Einführung in die Wahrscheinlichkeitstheorie

Andreas Eberle

8. Februar 2014

Inhaltsverzeichnis

Teil I

Allgemeine Zustandsräume

Kapitel 1

Stetige und Allgemeine Modelle

1.1 Unendliche Kombinationen von Ereignissen

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Ist $(A_n)_{n\in\mathbb{N}}$ eine Folge von bzgl. P unabhängigen Ereignissen, $A_n \in \mathcal{A}$ mit fester Wahrscheinlichkeit

$$P[A_n] = p \in [0, 1]$$

und

$$S_n(\omega) = \sum_{i=1}^{n} I_{A_i}(\omega) = |\{1 \leq i \leq n : \omega \in A_i\}|$$

die Anzahl der Ereignisse unter den ersten n, die eintreten, dann ist S_n binomialverteilt mit den Parametern n und p. Für die relative Häufigkeit $\frac{S_n}{n}$ der Ereignisse A_i gilt die Bernstein-Chernoff-Ungleichung

$$P\left[\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right] \leq 2 \cdot e^{-2\varepsilon^2 n}, \tag{1.1.1}$$

d.h. die Verteilung von $\frac{S_n}{n}$ konzentriert sich für $n \to \infty$ sehr rasch in der Nähe von p, siehe Abschnitt **??**. Insbesondere ergibt sich ein Spezialfall des schwachen Gesetzes der großen Zahlen: die Folge der Zufallsvariablen $\frac{S_n}{n}$ konvergiert P-stochastisch gegen p, d.h.

$$P\left[\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right] \overset{n\to\infty}{\to} 0 \text{ für alle } \varepsilon > 0.$$

Definition (Nullmengen und fast sichere Ereignisse). *(1). Eine P-***Nullmenge** *ist ein Ereignis $A \in \mathcal{A}$ mit $P[A] = 0$.*

*(2). Ein Ereignis $A \in \mathcal{A}$ tritt P-***fast sicher** *bzw. für P-***fast alle** *$\omega \in \Omega$ ein, falls $P[A] = 1$ gilt, d.h. falls A^C eine P-Nullmenge ist.*

Wir wollen nun Methoden entwickeln, die es uns ermöglichen, zu zeigen, dass aus (1.1.1) sogar

$$\lim_{n\to\infty} \frac{S_n(\omega)}{n} = p \qquad \text{für } P\text{-fast alle } \omega \in \Omega \tag{1.1.2}$$

folgt. Das relevante Ereignis

$$L := \left\{ \omega \in \Omega : \lim_{n\to\infty} \frac{S_n(\omega)}{n} = p \right\}$$

lässt sich offensichtlich nicht durch endlich viele der A_i beschreiben.

Seien nun allgemein $A_1, A_2, \ldots \in \mathcal{A}$ beliebige Ereignisse. Uns interessieren zusammengesetzte Ereignisse wie z.B.

$\bigcup_{n=1}^{\infty} A_n$ („Eines der A_n tritt ein“)

$\bigcap_{n=1}^{\infty} A_n$ („Alle der A_n treten ein“)

$\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n = \{\omega \in \Omega : \forall m \quad \exists n \geq m : \omega \in A_n\}$ („Unendlich viele der A_n treten ein“ oder „A_n tritt immer mal wieder ein“)

$\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n = \{\omega \in \Omega : \exists m \quad \forall n \geq m : \omega \in A_n\}$ („A_n tritt schließlich ein“)

Aufgrund der Eigenschaften einer σ-Algebra liegen alle diese Mengen wieder in $\mathcal{A}$. Das Ereignis L lässt sich wie folgt als abzählbare Kombination der A_i ausdrücken:

$$\begin{aligned}
\omega \in L \quad &\iff \quad \lim_{n\to\infty} \frac{S_n}{n} = p \\
&\iff \quad \forall \varepsilon \in \mathbb{Q}_+ : \left| \frac{S_n}{n} - p \right| \leq \varepsilon \text{ schließlich} \\
&\iff \quad \forall \varepsilon \in \mathbb{Q}_+ \quad \exists m \in \mathbb{N} \quad \forall n \geq m : \left| \frac{S_n}{n} - p \right| \leq \varepsilon
\end{aligned}$$

