X_i - \mu)/\sigma$. (Rechnen Sie für sich nach, dass $\mathrm{E}[X_i^*] = 0$ und $\mathrm{Var}[X_i^*] = 1$ ist, und beachten Sie, dass man (6.7) als $1/\sqrt{n}\sum_{i=1}^n X_i^* \Rightarrow Z$ schreiben kann.)

Wir definieren

$$W_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \tag{6.14}$$

Sei f eine 3-mal stetig differenzierbare Funktionen f mit beschränkten Ableitungen (wie in Lemma 6.13). Zu zeigen ist, dass

$$\mathrm{E}[f(W_n)] \xrightarrow{n\to\infty} \mathrm{E}[f(Z)] \tag{6.15}$$

gilt. Nach (6.9) hat Z dieselbe Verteilung wie $\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i$ und die Idee des Beweises besteht darin in W_n die Summanden X_i nach und nach durch die normalverteilten Z_i zu ersetzen und dabei die Fehler zu kontrollieren.

Nach Lemma 6.13 und insbesondere Korollar 6.14 gilt für $h_1, h_2 \in \mathbb{R}$

$$\left| f(x+h_1) - f(x+h_2) - f'(x)(h_1 - h_2) - \frac{1}{2} f''(x)(h_1^2 - h_2^2) \right| \le g(h_1) + g(h_2) \tag{6.16}$$

wobei $K_f \in (0, \infty)$ eine Konstante ist mit

$$g(h) \le K_f \min\{h^2, |h|^3\}. \tag{6.17}$$

Wir setzen

$$T_k := X_1 + \dots X_{k-1} + Z_{k+1} + \dots + Z_n.$$

Dann ist $\frac{1}{\sqrt{n}}(T_n + X_n) = W_n$ und $\frac{1}{\sqrt{n}}(T_1 + Z_1) = Z$.

Außerdem gilt

$$\begin{aligned} T_k + Z_k &= X_1 + \dots + X_{k-2} + X_{k-1} + Z_k + Z_{k+1} + \dots + Z_n \\ &= X_1 + \dots + X_{k-2} + X_{k-1} + Z_k + Z_{k+1} + \dots + Z_n \\ &= T_{k-1} + X_k \end{aligned}$$

Mit Teleskopsummenargument erhalten wir

$$\begin{aligned} \mathrm{E}[f(W_n) - f(Z)] &= \mathrm{E}\Big[f\Big(\frac{1}{\sqrt{n}}(T_n + X_n)\Big) - f\Big(\frac{1}{\sqrt{n}}(T_1 + Z_1)\Big)\Big] \\ &= \sum_{k=1}^{n} \mathrm{E}\Big[f\Big(\frac{1}{\sqrt{n}}(T_k + X_k)\Big) - f\Big(\frac{1}{\sqrt{n}}(T_k + Z_k)\Big)\Big]. \end{aligned}$$

Es folgt mit Dreicksungleichung und „Addition einer 0“

$$|\mathrm{E}[f(W_n) - f(Z)]| \le \sum_{k=1}^{n} \Big| \mathrm{E}\Big[f\Big(\frac{1}{\sqrt{n}}(T_k + X_k)\Big) - f\Big(\frac{1}{\sqrt{n}}(T_k + Z_k)\Big)\Big] \\ - f'\Big(\frac{T_k}{\sqrt{n}}\Big)\frac{1}{\sqrt{n}}(X_k - Z_k) - \frac{1}{2}f''\Big(\frac{T_k}{\sqrt{n}}\Big)\frac{1}{n}(X_k^2 - Z_k^2)\Big]\Big|.$$

Hier haben wir benutzt, dass $\mathrm{E}[X_k] = \mathrm{E}[Z_k] = 0$ und $\mathrm{E}[X_k^2] = \mathrm{E}[Z_k^2] = 1$ gilt, und dass T_k und Z_k sowie T_k und X_k unabhängig sind. Insbesondere ist z.B.

$$\mathrm{E}\Big[f'\Big(\frac{T_k}{\sqrt{n}}\Big)\frac{1}{\sqrt{n}}X_k\Big] = \mathrm{E}\Big[f'\Big(\frac{T_k}{\sqrt{n}}\Big)\Big]\,\mathrm{E}\Big[\frac{1}{\sqrt{n}}X_k\Big] = 0.$$

Mit der Abschätzung (6.16) und weil jeweils $X_1, \dots, X_n$ und $Z_1, \dots, Z_n$ identisch verteilt sind, erhalten wir

$$|\mathrm{E}[f(W_n) - f(Z)]| \le n\,\mathrm{E}\Big[g\Big(\frac{X_1}{\sqrt{n}}\Big)\Big] + n\,\mathrm{E}\Big[g\Big(\frac{Z_1}{\sqrt{n}}\Big)\Big].$$

Wir zeigen nun, dass beide Summanden auf der rechten Seite für $n \to \infty$ verschwinden. Das Argument ist in beiden Fällen analog, daher betrachten wir nur den ersten Summanden.

Mit (6.17) und Zerlegung des Integrals (Erwartungswertes) gilt für jedes $\varepsilon > 0$

$$\begin{aligned} n\,\mathrm{E}\Big[g\Big(\frac{X_1}{\sqrt{n}}\Big)\Big] &\le nK_f\,\mathrm{E}\Big[\frac{X_1^2}{n}\mathbb{1}_{\{|X_1| \ge \varepsilon\sqrt{n}\}}\Big] + nK_f\,\mathrm{E}\Big[\Big|\frac{X_1}{\sqrt{n}}\Big|^3\mathbb{1}_{\{|X_1| \le \varepsilon\sqrt{n}\}}\Big] \\ &\le K_f\,\mathrm{E}\big[X_1^2\mathbb{1}_{\{|X_1| \ge \varepsilon\sqrt{n}\}}\big] + \varepsilon K_f\,\mathrm{E}\big[X_1^2\big] \\ &= K_f\,\mathrm{E}\big[X_1^2\mathbb{1}_{\{|X_1| \ge \varepsilon\sqrt{n}\}}\big] + \varepsilon K_f. \end{aligned}$$

Der erste Summand verschwindet für $n \to \infty$, da $\mathrm{E}[X^2]$ endlich ist und die Folge $\{|X_1| \ge \varepsilon\sqrt{n}\}$ gegen die leere Menge absteigt. Das ist das allgemeine maßtheoretische Argument. Wenn X_1 z.B. eine ganzzahlige Zufallsvariable mit $\mathrm{E}[X^2] = \sum_{k\in\mathbb{N}} k^2\,\mathrm{P}(X_1 = \pm k) < \infty$ ist, dann konvergieren die Partialsummen

$$\mathrm{E}\big[X_1^2\mathbb{1}_{\{|X_1| \ge \varepsilon\sqrt{n}\}}\big] = \sum_{k \ge \varepsilon\sqrt{n}} k^2\,\mathrm{P}(X_1 = \pm k)$$

für $n \to \infty$ gegen 0. Analog kann man sich das für Zufallsvariablen mit Dichten überlegen.

Da $\varepsilon > 0$ in der obigen Abschätzung beliebig war, sehen wir, dass $n\,\mathrm{E}\Big[g\Big(\frac{X_1}{\sqrt{n}}\Big)\Big] \to 0$ und analog $n\,\mathrm{E}\Big[g\Big(\frac{Z_1}{\sqrt{n}}\Big)\Big] \to 0$ für $n \to \infty$ gilt. Damit ist (6.15) gezeigt. □

Beweis von Satz 6.10. Für W_n wie in (6.14) ist zu zeigen $\mathrm{P}(W_n \le x) \xrightarrow{n\to\infty} \mathrm{P}(Z \le x)$, $x \in \mathbb{R}$, bzw. äquivalent dazu

$$\mathrm{E}[\mathbb{1}_{\{W_n \le x\}}] \xrightarrow{n\to\infty} \mathrm{E}[\mathbb{1}_{\{Z \le x\}}], \quad x \in \mathbb{R}. \tag{6.18}$$

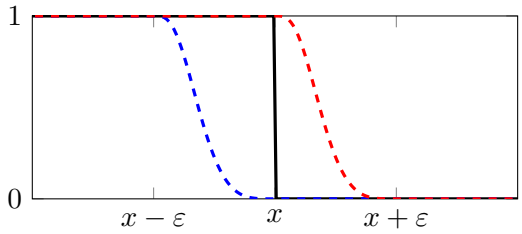

Abbildung 6.1: Approximationen der Indikatorfunktion $\mathbb{1}_{(-\infty,x]}$.

Für festes $x \in \mathbb{R}$ approximieren wir dafür die Indikatorfunktion $\mathbb{1}_{[x,\infty)}$ von oben und von unten durch 3-mal stetig differenzierbare Funktionen mit beschränkten Ableitungen und benutzen Lemma 6.15; siehe Abbildung 6.1.

Für $\varepsilon > 0$ wählen wir Funktionen $f_{\mathrm{u}}^{(\varepsilon)} \leq \mathbb{1}_{[x,\infty)} \leq f_{\mathrm{o}}^{(\varepsilon)}$, sodass die drei Funktionen außerhalb des Intervalls $(x-\varepsilon, x+\varepsilon)$ gleich sind. Ein konkrete Wahl solcher Funktionen wäre z.B.

$$f_{\mathrm{u}}^{(\varepsilon)}(t) = \begin{cases} 1 & : t < x-\varepsilon, \\ 1-(1-(\frac{t-x}{\varepsilon})^4)^4 & : x-\varepsilon \leq t \leq x, \\ 0 & : t > x, \end{cases}$$

und

$$f_{\mathrm{o}}^{(\varepsilon)}(t) = \begin{cases} 1 & : t < x, \\ 1-(1-(\frac{t-x-\varepsilon}{\varepsilon})^4)^4 & : x \leq t \leq x+\varepsilon, \\ 0 & : t > x+\varepsilon. \end{cases}$$

Es gilt nun

$$\mathrm{E}[f_{\mathrm{u}}^{(\varepsilon)}(W_n)] \leq \mathrm{P}(W_n \leq x) \leq \mathrm{E}[f_{\mathrm{o}}^{(\varepsilon)}(W_n)]$$

und

$$\begin{aligned} \mathrm{P}(Z \leq x-\varepsilon) \leq \mathrm{E}[f_{\mathrm{u}}^{(\varepsilon)}(Z)] &= \lim_{n\to\infty} \mathrm{E}[f_{\mathrm{u}}^{(\varepsilon)}(W_n)], \\ \mathrm{P}(Z \leq x+\varepsilon) \geq \mathrm{E}[f_{\mathrm{o}}^{(\varepsilon)}(Z)] &= \lim_{n\to\infty} \mathrm{E}[f_{\mathrm{o}}^{(\varepsilon)}(W_n)]. \end{aligned}$$

Dabei folgt jeweils die Gleichheit auf der rechten Seit mit Lemma 6.15. Wir erhalten

$$\mathrm{P}(Z \leq x-\varepsilon) \leq \liminf_{n\to\infty} \mathrm{P}(W_n \leq x-\varepsilon) \leq \limsup_{n\to\infty} \mathrm{P}(W_n \leq x+\varepsilon) \leq \mathrm{P}(Z \leq x+\varepsilon).$$

Die Verteilungsfunktion der Normalverteilung ist stetig. Mit $\varepsilon \to 0$ konvergiert daher sowohl die linke als auch die rechte Seite im letzten Display jeweils gegen $\mathrm{P}(Z \leq x)$, was den Beweis abschließt. □

Das folgende Resultat liefert unter Annahme von endlichen dritten Momenten eine Abschätzung der Konvergenzgeschwindigkeit im zentralen Grenzwertsatz. Wir geben es hier ohne Beweis an.

Satz 6.16 (Satz von Berry-Esseen)**.** *Es seien $X_1, X_2, \dots$ unabhängige, identische verteilte Zufallsvariablen mit $\mu = \mathrm{E}[X_1]$, $\sigma^2 = \mathrm{Var}[X_1]$ und sei $S_n = \sum_{k=1}^n X_k$. Ist $\gamma^3 = \mathrm{E}[|X - \mu|^3] < \infty$, dann gilt*

$$\sup_{x\in\mathbb{R}} \Big| \mathrm{P}\Big(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\Big) - \Phi(x)\Big| \le C \cdot \frac{\gamma^3}{\sigma^3\sqrt{n}}. \tag{6.19}$$

Dabei ist Φ die Verteilungsfunktion der Standardnormalverteilung und C ist eine von keinen Parametern abhängige Konstante mit $0.4097 \le C \le 0.4784$.

Beweis. Siehe z.B. Theorem 6.1 in Gut (2013). □

Beispiel 6.17 (Wahl von n bei Approximation der Binomialverteilung)**.** Wenn $X_1, X_2, \dots$ im obigen Satz Bernoulli verteilt sind, dann gilt

$$\gamma^3 = \mathrm{E}[|X_1 - p|^3] = p^3\,\mathrm{P}(X_1 = 0) + (1-p)^3\,\mathrm{P}(X_1 = 1) = p^3(1-p) + p(1-p)^3.$$

Hier ist $\mu = p$ und $\sigma^2 = p(1-p)$ und mit (6.19) folgt

$$\begin{aligned}
\sup_{x\in\mathbb{R}} \Big| \mathrm{P}\Big(\frac{S_n - np}{\sqrt{np(1-p)}} \le x\Big) - \Phi(x)\Big| &\le C \cdot \frac{p^3(1-p) + p(1-p)^3}{(p(1-p))^{3/2}\sqrt{n}} \\
&= C \cdot \frac{p^2 + (1-p)^2}{\sqrt{np(1-p)}} \\
&\le C \cdot \frac{1}{\sqrt{np(1-p)}} \\
&\le \frac{0.4784}{\sqrt{np(1-p)}}.
\end{aligned}$$

Die vorletzte Ungleichung gilt, weil die Funktion $p \mapsto p^2 + (1-p)^2 = 1 - 2p + 2p^2$ im Punkt $p = 1/2$ ein lokales Minimum mit Wert $1/2$ hat und an den Rändern des Intervalls $[0, 1]$ jeweils den Wert 1 annimmt.

Angenommen wir wollen die Verteilung von $\frac{S_n - np}{\sqrt{np(1-p)}}$ so durch die Normalverteilung approximieren, dass sich die Verteilungsfunktionen höchstens einen Abstand von 0.05 zueinander haben. Das ist dann der Fall wenn

$$\frac{0.4784}{\sqrt{np(1-p)}} \le 0.05,$$

bzw.

$$np(1-p) \ge \Big(\frac{0.4784}{0.05}\Big)^2 \approx 91.55.$$

Nun kann man für ein gegebenes $p \in (0, 1)$ die Anzahl der Versuche n so wählen, dass die Approximation die geforderte Genauigkeit hat. So würde man z.B. $n \ge 366$ im symmetrischen Fall $p = 1/2$ benötigen.

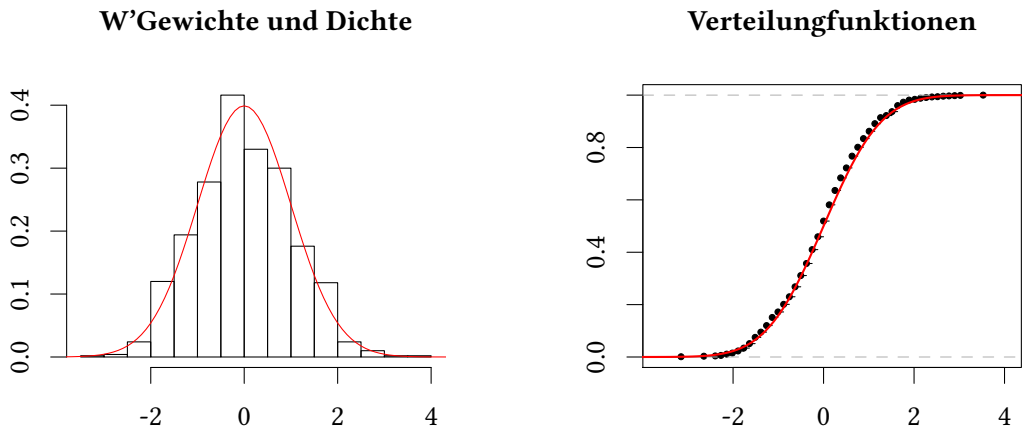

Abbildung 6.2: Approximation der Binomialverteilung im Satz von de Moivre-Laplace. Hier ist $p = 0.3$, $n = 300$ und somit ist die Berry-Esseen Schranke aus Beispiel 6.17 gegeben durch ≈ 0.06.

Man beachte, dass die Funktion $p \mapsto p(1-p)$ an den Rändern des Intervalls $[0,1]$ den Wert 0 annimmt und in $1/2$ ein lokales Maximum mit Wert $1/4$ besitzt. Dadurch erklärt sich, dass die Approximation in dem Satz von de Moivre-Laplace für $p \approx 1/2$ am besten ist und für p nah an 0 oder 1 schlechter wird.

7 Statistik

Ein typisches statistisches Problem kann wie folgt beschrieben werden: Eine Reihe von Zufallsexperimenten werden durchgeführt und deren Realisierungen als Daten aufgefasst. Die Aufgabe des Statistikers ist es wichtige bzw. relevante Informationen aus den Daten herauszufiltern und die Resultate zu interpretieren. Wir gehen hier davon aus, dass die Daten bereits vorliegen und bestimmte Voraussetzungen wie Verteilungsannahmen und Ähnliches erfüllen.

In diesem Kapitel geht um Anwendungen der Theorie der vorherigen Kapitel. Insbesondere werden wir sehen was für eine wichtige Rolle sowohl die Gesetze der großen Zahlen als auch der zentrale Grenzwertsatz in der Statistik spielen. Wir beginnen mit einem Abschnitt in dem wir verschiedene statistische Methoden anhand eines konkreten Beispiels behandeln (angelehnt an Abschnitt V.18 aus Kersting and Wakolbinger (2010)).

7.1 Beispiel: Schätzen und Testen von Anteilen

In einem See werden aus einer großen und gut durchmischten Population von Fischen 50 Fische gefangen. Davon sind 30 Männchen und 20 Weibchen. Also ist bei dem Fang der Weibchenanteil $2/5$, was auf ein unausgeglichenes Geschlechterverhältnis in der Population deuten könnte. Ob das wirklich so ist, oder ob es sich eher um eine zufällige Schwankung handelt wollen wir in diesem Abschnitt beantworten.

Es gib einen wahren aber unbekannten Anteil p von Weibchen im See. Da die Population groß ist können wir bei dem Fang von 50 Fischen vom Ziehen mit Zurücklegen ausgehen (vgl. Satz 3.5). Die Realisierung, die wir beobachten ist $X = (X_1, \dots, X_n)$ mit $X_i = 1$ wenn der i-te Fisch ein Weibchen ist, und 0 wenn es sich um ein Männchen handelt.

Nach den Gesetzen der Großen Zahlen ist

$$\hat{p}_n \coloneqq \hat{p}_n(X_1, \dots, X_n) \coloneqq \frac{1}{n} \sum_{i=1}^{n} X_i \tag{7.1}$$

ein plausibler *Schätzer* von p. Man beachte, dass der Schätzer (bzw. Schätzfunktion) eine Zufallsvariable ist. Setzt man in den Schätzer die konkreten Realisierungen von $X_1, \dots, X_n$ ein so erhält man eine *Schätzung*, also eine Realisierung des Schätzers. In unserem Beispiel ist die Schätzung gegeben durch $\hat{p}_n = 2/5 = 0.4$.

Bemerkung 7.1 (Konvention zu Notation von Schätzern)**.** Meistens bezeichnen wir Zufallsvariablen mit Großbuchstaben, in Statistik werden Schätzer typischerweise jedoch mit kleinen Buchstaben bezeichnet auch wenn es sich um Zufallsvariablen handelt. So wird beispielsweise

Schätzer eines Parameters θ oft $\hat{\theta}$ bezeichnet. Will man noch die Stichprobengröße n andeuten von der der Schätzer abhängt, so schreibt man $\hat{\theta}_n$.

Da wir mit einem unbekanntem Parameter p rechnen deuten wie die Abhängigkeit von diesem Parameter durch Indizes an, wir schreiben also P_p, E_p, Var_p etc. Eine der guten Eigenschaften des Schätzers $\hat{p}_n$ ist, dass der Schätzer in Wahrscheinlichkeit und fast sicher gegen den wahren Parameter konvergiert. Solche Schätzer heißen *konsistent* bzw. *stark konsistent*. Genauer gilt

$$\lim_{n\to\infty} \mathrm{P}_p(|\hat{p}_n - p| > \varepsilon) = 0, \quad \text{für alle } \varepsilon > 0$$

und

$$\mathrm{P}_p(\hat{p}_n \xrightarrow{n\to\infty} p) = 1.$$

Außerdem ist $\hat{p}_n$ ein *erwartungstreuer* Schätzer, d.h. der Erwartungswert des Schätzers ist gleich dem zu schätzenden Parameter

$$\mathrm{E}_p[\hat{p}_n] = p.$$

Erwartungstreue und Konsistenz sagen nicht wirklich etwas darüber aus wie gut der Schätzer für endliches (nicht zu großes) n ist. Die Varianz des Schätzers $\hat{p}_n$ ist gegeben durch

$$\mathrm{Var}_p[\hat{p}_n] = \frac{1}{n}\sigma^2,$$

wobei $\sigma^2 = \sigma^2(p) = \mathrm{Var}_p[X_1] = p(1-p)$ ist. Die Standardabweichung des Schätzers $\hat{p}_n$ von seinem Mittelwert p ist also $\sigma/\sqrt{n}$. Für eine Schätzung von σ können wir p durch $\hat{p}_n$ ersetzen und erhalten

$$\hat{\sigma}_n \coloneqq \sqrt{\hat{p}_n(1-\hat{p}_n)}.$$

Oft werden Schätzer zusammen mit der geschätzten Standardabweichung angegeben: Hier ergibt sich

$$\hat{p}_n \pm \hat{\sigma}_n/\sqrt{n} = 0.4 \pm 0.069.$$

Das gibt schon einen einigermaßen guten Eindruck über die Qualität des Schätzers. Man kann es aber noch verbessern indem man einen *Konfidenzintevall* angibt. Das ist ein *zufälliges* Intervall in dem der wahre Parameter p mit einer vorgegebenen Wahrscheinlichkeit – genannt Konfidenzniveau oder einfach Niveau – liegt. Bei diesen Überlegungen ist der Satz de Moivre-Laplace hilfreich.

Angenommen wir wollen ein Konfidenzintervall I angeben in dem der wahre Parameter p mit Wahrscheinlichkeit ≥ 0.95 liegt. Für $Z \sim \mathcal{N}_{0,1}$ ist (dafür nimmt man entweder Tabellen oder verwendet Software, R z.B.)

$$\mathrm{P}(-1.96 \leq Z \leq 1.96) \approx 0.95.$$

Nach Korollar 6.12 ist $\sqrt{n}(\hat{p}_n - p)/\sigma$ annähernd normalverteilt. Nach folgender Übung ist auch $\sqrt{n}(\hat{p}_n - p)/\hat{\sigma}_n$ annähernd normalverteilt.