Somit gilt

$$\begin{aligned}
L &= \bigcap_{\varepsilon \in \mathbb{Q}_+} \left\{ \left| \frac{S_n}{n} - p \right| \leq \varepsilon \text{ schließlich} \right\} \\
&= \bigcap_{\varepsilon \in \mathbb{Q}_+} \bigcup_{m \in \mathbb{N}} \bigcap_{n \geq m} \left\{ \left| \frac{S_n}{n} - p \right| \leq \varepsilon \right\}.
\end{aligned}$$

Um Wahrscheinlichkeiten von solchen Ereignissen berechnen zu können, ist es wesentlich, dass eine Wahrscheinlichkeitsverteilung P nicht nur endlich additiv, sondern sogar σ-additiv ist.

1.1.1 Konsequenzen der σ-Additivität

Der folgende Satz gibt eine alternative Charakterisierung der σ-Additivität:

Satz 1.1 (**σ-Additivität und monotone Stetigkeit**). *Sei $\mathcal{A}$ eine σ-Algebra und $P : \mathcal{A} \to [0, \infty]$ additiv, d.h.*

$$A \cap B = \emptyset \;\Rightarrow\; P[A \cup B] = P[A] + P[B].$$

(i) P ist σ-additiv genau dann, wenn:

$$A_1 \subseteq A_2 \subseteq \ldots \;\Rightarrow\; P\left[\bigcup_{n=1}^{\infty} A_n\right] = \lim_{n\to\infty} P[A_n]$$

(ii) Gilt $P[\Omega] < \infty$, dann ist dies auch äquivalent zu:

$$A_1 \supseteq A_2 \supseteq \ldots \;\Rightarrow\; P\left[\bigcap_{n=1}^{\infty} A_n\right] = \lim_{n\to\infty} P[A_n]$$

Beweis. (i) Sei P σ-additiv und $A_1 \subseteq A_2 \subseteq \ldots$. Die Mengen $B_1 := A_1$, $B_2 := A_2 \backslash A_1$, $B_3 := A_3 \backslash A_2$, $\ldots$ sind disjunkt mit

$$\bigcup_{i=1}^{n} B_i = \bigcup_{i=1}^{n} A_i = A_n \quad \text{und} \quad \bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i.$$

Also gilt:

$$\begin{aligned} P\left[\bigcup_{i=1}^{\infty} A_i\right] &= P\left[\bigcup_{i=1}^{\infty} B_i\right] \\ &\overset{\sigma-add.}{=} \sum_{i=1}^{\infty} P[B_i] \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} P[B_i] \\ &= \lim_{n\to\infty} P\left[\bigcup_{i=1}^{n} B_i\right] \\ &= \lim_{n\to\infty} P[A_n]. \end{aligned}$$

Der Beweis der umgekehrten Implikation wird dem Leser als Übungsaufgabe überlassen.

(ii) Gilt $P[\Omega] < \infty$, dann folgt

$$P\left[\bigcap_{i=1}^{\infty} A_i\right] \;=\; P\left[\left(\bigcup_{i=1}^{\infty} A_i^C\right)^C\right] \;=\; P\left[\Omega\right] - P\left[\bigcup_{i=1}^{\infty} A_i^C\right].$$

Die Behauptung folgt nun aus (i).

□

Ab jetzt setzen wir wieder voraus, dass P eine Wahrscheinlichkeitsverteilung ist. Eine weitere Folgerung aus der σ-Additivität ist:

Satz 1.2 (σ-Subadditivität). *Für beliebige Ereignisse $A_1, A_2, \ldots \in \mathcal{A}$ gilt:*