Übung 7.2. *Zeigen Sie mit mehrfacher Anwendung des Satzes von Slutsky (siehe Satz 6.8)*

$$\frac{\hat{p}_n - p}{\hat{\sigma}_n/\sqrt{n}} \Rightarrow Z,$$

wobei $Z \sim \mathcal{N}_{0,1}$

Es folgt

$$0.95 \approx \mathrm{P}_p\Big(-1.96 \leq \frac{\hat{p}_n - p}{\hat{\sigma}_n/\sqrt{n}} \leq 1.96\Big) = \mathrm{P}_p\Big(\hat{p}_n - 1.96\frac{\hat{\sigma}_n}{\sqrt{n}} \leq p \leq \hat{p}_n + 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}\Big).$$

Das zufällige Intervall $[\hat{p}_n - 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}, \hat{p}_n + 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}]$ ist das *Konfidenzintervall für* p *zum (approximativen) Niveau* 0.95. In unserem Anfangsbeispiel ergibt sich für den wahren Anteil der Weibchen im See das Konfidenzintervall $[0.26, 0.54]$.

Wir haben mit *Schätzen* und *Konfidenzintervallen* bereits zwei Begriffe bzw. Verfahren aus der Statistik kennengelernt. Ein weiteres wichtiges Verfahren sind *statistische Tests*. Dabei wird ein *Hypothese* aufgestellt und geprüft ob diese zu den erhobenen Daten passt. Die Hypothese wird abgelehnt, wenn die Daten mit einer zu kleinen Wahrscheinlichkeit (typischerweise $< 5\%$, manchmal $< 1\%$) dazu passen.

In unserem Beispiel kann man die Hypothese aufstellen, dass der Anteil von Männchen und Weibchen im See gleich ist, dass also $p = 1/2$ gilt. Wir bestimmen zunächst allgemein die Wahrscheinlichkeit $\{|\hat{p}_n - p| \geq a\}$ für ein $a > 0$. Es gilt mit $Z \sim N_{0,1}$

$$\begin{aligned}\mathrm{P}_p(|\hat{p}_n - p| \geq a) &= \mathrm{P}_p\Big(\frac{|\hat{p}_n - p|}{\sigma/\sqrt{n}} \geq \frac{a}{\sigma/\sqrt{n}}\Big) \\ &\approx \mathrm{P}_p\Big(|Z| \geq \frac{a}{\sigma/\sqrt{n}}\Big) = 2\,\mathrm{P}_p\Big(Z \leq -\frac{a}{\sigma/\sqrt{n}}\Big) = 2\Phi\Big(-\frac{a}{\sigma/\sqrt{n}}\Big).\end{aligned}$$

Dabei Bezeichnet Φ die Verteilungsfunktion von Z.

In unserem Beispiel ist $n = 50$, $p = 1/2$, $a = |2/5 - 1/2| = 1/10$, $\sigma = \sqrt{1/4} = 1/2$. Wir erhalten

$$\mathrm{P}_{1/2}(|\hat{p}_n - 1/2| \geq 0.1) \approx 2\Phi\Big(-0.2\sqrt{50}\Big) \approx 0.29.$$

Die Hypothese $p = 1/2$, würde man in diesem Fall nicht ablehnen. Wenn wir mit einer größeren Stichprobengröße n zur selben Schätzung $\hat{p}_n = 2/5$ gekommen wären, dann würden wir die Hypothese nicht ablehnen, wenn $2\Phi\big(-\frac{a}{\sigma/\sqrt{n}}\big) \leq 0.05$ wäre. Es gilt

$$2\Phi\Big(-\frac{a}{\sigma/\sqrt{n}}\Big) \leq 0.05 \iff -\frac{a}{\sigma/\sqrt{n}} \leq \Phi^{-1}(0.025) \approx -1.96 \iff n \geq \Big(1.96\frac{\sigma}{a}\Big)^2.$$

In unserem Beispiel mit $p = 1/2$, $a = |2/5 - 1/2| = 1/10$ und $\sigma = \sqrt{1/4} = 1/2$ würden wir die Hypothese ablehnen, wenn $n \geq 97$ wäre.

7.2 Schätzen des Erwartungswertes und Konfidenzintervalle

Oft kann man in der Statistik plausibel begründen (annehmen), dass erhobene Daten Realisierungen von Zufallsvariablen aus bestimmten Verteilungsfamilien sind.

Ein *statistisches Modell* ist gegeben durch einen Wahrscheinlichkeitsraum $(\mathcal{X}, \mathcal{B}, P)$. Dabei heißt $\mathcal{X}$ der *Beobachtungsraum* oder *Ergebnisraum*; $\mathcal{B}$ ist die σ-Algebra (die die Menge der beobachtbaren Ereignisse darstellt); P ist das unbekannte zugrunde liegende Wahrscheinlichkeitsgesetz.

Die Daten, auch *Stichproben* genannt, werden repräsentiert durch Realisierungen einer Zufallsgröße X mit Werten in $\mathcal{X}$ und Verteilung P. Oft ist $X = (X_1, \dots, X_n)$, $\mathcal{X} = \mathbb{R}^n$, $\mathcal{B} = \mathcal{B}(\mathbb{R}^n)$, $P = \mathrm{P}_X$. Wir gehen hier von *parametrischen Familien* aus, d.h. wir nehmen an, dass es einen *Parameterraum* $\Theta \subset \mathbb{R}^d$ gibt und eine Familie $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ mit $P \in \mathcal{P}$. Das Ziel ist, es aus den Daten, also aus Realisierungen von X, Informationen über die unbekannte Verteilung P oder über einige ihrer unbekannten Charakteristiken, wie z.B. Erwartungswert oder Varianz, herzuleiten.

Beispiel 7.3 (Parametrische Familien)**.**

(i) $P_\theta = \mathcal{N}_{\mu,\sigma^2}$ mit $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_{>0}$;

(ii) $P_\theta = \mathcal{N}_{\mu,\sigma^2}$, mit $\theta = \mu \in \Theta = \mathbb{R}$, σ^2 ist also bekannt;

(iii) $P_\theta = \mathrm{Bin}_{n,p}$, $\theta = p \in \Theta = [0, 1]$;

(iv) $P_\theta = \mathcal{U}_{[a,b]}$, $\theta = (a, b) \in \Theta = \{(x, y) \in \mathbb{R}^2 : x < y\}$.

Für messbare (feste) Funktionen T auf $\mathcal{X}$ werden die Zufallsvariablen $T(X)$ *Statistiken* bezeichnet. Beachten Sie, dass der Zufall in der Beobachtung X steckt und nicht in T. Typischerweise möchte man mit Statistiken Daten reduzieren. Die Abbildung $T = \mathrm{id}$ ist eine triviale Statistik, bei der keine Datenreduktion stattfindet.

Beispiel 7.4 (Statistiken)**.** Sei $X = (X_1, \dots, X_n)$ mit $X_1, \dots, X_n$ u.i.v.

(i) *Stichprobenmittelwert* ist die Statistik

$$T_1(X) \coloneqq \overline{X}_n \coloneqq \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{7.2}$$

Es gilt

$$\mathrm{E}[T_1(X)] = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[X_i] = \mathrm{E}[X_1].$$

(ii) *Stichprobenvarianz* ist die Statistik

$$T_2(X) \coloneqq \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2, \quad n \geq 2. \tag{7.3}$$

Es gilt (Übung!)

$$\mathrm{E}[T_2(X)] = \mathrm{Var}[X_1].$$

Bei einer sinnvollen Datenreduktion möchte man keine wichtigen Informationen verlieren, die man für Rückschlüsse auf die unbekannte Verteilung benötigt.

Definition 7.5 (Suffiziente Statistiken)**.** Sei X eine Stichprobe aus $P \in \mathcal{P}$. Eine Statistik T heißt *suffizient für P (oder für θ wenn $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$)*, wenn die bedingte Verteilung von X gegeben $T(X)$ bekannt ist (ist unabhängig von θ).

Beispiel 7.6. Sei $X = (X_1, \ldots, X_n)$, $X_1, \ldots, X_n$ u.i.v. mit $X_i \sim \mathrm{Ber}_\theta$, $\theta \in \Theta = [0, 1]$. Also ist

$$P_\theta(X_i = x_i) = \theta^{x_i}(1-\theta)^{1-x_i}, \quad x_i \in \{0, 1\}.$$

Intuitiv ist klar, dass bei bekanntem n enthält $T(x) = \sum_{i=1}^n x_i$ bis auf die Anordnung der 0-er und 1-er genauso viel Information über p wie der beobachtete Datensatz $x = (x_1, \ldots, x_n)$.

Das kann man natürlich auch formal nachrechnen. Für $t \in \{0, \ldots, n\}$ und $x \in \{0, 1\}^n$ gilt

$$P_\theta(X = x | T = t) = \frac{P_\theta(X = x, T = t)}{P_\theta(T = t)}.$$

Die Zufallsvariable $T(X)$ ist natürlich $\mathrm{Bin}_{n,\theta}$ verteilt und somit ist

$$P_\theta(T = t) = \binom{n}{t}\theta^t(1-\theta)^{n-t}.$$

Für x mit $\sum_{i=1}^n x_i \neq t$ gilt $P_\theta(X = x, T = t) = 0$. Für x mit $\sum_{i=1}^n x_i = t$ erhalten wir

$$\begin{aligned} P_\theta(X = x, T = t) = P_\theta(X_1 = x_1, \ldots, X_n = x_n) &= \prod_{i=1}^n P_\theta(X_i = x_i) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} = \theta^t(1-\theta)^{n-t}. \end{aligned}$$

Es folgt

$$P_\theta(X = x | T = t) = \begin{cases} 0 & : \text{ falls } \sum_{i=1}^n x_i \neq t, \\ \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}} & : \text{ falls } \sum_{i=1}^n x_i \neq t. \end{cases}$$

Die bedingte Verteilung von X gegeben T hängt nicht von θ ab. Also ist T eine suffiziente Statistik für θ.

Das Ziel einer *Punkt- oder Parameterschätzung* ist das Schätzen eines Parameters oder Aspekts $\vartheta = h(\theta)$ der Verteilung $P \in \{P_\theta : \theta \in \Theta\}$. Die Funktion h könnte die Identität sein, dann würde man θ selbst schätzen. Sie könnte bei mehrdimensionalen Parametern aber auch Projektion auf einen Teilraum sein, z.B. $\vartheta = \mu = h(\theta)$ für $\theta = (\mu, \sigma^2)$.

Gesucht ist eine Statistik, in diesem Fall auch Schätzfunktion oder Schätzer genannt, $\hat{\vartheta} : \mathcal{X} \to h(\Theta)$. Natürlich ist dann $\hat{\vartheta}(X)$ eine $h(\Theta)$-wertige Zufallsvariable und man sucht möglichst *(asymptotisch) erwartungstreue* Schätzer mit hoher Konzentration um den wahren Parameter ϑ. Ein übliches Kriterium ist, dass der *mittlere quadratische Fehler*

$$\mathrm{E}_\vartheta[(\hat{\vartheta}(X) - \vartheta)^2] = \big(\mathrm{E}_\vartheta[\hat{\vartheta}(X) - \vartheta]\big)^2 + \mathrm{Var}_\vartheta[\hat{\vartheta}(X)] \tag{7.4}$$

möglichst klein ist und mit wachsender Stichprobengröße verschwindet.