$$P\left[\bigcup_{n=1}^{\infty} A_n\right] \leq \sum_{n=1}^{\infty} P[A_n]$$

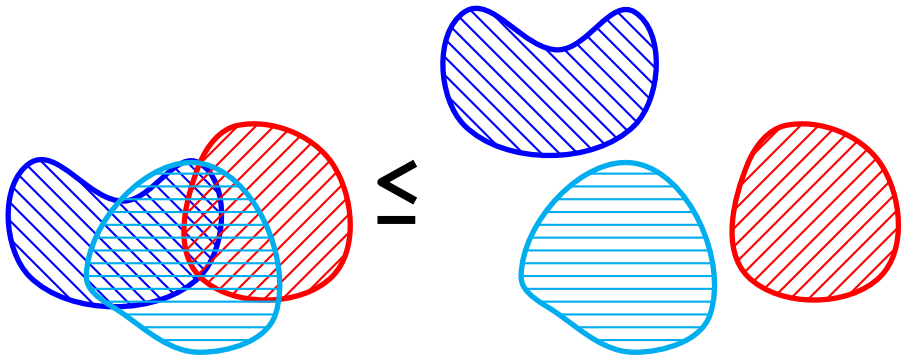


Abbildung 1.1: Darstellung von drei Mengen. Das Maß der Vereinigung von Mengen ist stets kleiner gleich als die Summe der Maße der einzelnen Mengen.

Beweis. Die Mengen

$$B_n = A_n \setminus (A_{n-1} \cup \cdots \cup A_1)$$

sind disjunkt mit $\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$. Also gilt:

$$P\left[\bigcup_{n=1}^{\infty} A_n\right] = P\left[\bigcup_{n=1}^{\infty} B_n\right] = \sum_{n=1}^{\infty} \underbrace{P[B_n]}_{\leq P[A_n]} \leq \sum_{n=1}^{\infty} P[A_n].$$

□

Bemerkung. Insbesondere ist eine Vereinigung von abzählbar vielen Nullmengen wieder eine Nullmenge.

1.1.2 Borel-Cantelli

Der folgende Satz spielt eine zentrale Rolle beim Beweis von Konvergenzaussagen für Zufallsvariablen:

Satz 1.3 (1. Borel - Cantelli - Lemma). *Für Ereignisse $A_1, A_2, \ldots \in \mathcal{A}$ mit*

$$\sum_{n=1}^{\infty} P[A_n] < \infty$$

gilt:

$$P[\text{„unendlich viele der } A_n \text{ treten ein“}] = P\left[\bigcap_m \bigcup_{n\geq m} A_n\right] = 0.$$

Beweis. Da die Folge $\bigcup_{n\geq m} A_n =: B_m$ von Ereignissen aus $\mathcal{A}$ monoton fallend ist, ergibt sich nach Satz 1.1 und 1.2:

$$\begin{aligned} P\left[\bigcap_m \bigcup_{n\geq m} A_n\right] &= P\left[\bigcap_m B_m\right] \\ &\overset{1.1}{=} \lim_{m\to\infty} P[B_m] \\ &= \lim_{m\to\infty} \underbrace{P\left[\bigcup_{n\geq m} A_n\right]}_{\overset{1.2}{\leq} \sum_{n=m}^{\infty} P[A_n]} \\ &\leq \liminf_{m\to\infty} \underbrace{\sum_{n=m}^{\infty} P[A_n]}_{\overset{m\to\infty}{\to} 0} = 0, \end{aligned}$$

da die Summe $\sum_{n=1}^{\infty} P[A_n]$ nach Voraussetzung konvergiert. □

Das erste Borel-Cantelli-Lemma besagt, dass mit Wahrscheinlichkeit 1 nur endlich viele der Ereignisse $A_n, n \in \mathbb{N}$ eintreten, falls $\sum P[A_n] < \infty$ gilt. Die Unabhängigkeit der Ereignisse ermöglicht die Umkehrung dieser Aussage. Es gilt sogar:

Satz 1.4 (2. Borel - Cantelli - Lemma). *Für unabhängige Ereignisse $A_1, A_2, \ldots \in \mathcal{A}$ mit*

$$\sum_{n=1}^{\infty} P[A_n] = \infty$$

gilt:

$$P[A_n \text{ unendlich oft}] = P\left[\bigcap_m \bigcup_{n\geq m} A_n\right] = 1$$

Bemerkung. Insbesondere ergibt sich ein ***0-1 Gesetz***:
Sind $A_1, A_2, \ldots \in \mathcal{A}$ unabhängige Ereignisse, dann beträgt die Wahrscheinlichkeit, dass unendlich viele der $A_n, n \in \mathbb{N}$, eintreten, entweder 0 oder 1 - je nachdem ob die Summe $\sum P[A_n]$ endlich oder unendlich ist.

Wir zeigen nun das zweite Borel-Cantelli-Lemma:

Beweis. Sind die Ereignisse $A_n, n \in \mathbb{N}$ unabhängig, so auch die Ereignisse A_n^C, siehe Lemma **??**. Zu zeigen ist:

$$P[A_n \text{ nur endlich oft}] = P\left[\bigcup_m \bigcap_{n\geq m} A_n^C\right] = 0$$

Nach Satz 1.1 gilt:

$$P\left[\bigcup_m \bigcap_{n\geq m} A_n^C\right] = \lim_{m\to\infty} P\left[\bigcap_{n\geq m} A_n^C\right] \tag{1.1.3}$$

Wegen der Unabhängigkeit der Ereignisse A_n^C erhalten wir zudem

$$\begin{aligned} P\left[\bigcap_{n\geq m} A_n^C\right] &\overset{\text{mon. Stetigkeit}}{=} \lim_{k\to\infty} P\left[\bigcap_{n=m}^{k} A_n^C\right] \\ &\overset{\text{unabh.}}{=} \lim_{k\to\infty} \prod_{n=m}^{k} \underbrace{P[A_n^C]}_{=1-P[A_n]\leq \exp(-P[A_n])} \\ &\leq \liminf_{k\to\infty} \prod_{n=m}^{k} e^{-P[A_n]} \\ &= \liminf_{k\to\infty} e^{-\sum_{n=m}^{k} P[A_n]} = 0, \end{aligned} \tag{1.1.4}$$

da $\lim_{k\to\infty} \sum_{n=m}^{k} P[A_n] = \sum_{n=m}^{\infty} P[A_n] = \infty$ nach Voraussetzung.
Aus 1.1.3 und 1.1.4 folgt die Behauptung. □

1.1.3 Ein starkes Gesetz der großen Zahlen

Mithilfe des 1. Borel-Cantelli-Lemmas können wir nun eine erste Version eines starken Gesetzes großer Zahlen beweisen. Sei $p \in [0, 1]$.

Satz 1.5 (Starkes Gesetz großer Zahlen I, Borel 1909, Hausdorff 1914, Cantelli 1917). *Sind $A_1, A_2, \ldots \in \mathcal{A}$ unabhängige Ereignisse mit Wahrscheinlichkeit $P[A_n] = p$ für alle $n \in \mathbb{N}$, dann gilt für $S_n = \sum_{i=1}^{n} I_{A_i}$:*

$$\underbrace{\lim_{n\to\infty} \frac{S_n(\omega)}{n}}_{\substack{\text{asymptotische relative Häufigkeit des Ereignisses}}} = \underbrace{p}_{\text{W'keit}} \quad \text{für } P\text{-fast alle } \omega \in \Omega$$

Beweis. Sei

$$L := \left\{ \omega \in \Omega : \frac{1}{n} S_n(\omega) \to p \text{ für } n \to \infty \right\}$$

Zu zeigen ist, dass $L^C \in \mathcal{A}$ mit $P[L^C] = 0$.
Wegen

$$\omega \in L^C \iff \frac{S_n(\omega)}{n} \not\to p \iff \exists \varepsilon \in \mathbb{Q}_+ : \left| \frac{S_n(\omega)}{n} - p \right| > \varepsilon \quad \text{unendlich oft}$$

gilt:

$$L^C = \bigcup_{\varepsilon \in \mathbb{Q}_+} \left\{ \left| \frac{S_n}{n} - p \right| > \varepsilon \quad \text{unendlich oft} \right\} = \bigcup_{\varepsilon \in \mathbb{Q}_+} \bigcap_{m} \bigcup_{n \geq m} \left\{ \left| \frac{S_n}{n} - p \right| > \varepsilon \right\} \in \mathcal{A}.$$

Zudem folgt aus der Bernstein-Chernoff-Abschätzung:

$$\sum_{n=1}^{\infty} P\left[\left| \frac{S_n}{n} - p \right| > \varepsilon \right] \leq \sum_{n=1}^{\infty} 2e^{-2n\varepsilon^2} < \infty$$

für alle $\varepsilon > 0$, also nach dem 1. Borel-Cantelli-Lemma:

$$P\left[\left| \frac{S_n}{n} - p \right| > \varepsilon \text{ unendlich oft} \right] = 0.$$

Also ist L^C eine Vereinigung von abzählbar vielen Nullmengen, und damit nach Satz 1.2 selbst eine Nullmenge. □

Das starke Gesetz großer Zahlen in obigem Sinn rechtfertigt nochmals im Nachhinein die empirische Interpretation der Wahrscheinlichkeit eines Ereignisses als asymptotische relative Häufigkeit bei unabhängigen Wiederholungen.