Definition 7.7 (Erwartungstreue und Konsistenz)**.** Ein Schätzer $\hat{\vartheta}_n = \hat{\vartheta}_n(X_1, \ldots, X_n)$ von $\vartheta = h(\theta)$ heißt *erwartungstreu (engl. unbiased)*, wenn für alle n

$$\mathrm{E}_\theta[\hat{\vartheta}_n] = \vartheta$$

gilt. Er heißt *asymptotisch erwartungstreu*, wenn

$$\mathrm{E}_\theta[\hat{\vartheta}_n] \xrightarrow{n\to\infty} \vartheta.$$

Die Differenz $\mathrm{E}_\theta[\hat{\vartheta}(X) - \vartheta]$ bezeichnet man als *Verfälschung (engl. bias)*.

Der Schätzer $\hat{\vartheta}_n$ heißt *konsistent*, wenn er in P_θ Wahrscheinlichkeit gegen den wahren Parameter $\vartheta = h(\theta)$ konvergiert, d.h. für alle $\varepsilon > 0$ gilt $\mathrm{P}_\theta(|\hat{\vartheta}_n - \vartheta| > \varepsilon) \xrightarrow{n\to\infty} 0$.

Der erste Summand auf der rechten Seite von (7.4) (asymptotisch) verschwindet wenn der Schätzer (asymptotisch) erwartungstreu ist. Also ist neben der (asymptotischen) Erwartungstreue eine kleine Varianz des Schätzers wünschenswert.

Beispiel 7.8. Es seien $X_1, \ldots, X_n$ u.i.v. $X_i \sim \mathrm{Ber}_p$, $p \in (0, 1)$. Ferner sei $x = (x_1, \ldots, x_n)$ eine Realisierung von $X = (X_1, \ldots, X_n)$. Gesucht ist ein möglichst „guter“ Schätzwert für p.

1. Möglichkeit: $\hat{p}_1(x) = \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Der Schätzer $\hat{p}_1$ ist erwartungstreu und nach dem schwachen Gesetz der großen Zahlen ist es auch konsistent. Beachten Sie, dass auch $\hat{p}_1 + 10000/n$ auch konsistent ist, also ist Konsistenz allein oft unbefriedigend.

2. Möglichkeit: $\hat{p}_2(x) = \frac{1+\sum_{i=1}^n x_i}{n+2}$. Dieser Schätzer ist asymptotisch erwartungstreu aber nicht erwartungstreu.

Intuitiver Vergleich von $\hat{p}_1$ und $\hat{p}_2$: Bei Realisierungen $\underline{0} = (0, \ldots, 0)$ und $\underline{1} = (1, \ldots, 1)$ gilt

$$\hat{p}_1(\underline{0}) = 0 \qquad \hat{p}_1(\underline{1}) = 1$$
$$\hat{p}_2(\underline{0}) = \frac{1}{n+2} \qquad \hat{p}_2(\underline{1}) = \frac{n+1}{n+2}.$$

Bei $n = 5$ haben wir $\hat{p}_2(\underline{0}) = 1/7$ und $\hat{p}_2(\underline{1}) = 6/7$. Für kleine n und Realisierungen $\underline{0}$ und $\underline{1}$ scheint $\hat{p}_2$ realistischere Schätzungen zu liefern.

Für die Varianzen der Schätzer gilt

$$\mathrm{Var}_p[\hat{p}_1(X)] = \frac{p(1-p)}{n}$$
$$\mathrm{Var}_p[\hat{p}_2(X)] = \frac{1}{(n+1)^2}\,\mathrm{Var}_p\Big[1 + \sum_{i=1}^n X_i\Big] = \frac{np(1-p)}{(n+1)^2}$$

und man überlegt sich leicht

$$\frac{p(1-p)}{n} > \frac{np(1-p)}{(n+1)^2}.$$

Also ist $\mathrm{Var}_p[\hat{p}_1(X)] > \mathrm{Var}_p[\hat{p}_2(X)]$ für alle $b \in (0,1)$ und alle $n \in \mathbb{N}$.

Das obige Beispiel zeigt, dass man asymptotisch erwartungstreue und konsistente Schätzer konstruieren kann, die kleinere Varianz haben als erwartungstreue Schätzer. Oft sucht man nach besten Schätzern unter den erwartungstreuen Schätzern. In der folgenden Definition geht es um ein klassisches Optimalitätskriterium für Schätzer.

Definition 7.9. Eine Schätzfunktion $\hat{\vartheta}^*$ heißt *beste erwartungstreue Schätzfunktion für* $\vartheta = h(\theta)$ *bezüglich* $\{P_\theta : \theta \in \Theta\}$ falls gilt

(i) $\hat{\vartheta}^*$ ist erwartungstreu für ϑ;

(ii) $\hat{\vartheta}^*$ hat unter allen erwartungstreuen Schätzern die kleinste Varianz, d.h. für alle $\theta \in \Theta$ und alle erwartungstreue Schätzer $\hat{\vartheta}$ von ϑ gilt

$$\mathrm{Var}_\theta[\hat{\vartheta}^*] \leq \mathrm{Var}_\theta[\hat{\vartheta}].$$

Solche Schätzern nennt man auch UMVUE (uniformly minimum variance unbiased estimator).

Bemerkung 7.10. Man kann zeigen, dass das arithmetische Mittel $\hat{\mu}(X) = \overline{X}$ und die Stichprobenvarianz $\hat{\sigma}^2(X) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ unter recht allgemeinen Annahmen beste erwartungstreue Schätzer für den Erwartungswert μ und die Varianz σ^2 sind.

7.2.1 Eigenschaften des Stichprobenmittels

Sei $X = (X_1, \ldots, X_n)$ eine Stichprobe von unabhängigen und identisch verteilten Zufallsvariablen mit $\mathrm{E}[X_i] = \mu \in \mathbb{R}$ und $\mathrm{Var}[X_i] = \sigma^2 \in (0, \infty)$. Wir setzen $\theta = (\mu, \sigma^2)$. Eine Schätzfunktion für μ ist

$$\hat{\mu}_n(X) = \overline{X}_n = \tfrac{1}{n}\sum_{i=1}^n X_i.$$

Dieser Schätzer $\hat{\mu}_n(X)$ folgende Eigenschaften:

(i) Es gilt $\mathrm{E}_\theta[\hat{\mu}_n(X)] = \mu$ und $\mathrm{Var}_\theta[\hat{\mu}_n(X)] = \frac{\sigma^2}{n}$.

(ii) Nach dem Gesetz der großen Zahlen ist $\hat{\mu}_n(X)$ konsistent, d.h. $\hat{\mu}_n(X) \xrightarrow{\mathrm{P}} \mu$.

(iii) Nach dem zentralen Grenzwertsatz ist $\hat{\mu}_n(X)$ asymptotisch normal: Für $n \to \infty$ gilt

$$\frac{\hat{\mu}_n(X) - \mathrm{E}_\theta[\hat{\mu}_n(X)]}{\sqrt{\mathrm{Var}_\theta[\hat{\mu}_n(X)]}} = \frac{\sqrt{n}(\hat{\mu}_n(X) - \mu)}{\sigma} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \Longrightarrow Z \sim \mathcal{N}_{0,1}.$$

Insbesondere gilt für große Stichprobenumfänge n

$$\mathrm{P}_\theta(\hat{\mu}_n(X) \leq x) = \mathrm{P}_\theta\Big(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq \frac{x-\mu}{\sigma/\sqrt{n}}\Big) \approx \Phi\Big(\frac{x-\mu}{\sigma/\sqrt{n}}\Big).$$

Auf den obigen Eigenschaften basieren sehr oft approximative Berechnungen von z.B. Konfidenzintervallen für den Erwartungswert. Wenn der Stichprobenumfang n nicht groß ist, dann kann man auch exakte Rechnungen sowohl für den Schätzer als auch für Konfidenzintervalle durchführen.

Beispiel 7.11. (i) *Binomialverteilung:* Sind $X_1, \dots, X_n$ u.i.v. mit $X_i \sim \mathrm{Ber}_p$, dann ist

$$n\hat{p}_n(X) = \sum_{i=1}^{n} X_i \sim \mathrm{Bin}_{n,p}\,.$$

Insbesondere ist $\hat{p}_n(X)$ eine Zufallsvariable mit Träger $\{\frac{k}{n} : k = 0, \dots, n\}$ und es gilt

$$\mathrm{P}_p\Big(\hat{p}_n(X) = \frac{k}{n}\Big) = \mathrm{P}_p\Big(\sum_{i=1}^{n} X_i = k\Big) = \binom{n}{k} p^k (1-p)^{n-k}.$$

(ii) *Poissonverteilung:* Sind $X_1, \dots, X_n$ u.i.v. mit $X_i \sim \mathrm{Poi}_\mu$, dann ist

$$n\hat{\mu}_n(X) = \sum_{i=1}^{n} X_i \sim \mathrm{Poi}_{n\mu}$$

und $\hat{\mu}_n(X)$ ist eine Zufallsvariable mit Träger $\{\frac{k}{n} : k = 0, 1, \dots\}$. Die Verteilung ist gegeben durch

$$\mathrm{P}_p\Big(\hat{\mu}_n(X) = \frac{k}{n}\Big) = \mathrm{P}_p\Big(\sum_{i=1}^{n} X_i = k\Big) = e^{-n\mu}\frac{(n\mu)^k}{k!}.$$

7.2.2 Konfidenzintervalle für den Erwartungswert der Normalverteilung

Sei $X = (X_1, \dots, X_n)$, wobei $X_1, \dots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $X_i \sim \mathcal{N}_{\mu,\sigma^2}$ mit unbekanntem $\mu \in \mathbb{R}$ und bekanntem σ^2. Gesucht ist ein (möglichst) kleines Konfidenzintervall zur Sicherheit $1-\alpha$, $\alpha \in (0,1)$. Wegen der Symmetrie der Normalverteilung machen wir den folgenden Ansatz:

$$I(X) = (\overline{X}_n - c, \overline{X}_n + c), \quad c > 0 \text{ konstant.}$$

Es gilt

$$\begin{aligned}\mathrm{P}_\mu(\mu \in I(X)) &= \mathrm{P}_\mu(-c \le \overline{X}_n - \mu \le c) = \mathrm{P}_\mu\Big(-c \le \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \le \frac{c}{\sigma/\sqrt{n}}\Big) \\ &= \Phi\Big(\frac{c}{\sigma/\sqrt{n}}\Big) - \Phi\Big(-\frac{c}{\sigma/\sqrt{n}}\Big) = 1 - 2\Phi\Big(-\frac{c}{\sigma/\sqrt{n}}\Big) = 2\Phi\Big(\frac{c}{\sigma/\sqrt{n}}\Big) - 1.\end{aligned}$$

Das kleinste Konfidenzintervall zur Sicherheit $1-\alpha$ erhält man, wenn man c so wählt, dass dieser Ausdruck gleich $1-\alpha$ ist. Es muss also gelten

$$2\Phi\Big(\frac{c}{\sigma/\sqrt{n}}\Big) - 1 = 1 - \alpha,$$

was gleichbedeutend mit

$$\Phi\Big(\frac{c}{\sigma/\sqrt{n}}\Big) = 1 - \frac{\alpha}{2}$$

bzw. mit

$$c = \Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big)\frac{\sigma}{\sqrt{n}}.$$

Wir setzen $z_\alpha = \Phi^{-1}(1-\alpha)$, z_α ist also das *obere α-Quantil* der Standardnormalverteilung. Mit dieser Notation haben wir

$$c = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

und das gesuchte Konfidenzintervall zur Sicherheit $1-\alpha$ ist

$$I(X) = (\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}).$$

Manchmal möchte man nur einseitige Konfidenzintervalle definieren. Das macht man ähnlich wie oben mit den Ansätzen $I(X) = (-\infty, \overline{X}_n + c)$ oder $I(X) = (\overline{X}_n - c, \infty)$.

7.2.3 Exakte Konfidenzgrenzen im Binomialmodell

Sei $X \sim \mathrm{Bin}_{n,p}$, z.B. $X = \sum_{i=1}^n X_i$, $X_i \sim \mathrm{Bin}_p$ unabhängig und sei $\alpha \in (0,1)$. Gesucht ist die obere Konfidenzgrenze $\hat{p}_{\mathrm{o}}(X)$ für den wahren Parameter p, d.h. für alle $p \in (0,1)$ muss gelten

$$\mathrm{P}_p(p \le \hat{p}_{\mathrm{o}}(X)) \ge 1-\alpha \quad \text{bzw.} \quad \mathrm{P}_p(\hat{p}_{\mathrm{o}}(X) \le p) \le \alpha.$$

Sei also $\alpha \in (0,1)$. Die Verteilungsfunktion

$$F_p(x) = \mathrm{P}_p(X \le x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$$

ist für $x < n$ streng monoton fallend in p, weil $\frac{\partial}{\partial p} F_p(x) < 0$ ist (Übung!).

Für jedes $x \in \{0,1,\ldots,n-1\}$ sei $\hat{p}_{\mathrm{o}}(x)$ implizit definiert durch $F_{\hat{p}_{\mathrm{o}}(x)}(x) = \alpha$. Dann gilt für $x \in \{0,1,\ldots,n-1\}$

$$\hat{p}_{\mathrm{o}}(x) \le p \iff \alpha = F_{\hat{p}_{\mathrm{o}}(x)}(x) \ge F_p(x).$$

Nun ist $F_p(x)$ monoton wachsend in x, d.h.

$$F_p(x) \le \alpha \iff x \le \max\{\ell \in \mathbb{N}_0 : F_p(\ell) \le \alpha\}.$$

Wir setzen $\hat{p}_{\mathrm{o}}(n) = 1$. Dann ist $\hat{p}_{\mathrm{o}}(n) > p$ und $n > \max\{\ell \in \mathbb{N}_0 : F_p(\ell) \leq \alpha\}$. Insgesamt folgt

$$\mathrm{P}_p(\hat{p}_{\mathrm{o}}(X) \leq p) = \mathrm{P}_p(X \leq \max\{\ell \in \mathbb{N}_0 : F_p(\ell) \leq \alpha\}) \leq \alpha.$$

Also ist durch $F_{\hat{p}_{\mathrm{o}}(x)}(x) = \alpha$ für $x \in \{0, 1, \ldots, n-1\}$ und $\hat{p}_{\mathrm{o}}(n) = 1$ die obere Konfidenzgrenze definiert.

Die untere Konfidenzgrenze $\hat{p}_{\mathrm{u}}(x)$ kann man mit ähnlichen Argumenten herleiten. Es muss gelten

$$\mathrm{P}_p(p \geq \hat{p}_{\mathrm{u}}(X)) \geq 1 - \alpha \quad \text{bzw.} \quad \mathrm{P}_p(\hat{p}_{\mathrm{u}}(X) \geq p) \leq \alpha.$$

Setze $G_p(x) = \mathrm{P}_p(X \geq x) = \sum_{k=x}^{n} \binom{n}{k} p^k (1-p)^{n-k}$. Für $x \in \{1, \ldots, n\}$ ist $G_p(x)$ streng monoton wachsend in p. Also können wir $\hat{p}_{\mathrm{u}}(x)$ für $x \in \{1, \ldots, n\}$ implizit durch $G_{\hat{p}_{\mathrm{u}}(x)}(x) = \alpha$ definieren und erhalten

$$\hat{p}_{\mathrm{u}}(x) \geq p \iff \alpha = G_{\hat{p}_{\mathrm{u}}(x)}(x) \geq G_p(x).$$

Da $G_p(x)$ monoton fallend in x ist, gilt

$$G_p(x) \leq \alpha \iff x \geq \min\{\ell \in \mathbb{N}_0 : G_p(\ell) \leq \alpha\}.$$

Für $x = 0$ setzen wir $\hat{p}_{\mathrm{u}}(0) = 0$. Insgesamt erhalten wir

$$\mathrm{P}_p(\hat{p}_{\mathrm{u}}(X) \geq p) = \mathrm{P}_p(X \geq \min\{\ell \in \mathbb{N}_0 : G_p(\ell) \leq \alpha\}) \leq \alpha.$$

Also ist durch $G_{\hat{p}_{\mathrm{u}}(x)}(x) = \alpha$ für $x \in \{1, \ldots, n\}$ und $\hat{p}_{\mathrm{u}}(0) = 0$ die untere Konfidenzgrenze zur Sicherheit $1 - \alpha$ definiert.

Ist man an einem *zweiseitigen Konfidenzintervall* zur Sicherheit $1 - \alpha$ interessiert, so bestimmen wir wie oben zunächst $\hat{p}_{\mathrm{u}}$ und $\hat{p}_{\mathrm{o}}$ jeweils zur Sicherheit $1 - \alpha/2$. Dann gilt

$$\begin{aligned}
\mathrm{P}_p\big(p \in (p_{\mathrm{u}}(X), \hat{p}_{\mathrm{o}}(X))\big) &= 1 - \mathrm{P}_p\big(p \notin (p_{\mathrm{u}}(X), \hat{p}_{\mathrm{o}}(X))\big) \\
&= 1 - \big[\mathrm{P}_p\big(p \leq p_{\mathrm{u}}(X)\big) + \mathrm{P}_p(p \geq \hat{p}_{\mathrm{o}}(X))\big] \\
&\geq 1 - \alpha.
\end{aligned}$$

Zur konkreten Berechnung der Konfidenzgrenzen kann man eine Beziehung der Verteilungsfunktion der Binomialverteilung mit einer bestimmten F-Verteilung ausnutzen. Es gilt (siehe z.B. Kapitel zu Binomialverteilung inJohnson et al. (1992)):

$$\sum_{k=x}^{n} \binom{n}{k} p^k (1-p)^{n-k} = \mathrm{P}\Big(Y > \frac{1-p}{p} \cdot \frac{x}{n-x+1}\Big), \tag{7.5}$$

wobei Y eine Zufallsvariable die $F_{2(n-x+1),2x}$ verteilt ist. Quantile der F_{ν_1,ν_2} Verteilungen sind in Formelsammlungen und Handbüchern zu Statistik oft in Tabellen zu finden.

Damit kann man z.B. die untere Konfidenzgrenze berechnen. Sei $\alpha \in (0,1)$. Dann ist nach (7.5)

$$\begin{aligned}\alpha = G_{\hat{p}_{\mathrm{u}}(x)}(x) &= \mathrm{P}\Big(Y > \frac{1-\hat{p}_{\mathrm{u}}(x)}{\hat{p}_{\mathrm{u}}(x)} \cdot \frac{x}{n-x+1}\Big) \\ &= 1 - \mathrm{P}\Big(Y \leq \frac{1-\hat{p}_{\mathrm{u}}(x)}{\hat{p}_{\mathrm{u}}(x)} \cdot \frac{x}{n-x+1}\Big).\end{aligned}$$

Also ist

$$1 - \alpha = F_Y\Big(\frac{1-\hat{p}_{\mathrm{u}}(x)}{\hat{p}_{\mathrm{u}}(x)} \cdot \frac{x}{n-x+1}\Big).$$

Nun ist F_Y streng monoton wachsend und damit invertierbar. Es folgt

$$F_Y^{-1}(1-\alpha) = \frac{1-\hat{p}_{\mathrm{u}}(x)}{\hat{p}_{\mathrm{u}}(x)} \cdot \frac{x}{n-x+1}.$$

Das können wir nach $\hat{p}_{\mathrm{u}}(x)$ auflösen und erhalten

$$\hat{p}_{\mathrm{u}}(x) = \frac{1}{1+a}, \quad \text{mit } \; a = F_Y^{-1}(1-\alpha) \cdot \frac{n-x+1}{x}.$$

7.3 Schätzen der Varianz

Sei $X = (X_1, \ldots, X_n)$ eine Stichprobe von unabhängigen und identisch verteilten Zufallsvariablen mit $\mathrm{E}[X_i] = \mu \in \mathbb{R}$, $\mathrm{Var}[X_i] = \sigma^2 \in (0,\infty)$ und $\mu_4 = \mathrm{E}[(X_i-\mu)^4] < \infty$. Wir wollen die Varianz σ^2 schätzen.

Wenn μ bekannt ist, so können wir σ^2 wie folgt schätzen: Wir setzen $Y_i = (X_i-\mu)^2$. Dann sind $Y_1, \ldots, Y_n$ unabhängig und identisch verteilt und es gilt $\mathrm{E}[Y_i] = \mathrm{Var}[X_i] = \sigma^2$, sowie $\mathrm{Var}[Y_i] < \infty$. Damit ist

$$\hat{\sigma}_1^2(X) \coloneqq \overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2$$

ein Schätzer von σ^2. Dieser Schätzer hat oben diskutierte Eigenschaften des Stichprobenmittels wie z.B. Erwartungstreue, Konsistenz etc.

Wenn μ bekannt ist so können wir dennoch die Stichprobenvarianz

$$\hat{\sigma}^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2 \tag{7.6}$$

als Schätzer für σ^2 verwenden. Auch dieser Schätzer ist erwartungstreu, denn es ist

$$\begin{aligned}\sum_{i=1}^n (X_i-\mu)^2 &= \sum_{i=1}^n \big((X_i-\overline{X}) + (\overline{X}-\mu)\big)^2 \\ &= \sum_{i=1}^n (X_i-\overline{X})^2 + \sum_{i=1}^n (\overline{X}-\mu)^2 + 2(\overline{X}-\mu)\sum_{i=1}^n (X_i-\overline{X}).\end{aligned}$$

Wegen $\sum_{i=1}^n X_i = n\overline{X}$ verschwindet der letzte Summand. Es folgt

$$\begin{aligned}\mathrm{E}\Big[\sum_{i=1}^n (X_i - \overline{X})^2\Big] &= \mathrm{E}\Big[\sum_{i=1}^n (X_i - \mu)^2\Big] - \mathrm{E}\Big[\sum_{i=1}^n (\overline{X} - \mu)^2\Big] \\ &= n\,\mathrm{Var}[X_1] - n\,\mathrm{Var}[\overline{X}] = n\sigma^2 - \sigma^2 = (n-1)\sigma^2\end{aligned}$$

und somit

$$\mathrm{E}[\hat{\sigma}^2] = \sigma^2.$$

Man kann weiter nachrechnen, dass

$$\mathrm{Var}[\hat{\sigma}^2] = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\sigma^4).$$

Schließlich ist $\hat{\sigma}^2$ auch konsistent, denn es ist

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{1}{n-1}\Big(\sum_{i=1}^n (X_i - \mu)^2 - n(\overline{X} - \mu)^2\Big) \\ &= \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 - \frac{n}{n-1}(\overline{X} - \mu)^2 \xrightarrow{\mathrm{P}} 1 \cdot \sigma^2 - 1 \cdot 0.\end{aligned}$$

Bemerkung 7.12 (Einige wichtige Verteilungen: χ^2_n, t_n, $F_{n,m}$)**.** Seien die Zufallsvariablen $Z_1, \ldots, Z_n$ unabhängig und identisch verteilt mit $Z_i \sim \mathcal{N}_{0,1}$.

(i) Die Zufallsvariable $\sum_{i=1}^n Z_i^2$ ist χ^2_n verteilt; man nennt diese Verteilung *(zentrale)* χ^2*-Verteilung* mit n Freiheitsgraden. Die χ^2 Verteilung ist eine Spezielle Gammaverteilung. Mit Bezeichnungen in Definition 4.5 gilt nämlich $\chi^2_n = \Gamma_{1/2,n/2}$. Die Dichte ist gegeben durch

$$f_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \quad x > 0. \tag{7.7}$$

(ii) Seien $U \sim \mathcal{N}_{0,1}$ und $V \sim \chi^2_n$ stochastisch unabhängig. Die Zufallsvariable $U/\sqrt{V/n}$ ist dann t_n verteilt; man nennt diese Verteilung *studentsche* t*-Verteilung*, oder einfach t*-Verteilung* mit n Freiheitsgraden. Die Dichte ist gegeben durch

$$f_n(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}\Big(1 + \frac{x^2}{n}\Big)^{-(n+1)/2}, \quad x \in \mathbb{R}. \tag{7.8}$$

Man kann zeigen, dass f_n für $n \to \infty$ punktweise gegen die Dichte der Standardnormalverteilung konvergiert.

(iii) Seien $X_1 \sim \chi^2_n$ und $X_2 \sim \chi^2_m$ unabhängig. Die Zufallsvariable $\frac{X_1/n}{X_2/m}$ ist dann $F_{n,m}$ verteilt; man nennt diese Verteilung F*-Verteilung mit* n *Freiheitsgraden im Zähler und* m *Freiheitsgraden im Nenner.* Die Dichte bekommt man aus den entsprechenden Dichten der χ^2 Verteilung. Sie ist gegeben durch

$$f_{n,m}(x) = \frac{n^{n/2}m^{m/2}\Gamma((n+m)/2)x^{n/2-1}}{\Gamma(n/2)\Gamma(m/2)(m+nx)^{(n+m)/2}}, \quad x > 0. \tag{7.9}$$

Satz 7.13 (Verteilung von $\hat{\mu}$ und $\hat{\sigma}^2$ bei Normalität)**.** *Seien $X_1, \dots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $X_i \sim \mathcal{N}_{\mu,\sigma^2}$, $0 < \sigma^2 < \infty$. Dann gilt*

(i) $\hat{\mu}(X) = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ *und* $\hat{\sigma}^2(X) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ *sind stochastisch unabhängig.*

(ii) $\hat{\mu}(X)$ *ist* $\mathcal{N}_{\mu,\sigma^2/n}$ *verteilt.*

(iii) $\frac{n-1}{\sigma^2}\hat{\sigma}^2(X)$ *ist* χ^2_{n-1} *verteilt.*

Beweis. Die Aussage (ii) und (iii) sind einfach. Die Aussage (i) kann man mit dem Transformationssatz für Dichten (siehe Satz 4.20) nachrechnen unter Ausnutzung von Eigenschaften der mehrdimensionalen Normalverteilung. Alternativ folgt die Aussage auch im etwas allgemeineren Rahmen mit dem Satz von Basu; vgl. Example 2.18 in Shao (2003). □

Korollar 7.14. *Seien $X_1, \dots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $X_i \sim \mathcal{N}_{\mu,\sigma^2}$, $0 < \sigma^2 < \infty$. Dann gilt*

$$\frac{\sqrt{n}(\hat{\mu}(X) - \mu)}{\hat{\sigma}} = \frac{\sqrt{n}(\hat{\mu}(X) - \mu)}{\sigma}\frac{\sigma}{\hat{\sigma}} = \frac{\sqrt{n}(\hat{\mu}(X) - \mu)}{\sigma}\frac{1}{\sqrt{(n-1)\frac{\hat{\sigma}^2}{\sigma^2}\frac{1}{n}}} \sim t_{n-1}. \tag{7.10}$$

Beispiel 7.15 (Anwendung auf Konfidenzgrenzen für μ bei geschätzter Varianz)**.**

(i) Für $X_1, \dots, X_n$ unabhängig mit $X_i \sim \mathcal{N}_{\mu,\sigma^2}$ mit bekanntem $\sigma^2 \in (0, \infty)$ haben wir in Abschnitt 7.2.2 gezeigt, dass das Konfidenzintervall für μ zur Sicherheit $1 - \alpha$ durch

$$I(X) = (\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

gegeben ist, wobei $z_{\alpha/2}$ das obere $\alpha/2$ Quantil der Standardnormalverteilung ist, d.h.

$$z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2) \quad \text{bzw.} \quad \mathrm{P}(Z \geq z_{\alpha/2}) = \frac{\alpha}{2},\ Z \sim \mathcal{N}_{0,1}.$$

Ist $\sigma^2 \in (0, \infty)$ unbekannt und wird es mit $\hat{\sigma}^2(X)$ geschätzt dann kann man sich analog zu Abschnitt 7.2.2 überlegen, dass

$$\tilde{I}(X) = (\overline{X} - t_{n-1,\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \overline{X} + t_{n-1,\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}})$$

ein Konfidenzintervall für μ zur Sicherheit $1 - \alpha$ ist. Dabei bezeichnet $t_{n-1,\alpha/2}$ das obere $\alpha/2$ Quantil der t_{n-1}-Verteilung. Die einseitigen Konfidenzintervalle können auch analog zu Abschnitt 7.2.2 hergeleitet werden.

(ii) Seien $X_1, \dots, X_n$ unabhängig und identisch verteilt mit $\mathrm{E}[X_i] = \mu$ und $\mathrm{Var}[X_i] = \sigma^2 \in (0, \infty)$ dann hat das Konfidenzitervall $\tilde{I}(X)$ aus (i) die asymptotische Sicherheit $1 - \alpha$,

denn es gilt

$$\begin{aligned}\mathrm{P}(\mu \in \tilde{I}(X)) &= \mathrm{P}\Big(\frac{\sqrt{n}|\overline{X}_n - \mu|}{\hat{\sigma}_n} \leq t_{n-1,\alpha/2}\Big) \\ &= \mathrm{P}\Big(\frac{\sqrt{n}|\overline{X}_n - \mu|}{\sigma}\frac{\sigma}{\hat{\sigma}_n}\frac{z_{\alpha/2}}{t_{n-1,\alpha/2}} \leq z_{\alpha/2}\Big) \\ &\xrightarrow{n\to\infty} \mathrm{P}(|Z| \leq z_{\alpha/2}) = \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1-\alpha.\end{aligned}$$

Bemerkung 7.16. (i) Statt $t_{n-1,\alpha/2}$ kann man $z_{\alpha/2}$ (auch im allgmeinen Modell) verwenden. Asymptotisch liefert es dieselbe Sicherheit.

(ii) In der Praxis ist es üblich mit t-Quantilen zu arbeiten. Sie sind exakt im Normalverteilungsmodell und liefern konservativere (größere) Konfidenzintervalle, da $z_{\alpha/2} < t_{n,\alpha/2}$ für alle $n \in \mathbb{N}$ und $\alpha < 1/2$.

7.4 Das Maximum-Likelihood Prinzip

In diesem Abschnitt wiederholen wir die Maximum-Liekelihood Methode und starten mit einem Beispiel.

Beispiel 7.17. Wir betrachten eine einzige Beobachtung einer Zufallsvariablen mit Werten in $\{0,1,2\}$ mit Verteilung P_θ, $\theta \in \{\theta_0, \theta_1\}$, wobei die Wahrscheinlichkeitsgewichte in der folgenden Tabelle angegeben sind:

	$x=0$	$x=1$	$x=2$
$\theta=\theta_0$	0.7	0.2	0.1
$\theta=\theta_1$	0.2	0.3	0.5

Wenn $X=0$ beobachtet wird, dann ist es plausibler, dass die Beobachtung von der Verteilung P_{θ_0} stammt, weil $P_{\theta_0}(\{0\}) > P_{\theta_1}(\{0\})$. Man würde in diesem Fall schätzen, dass θ durch θ_0 gegeben ist. In dem Fall einer Beobachtung $X=1$ oder $X=2$ ist umgekehrt $\theta=\theta_1$ plausibler. Folgende Schätzfunktion bietet sich also an

$$\hat{\theta}(X) = \begin{cases}\theta_0 & : X = 0, \\ \theta_1 & : X \neq 0.\end{cases}$$

Das Beispiel kann leicht auf den Fall von diskreten Verteilungen P_θ mit $\theta \in \Theta \subset \mathbb{R}^k$ verallgemeinern. Wenn $X = x$ beobachtet wird, dann ist θ_1 plausibler als θ_2 genau dann wenn $P_{\theta_1}(\{x\}) > P_{\theta_2}(\{x\})$. Als Schätzer $\hat{\theta}$ würde man also das $\theta \in \Theta$ wählen, das $\theta \mapsto P_\theta(\{x\})$ maximiert, wenn es existiert.

Im Folgenden sei $X = (X_1, \ldots, X_n) \in \mathcal{X}$, $X \sim P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Wir nehmen an, dass die Dichten bzw. Wahrscheinlichkeitsgewichte von P_θ durch f_θ gegeben sind.

Definition 7.18 (Likelihood Funktion, Maximum-Likelihood Schätzer)**.**

(i) Für jedes $x \in \mathcal{X}$ heißt die Funktion

$$L(\cdot, x) \mapsto L(\theta, x) = f_\theta(x) \in \mathbb{R}$$

die *Likelihood-Funktion* bei Beobachtung $x \in \mathcal{X}$.

(ii) Wir nennen $\hat{\theta} \in \Theta$ die Maximum-Likelihood Schätzung (ML-Schätzung) für θ falls gilt

$$L(\hat{\theta}, x) = \sup_{\theta \in \Theta} L(\theta, x).$$

Fasst man $\hat{\theta}$ als Funktion von X auf, dann nennt man $\hat{\theta} = \hat{\theta}(X)$ *ML-Schätzfunktion bzw. ML-Schätzer* für θ.

Bemerkung 7.19 (Konkrete Bestimmung einer ML-Schätzung)**.** Da der Logarithmus eine streng monoton wachsende Funktion ist, kann man anstelle von $L(\cdot, x)$ äquivalent auch $\log L(\cdot, x)$ maximieren. Die letztere Funktion nennt man *Log-Likelhood-Funktion.* Das ist besonders dann nützlich, wenn $L(\cdot, x)$ eine Produktform hat. Dies gilt wenn $X_1, \ldots, X_n$ unabhängig sind.

Ist $L(\theta, x)$ bzw. $\log L(\theta, x)$ differenzierbar in θ auf $\Theta_0 \subset \Theta$, Θ_0 offen, so kann man oft (aber nicht immer) die ML-Schätzung durch explizites Lösen der *Likelihood-Gleichung*

$$\frac{\partial}{\partial \theta} L(\theta, x) = 0$$

bzw. der *Log-Likelihood-Gleichung*

$$\frac{\partial}{\partial \theta} \log L(\theta, x) = 0$$

bestimmen. Natürlich muss dann noch geprüft werden (mit 2. Ableitungen beispielsweise), ob es sich bei den Lösungen wirklich um Maxima handelt, bzw. es muss das globale Maximum bestimmt werden.

Wir schauen uns nun einige Beispiele an.

Beispiel 7.20 (Bernoulli Verteilung)**.** Seien $X_1, \ldots, X_n$ unabhängig und identisch verteilt mit $X_i \sim \mathrm{Ber}_\theta$, $\theta \in (0, 1)$. Es gilt

$$L(\theta, x) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i},$$

$$\log L(\theta, x) = \sum_{i=1}^{n} x_i \log \theta + (n - \sum_{i=1}^{n} x_i) \log(1-\theta).$$

Die Log-Likelihood-Gleichung ist

$$0 = \frac{\partial}{\partial \theta} \log L(\theta, x)\Big|_{\theta=\hat{\theta}} = \frac{1}{\hat{\theta}} \sum_{i=1}^{n} x_i - \frac{1}{1-\hat{\theta}} (n - \sum_{i=1}^{n} x_i).$$

Die Gleichung ist äquivalent zu

$$\sum_{i=1}^{n} x_i - \hat{\theta} \sum_{i=1}^{n} x_i = \hat{\theta} n - \hat{\theta} \sum_{i=1}^{n} x_i$$

und wir erhalten

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}.$$

Für die zweite Ableitung gilt

$$\frac{\partial^2}{\partial^2 \theta} \log L(\theta, x)\Big|_{\theta=\hat{\theta}} = -\frac{1}{\hat{\theta}^2} \sum_{i=1}^{n} x_i - \frac{1}{(1-\hat{\theta})^2}(n - \sum_{i=1}^{n} x_i) < 0.$$

Also ist $\hat{\theta} = \bar{x}$ Maximum wenn $\bar{x} \in (0, 1)$.

Ist $\bar{x} = 0$, so gilt $\sup_{\theta \in (0,1)} \log L(\theta, \underline{0}) = \sup_{\theta \in (0,1)} n \log(1-\theta)$. Da $\log(1-\theta)$ fallend in θ ist, ist $\hat{\theta} = 0 = \frac{1}{n} \sum_{i=1}^{n} 0$.

Ist $\bar{x} = 1$, so gilt $\sup_{\theta \in (0,1)} \log L(\theta, \underline{1}) = \sup_{\theta \in (0,1)} n \log \theta$. Da $\log \theta$ wachsend in θ ist, ist $\hat{\theta} = 1 = \frac{1}{n} \sum_{i=1}^{n} 1$.

Insgesamt ist also $\hat{\theta}(x) = \bar{x}$ die ML-Schätzung für θ.

Beispiel 7.21 (Normalverteilung)**.** Seien $X_1, \ldots, X_n$ unabhängig und identisch verteilt mit $X_i \sim \mathcal{N}_{\mu,\sigma^2}$, $\theta = (\mu, \sigma^2)$

Es gilt

$$\begin{aligned} L(\theta, x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\Big) \end{aligned}$$

und

$$\log L(\theta, x) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}.$$

Die log-Likelhood Gleichung besteht aus den Gleichungen (mit $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$)

$$0 = \frac{\partial}{\partial \mu} \log L(\theta, x)\Big|_{\hat{\theta}} = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})}{\hat{\sigma}^2}$$

und (wir leiten hier nach σ^2 ab nicht nach σ)

$$0 = \frac{\partial}{\partial \sigma^2} \log L(\theta, x)\Big|_{\hat{\theta}} = -\frac{n}{2\hat{\sigma}^2} + \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{2\hat{\sigma}^4}.$$

Auf $\Theta = \mathbb{R} \times (0, \infty)$ erhalten wir aus der ersten Gleichung die Lösung

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

und einsetzen in die zweite und Auflösen nach $\hat{\sigma}^2$ liefert

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Die Menge $\Theta = \mathbb{R} \times (0, \infty)$ ist offen und für $\sigma^2 \to 0$, sowie für $\|\theta\| \to \infty$ gilt $L(\theta, x) \to 0$, das Maximum wird also im inneren angenommen.

Die Hessematrix ist gegeben durch

$$\frac{\partial^2}{\partial\theta\partial\theta^\mathsf{T}} \log L(\theta, x) = -\begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) \\ \frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) & -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2 \end{pmatrix}$$

und an der Stelle $(\hat{\mu}, \hat{\sigma}^2)$ durch

$$\frac{\partial^2}{\partial\theta\partial\theta^\mathsf{T}} \log L(\theta, x)\Big|_{\theta=(\hat{\mu},\hat{\sigma}^2)} = -\begin{pmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4}. \end{pmatrix}$$

Die Matrix ist negativ definit und somit ist $\hat{\theta}(X) = (\hat{\mu}(X), \hat{\sigma}^2(X))$ mit

$$\hat{\mu}(X) = \overline{X} \quad \text{und} \quad \hat{\sigma}^2(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

eindeutiger ML-Schätzer.

Beispiel 7.22 (Gleichverteilung auf $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$)**.** Seien $X_1, \ldots, X_n$ unabhängig und identisch verteilt mit $X_i \sim \mathcal{U}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\theta \in \Theta = \mathbb{R}$. Es gilt

$$\begin{aligned} L(\theta, x) &= \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \mathbb{1}_{(\theta-\frac{1}{2}, \theta+\frac{1}{2})}(x_i) \\ &= \begin{cases} 1 & : \theta - \frac{1}{2} < x_i < \theta + \frac{1}{2} \ \forall i \\ 0 & : \text{ sonst,} \end{cases} \\ &= \mathbb{1}_{(x_{(n)}-\frac{1}{2}, x_{(1)}+\frac{1}{2})}(\theta). \end{aligned}$$

Dabei ist $x_{(1)} = \min\{x_1, \ldots, x_n\}$ und $x_{(n)} = \max\{x_1, \ldots, x_n\}$.

Für jede Schätzung $\hat{\theta} \in (x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2})$ gilt $L(\hat{\theta}, x) = 1 = \sup_{\theta\in\mathbb{R}} L(\theta, x)$. Jedes $\hat{\theta} \in (x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2})$ ist also eine ML Schätzung, die aber nicht eindeutig ist.

In der folgenden Bemerkung geben wir ohne Beweis eine untere Schranke für die Varianz von Schätzern in Modellen die bestimmte Regularitätsbedingungen erfüllen. Für Details und weitere Diskussion der Ungleichung siehe z.B. Abschnitt 4.3 in Czado and Schmidt (2011).

Bemerkung 7.23 (Informationsungleichung von Cramér Rao)**.** Für $\vartheta = h(\theta)$ sei $\hat{\vartheta}(X)$ ein Schätzer mit

$$m(\theta) = \mathrm{E}_\theta[\hat{\vartheta}(X)].$$

Wir definieren die *Fisher Information* durch

$$I(\theta) = \mathrm{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}\log f_\theta(X)\Big)^2\Big].$$

Unter bestimmten Regularitätsvoraussetzungen und falls $0 < I(\theta) < \infty$ ist gilt die *Cramér-Rao Ungleichung*

$$\mathrm{Var}_\theta[\hat{\vartheta}(X)] \geq \frac{(m'(\theta))^2}{I(\theta)}.$$

Schätzer, für die asymptotisch die Gleichheit gilt, heißen *effizient.* ML-Schätzer sind unter relativ milden Bedingungen an das Modell effizient.

7.5 Elemente der Testtheorie

Es sei ein Beobachtungsmodell $\{(\mathcal{X}, \mathcal{B}, P_\theta) : \theta \in \Theta\}$ gegeben. Ferner sei $\Theta = \Theta_0 \cup \Theta_1$ eine disjunkte Vereinigung. Aufgrund von gemachten Beobachtungen möchte man entscheiden, ob das wahre θ nun in Θ_0 oder in Θ_1 liegt. Wir betrachten folgende Hypothesen

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \quad \textit{(Nullhypothese),} \\ H_1 &: \theta \in \Theta_1 \quad \textit{(Alternative).} \end{aligned}$$

Eine Funktion d mit $d : \mathcal{X} \to \{0, 1\}$ heißt *Test für H_0 gegen H_1* wenn sie messbar ist, und wie folgt interpretiert werden kann:

$$\begin{aligned} d(x) = 1 \quad &\Longleftrightarrow \quad \textit{Verwerfen der Nullhypothese } H_0 \textit{ bzw. Annahme der Alternative } H_1 \\ d(x) = 0 \quad &\Longleftrightarrow \quad \textit{Annahme der Nullhypothese } H_0. \end{aligned}$$

Insbesondere gibt es einen *kritischen Bereich* $K \in \mathcal{B}$ mit $d = \mathbb{1}_K$.

(Es gibt allgemeinere *randomisiserte* Tests mit Werten in dem Intervall $[0, 1]$, wir werden solche Tests hier jedoch nicht betrachten.)

Beim Testen können im Prinzip folgende Fehler enstehen:

	$d(x) = 0$	$d(x) = 1$
$\theta \in \Theta_0$	*kein Fehler*	*Fehler 1. Art*
$\theta \in \Theta_1$	*Fehler 2. Art*	*kein Fehler*

Wir bezeichnen die entsprechenden Irrtumswahrscheinlichkeiten mit

$$\alpha_d(\theta) = P_\theta(d(X) = 1) \qquad \text{für } \theta \in \Theta_0 \qquad \text{(Wahrsch. für Fehler 1. Art)}$$
$$\beta_d(\theta) = 1 - P_\theta(d(X) = 1) = P_\theta(d(X) = 0) \qquad \text{für } \theta \in \Theta_1 \qquad \text{(Wahrsch. für Fehler 2. Art)}$$

Definition 7.24 (Gütefunktion)**.** Die Funktion $G_d : \Theta \to [0, 1]$ mit

$$\theta \mapsto G_d(\theta) = P_\theta(d(X) = 1) = E_\theta[d(X)]$$

heißt *Gütefunktion der Tests d (auch Schärfe, Trennschärfe, Power).* Dabei gilt

$$\alpha_d(\theta) = G_d(\theta) \quad \text{für } \theta \in \Theta_0$$
$$\beta_d(\theta) = 1 - G_d(\theta) \quad \text{für } \theta \in \Theta_1.$$

Typischerweise wird die Hypothese H_0 so gewählt, dass man sie durch Experimente ablehnen will. Man möchte also in erster Linie den Fehler 1. Art kontrollieren. Dieser Fehler soll eine vorgegebene Grenze nicht überschreiten. Unter solchen Tests sucht man nach Tests mit möglichst kleiner Fehlerwahrscheinlichkeit 2. Art.

Definition 7.25. (i) Ein Test d für H_0 gegen H_1 heißt *Test zum Signifikanzniveau oder einfach Niveau* α, wenn $\alpha_d \leq \alpha$. Mit anderen Worten es gilt $G_d(\theta) \leq \alpha$ für alle $\theta \in \Theta_0$.

(ii) Ein Test d^* zum Niveau α heißt *trennschärfster Test* (engl. uniformly most powerfull (UMP)) für H_0 gegen H_1 wenn für jedes $\theta \in \Theta_1$ gilt

$$G_{d^*}(\theta) = \sup\{G_d(\theta) : d \text{ Niveau } \alpha \text{ Test für } H_0 \text{ gegen } H_1\}.$$

(iii) Ein Niveau α Test d heißt *unverfälscht* zum Niveau α, wenn $\beta_d \leq 1 - \alpha$, d.h. wenn $G_d(\theta) \geq \alpha$ für alle $\theta \in \Theta_1$.

Lemma 7.26. *Ein trennschärfster Niveau α Test d ist unverfälscht.*

Beweis. Sei d' eine Ber_α verteilt Zufallsvariable (unabhängig von der Beobachtung und von dem Parameter θ).

Dann gilt $G_{d'}(\theta) = E_\theta[d'] = \alpha$ für alle $\theta \in \Theta$. Also ist d' ein Niveau α Test.

Ist d trennschärfster Niveau α Test, so gilt für alle $\theta \in \Theta_1$

$$G_d(\theta) \geq G_{d'}(\theta) = \alpha.$$

Damit ist d unverfälscht. □

Ein klassisches Ziel ist es einen trennschärfsten Niveau α Test unter allen Niveau α Tests zu finden. Falls kein trennschärfster existiert, dann sucht man nach dem trennschärfsten unter den unverfälschten Niveau α Tests.

Beispiel 7.27 (Einseitiger Gauß-Test)**.** Sei $X = (X_1, \dots, X_n)$, wobei $X_1, \dots, X_n$ unabhängig und identisch verteilt sind mit $X_i \sim \mathcal{N}_{\mu,\sigma^2}$, $\theta = \mu \in \mathbb{R}$, $\sigma^2 > 0$ bekannt.

Wir betrachten einen *einseitigen Gauß-Test.* Für $\mu_0 \in \mathbb{R}$ seien die Nullhypothese und die Alternative gegeben durch

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0.$$

Es ist plausibel die Nullhypothese abzulehnen, wenn der Mittelwert $\overline{X}$ zu groß im Vergleich zu μ_0 ist. Wir machen den folgenden Ansatz

$$d(X) = \mathbb{1}_{(c,\infty)}(\overline{X}).$$

Für solchen Test gilt

$$\begin{aligned} G_d(\mu) &= P_\mu(d(X) = 1) = P_\mu(\overline{X} > c) = P_\mu\Big(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > \frac{c - \mu}{\sigma/\sqrt{n}}\Big) \\ &= 1 - \Phi\Big(\frac{c - \mu}{\sigma/\sqrt{n}}\Big) = \Phi\Big(\frac{\mu - c}{\sigma/\sqrt{n}}\Big). \end{aligned}$$

Die rechte Seite ist wachsend in μ und fallend in c. Für ein Test zum Niveau α muss gelten

$$\alpha \geq \sup_{\mu \leq \mu_0} G_d(\mu) = G_d(\mu_0) = \Phi\Big(\frac{\mu_0 - c}{\sigma/\sqrt{n}}\Big).$$

Das ist äquivalent zu

$$1 - \alpha \leq \Phi\Big(\frac{c - \mu_0}{\sigma/\sqrt{n}}\Big),$$

bzw.

$$z_\alpha = \Phi^{-1}(1 - \alpha) \leq \frac{c - \mu_0}{\sigma/\sqrt{n}}.$$

Insbesondere ist für jedes $c \geq \mu_0 + \frac{z_\alpha \sigma}{\sqrt{n}}$ der Test $d(X) = \mathbb{1}_{(c,\infty)}(\overline{X})$ ein Niveau α Test. Das muss es auch für den trennschärfsten Test d^* gelten. Außerdem muss für d^* auch

$$G_{d^*}(\mu) = \sup\Big\{G_d(\mu) : d(X) = \mathbb{1}_{(c,\infty)}(\overline{X}), c \geq \mu_0 + \frac{z_\alpha \sigma}{\sqrt{n}}\Big\}.$$

Dies ist für $d^*(X) = \mathbb{1}_{(c^*,\infty)}(\overline{X})$ mit $c^* = \mu_0 + \frac{z_\alpha \sigma}{\sqrt{n}}$ der Fall, denn für $c > c^*$ gilt

$$G_{d^*}(\mu) = \Phi\Big(\frac{\mu - c^*}{\sigma/\sqrt{n}}\Big) > \Phi\Big(\frac{\mu - c}{\sigma/\sqrt{n}}\Big) = G_d(\mu)$$

für alle μ und insbesondere für alle $\mu > \mu_0$.

Insgesamt ist also $d^*(X) = \mathbb{1}_{(c^*,\infty)}(\overline{X})$ mit $c^* = \mu_0 + \frac{z_\alpha \sigma}{\sqrt{n}}$ trennschärfster Niveau α Test in der Klasse der Niveau α Tests der Gestalt $d(X) = \mathbb{1}_{(c,\infty)}(\overline{X})$.

Die Gütefunktion ist gegeben durch

$$G_{d^*}(\mu) = \Phi\Big(\frac{\mu - c^*}{\sigma/\sqrt{n}}\Big) = \Phi\Big(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - z_\alpha\Big).$$

Die maximalen Fehlerwahrscheinlichkeiten sind

$$\sup_{\mu\le\mu_0} \alpha_{d^*}(\mu) = \sup_{\mu\le\mu_0} G_{d^*}(\mu) = \sup_{\mu\le\mu_0} \Phi\Big(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - z_\alpha\Big) = \Phi(-z_\alpha) = \alpha$$

und

$$\begin{aligned}\sup_{\mu\le\mu_0} \beta_{d^*}(\mu) &= \sup_{\mu>\mu_0} (1 - G_{d^*}(\mu)) = 1 - \inf_{\mu>\mu_0} G_{d^*}(\mu)\\ &= 1 - G_{d^*}(\mu_0) = 1 - \Phi(-z_\alpha) = 1 - \alpha.\end{aligned}$$

Man kann sich analog überlegen, dass für den einseitigen Gauß-Test mit $H_0 : \mu \ge \mu_0$ vs. $H_1 : \mu < \mu_0$ der Test d^* mit

$$d^*(X) = \mathbb{1}_{(-\infty,c^*)}(\overline{X}), \quad c^* = \mu_0 - \frac{z_\alpha\sigma}{\sqrt{n}}$$

trennschärfster Test unter allen Niveau α Tests der Form $d(X) = \mathbb{1}_{(-\infty,c)}(\overline{X})$ ist.

Der zweiseitige Gauß-Test mit $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \ne \mu_0$ wird in der Übung behandelt. Man kann zeigen, dass unter allen Niveau α Tests der Form $d(X) = \mathbb{1}_{(\mu_0-c,\mu_0+c)}(\overline{X})$ der Test

$$d^*(X) = \mathbb{1}_{((\mu_0-c^*,\mu_0+c^*)}(\overline{X}), \quad c^* = \frac{z_{\alpha/2}\sigma}{\sqrt{n}},\ z_{\alpha/2} = \Phi^{-1}(1-\alpha/2),$$

trennschärfster Niveau α Test ist.

Beispiel 7.28 (t-Test)**.** Sei $X = (X_1, \dots, X_n)$ wobei $X_1, \dots, X_n$ unabhängig und identisch verteilt sind mit unbekannten $\mu = \mathrm{E}[X_i] \in \mathbb{R}$ und $\sigma^2 \in (0, \infty)$. Wir wollen für ein $\mu_0 \in \mathbb{R}$

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \ne \mu_0$$

testen. Wir haben in Beispiel 7.15 gesehen, dass

$$\tilde{I}(X) = \Big(\overline{X} - t_{n-1,\alpha/2}\frac{\hat\sigma}{\sqrt{n}}, \overline{X} + t_{n-1,\alpha/2}\frac{\hat\sigma}{\sqrt{n}}\Big)$$

ein (asymptotisches) Konfidenzintervall für μ zur Sicherheit $1 - \alpha$ ist. Dabei ist

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

die Stichprobenvarianz und

$$t_{n-1,\alpha/2} = F_{n-1}^{-1}(1 - \alpha/2)$$

das obere $\alpha/2$-Quantil der t_{n-1}-Verteilung.

Das Intervall $\tilde{I}(X)$ ist „exakt“ wenn $X_i \sim \mathcal{N}_{\mu,\sigma^2}$ und approximativ sonst. Wir machen folgenden Ansatz

$$d(X) = 1 \iff \mu_0 \notin \tilde{I}(X).$$

Dann gilt für den Fehler 1. Art

$$P_{\mu_0}(d(X) = 1) = P_{\mu_0}(\mu_0 \notin \tilde{I}(X)) \leq \alpha$$

Also ist d ein Test für $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ zum Niveau α.

Beispiel 7.29 (Fishers exakter Test auf Unabhängigkeit)**.** Seien A und B zwei Ereignisse auf einem Wahscheinlichkeitsraum. Wir nehmen an, dass wir n Versuche beobachten und zählen wie oft das Ereignis jeweils in den folgenden Mengen landet

$$A \cap B,\ A \cap B^C,\ A^C \cap B,\ A^C \cap B^C.$$

Man kann die Ergebnisse in einer *Kontingenztafel* zusammenfassen

	A	A^C	total
B	X_{11}	X_{12}	n_1
B^C	X_{21}	X_{22}	n_2
total	m_1	m_2	n

Wir betrachten den folgenden Test

$$H_0 : A \text{ und } B \text{ sind unabhängig} \quad \text{vs.} \quad H_1 : A \text{ und } B \text{ sind abhängig.}$$

Zunächst bestimmen wir die Verteilung von X_{11} gegeben

- $X_{11} + X_{12} =$ Gesamtanzahl der Versuche in B,
- $X_{11} + X_{21} =$ Gesamtanzahl der Versuche in A.

Dafür interpretieren wir X_{11} als eine Zufallsvariable in einem Urnenexperiment:

- Sei n die Gesamtanzahl der Kugeln in der Urne.
- Davon sind $n_1 = X_{11} + X_{12}$ markiert,
- Es werden $m_1 = X_{11} + X_{21}$ Kugeln gezogen, wovon X_{11} markiert sind.

Unter der Hypothese H_0 ist X_{11} hypergeometrisch verteilt:

$$P_0(X_{11} = k) = \binom{n_1}{k}\binom{n - n_1}{m_1 - k} \Big/ \binom{n}{m_1}.$$

Liegen konkrete Beobachtungen vor so kann man obige Wahrscheinlichkeit (p-Wert) ausrechnen. Ist diese zu klein, dann lehnt man die Nullhypothese ab. Für ein konkretes Beispiel des Tests auf Unabhängigkeit verweisen wir auf Abschnitt 21 in Kersting and Wakolbinger (2010).

Bemerkung 7.30 (p-Wert)**.** In Anwendungen interessiert man sich beim Testen oft für den p-Wert, bzw. das beobachtete Signifikanzniveau. Es ist das kleinste Signifikanzniveau zu dem die Hypothese noch abgelehnt wird. In Statistikpaketen wie z.B. R geben Tests standardmäßig den p-Wert der Beobachtung mit aus. Man lehnt die Hypothese dann ab, wenn der p-Wert unter dem vorgegebenen Signifikanzniveau α liegt.

Literaturverzeichnis

Czado, C. and Schmidt, T.: 2011, *Mathematische Statistik.*, Berlin: Springer.

Eichelsbacher, P. and Löwe, M.: 2014, 90 Jahre Lindeberg-Methode, *Math. Semesterber.* **61**(1), 7–34.
URL: *http://dx.doi.org/10.1007/s00591-013-0118-9*

Feller, W.: 1968, *An introduction to probability theory and its applications. Vol. I*, Third edition, John Wiley & Sons, Inc., New York-London-Sydney.

Gut, A.: 2013, *Probability: a graduate course*, Springer Texts in Statistics, second edn, Springer, New York.
URL: *http://dx.doi.org/10.1007/978-1-4614-4708-5*

Johnson, N. L., Kotz, S. and Kemp, A. W.: 1992, *Univariate discrete distributions*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, second edn, John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication.

Kersting, G. and Wakolbinger, A.: 2010, *Elementare Stochastik. 2nd revised ed.*, Mathematik Kompakt. Basel: Birkhäuser.

Klenke, A.: 2013, *Wahrscheinlichkeitstheorie.*, 3nd revised ed. edn, Springer.

Pfanzagl, J.: 1991, *Elementare Wahrscheinlichkeitsrechnung.*, 2., überarb. und erweit. aufl. edn, Berlin etc.: de Gruyter.

Shao, J.: 2003, *Mathematical statistics*, Springer Texts in Statistics, second edn, Springer-Verlag, New York.
URL: *http://dx.doi.org/10.1007/b97553*