Beispiel (Random Walk/Irrfahrt). Wir betrachten einen Random Walk

$$Z_n = X_1 + X_2 + X_3 + \ldots + X_n \quad (n \in \mathbb{N})$$

mit unabhängigen identisch verteilten Inkrementen $X_i, i \in \mathbb{N}$, mit

$$P[X_i = 1] = p \quad \text{und} \quad P[X_i = -1] = 1 - p, \quad p \in (0,1) \text{ fest.}$$

Die Ereignisse $A_i := \{X_i = 1\}$ sind unabhängig mit $P[A_i] = p$ und es gilt:

$$X_i = I_{A_i} - I_{A_i^C} = 2I_{A_i} - 1,$$

also

$$Z_n = 2S_n - n, \quad \text{wobei} \quad S_n = \sum_{i=1}^{n} I_{A_i}.$$

Nach Satz 1.5 folgt:

$$\lim_{n\to\infty} \frac{Z_n}{n} = 2 \lim_{n\to\infty} \frac{S_n}{n} - 1 = 2p - 1 \qquad P\text{-fast sicher.}$$

Für $p \neq \frac{1}{2}$ wächst (bzw. fällt) Z_n also mit Wahrscheinlichkeit 1 asymptotisch linear (siehe Abbildung 1.2):

$$Z_n \sim (2p-1) \cdot n \quad P\text{-fast sicher}$$

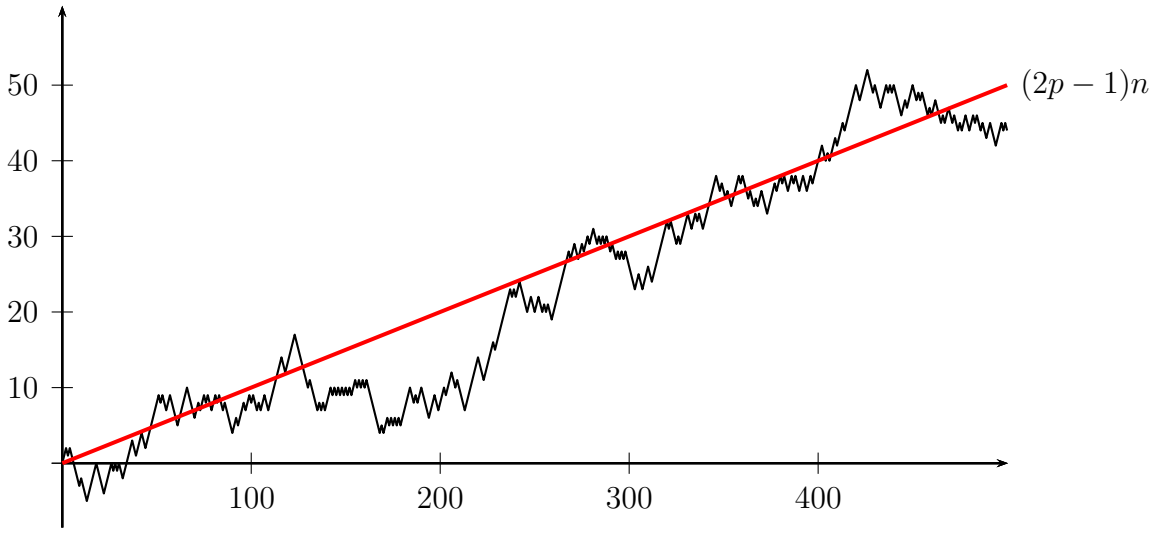


Abbildung 1.2: Random Walk mit Drift: $p = 0.55, n = 500$

Für $p = \frac{1}{2}$ dagegen wächst der Random Walk sublinear, d.h. $\frac{Z_n}{n} \to 0$ P-fast sicher. In diesem Fall liegt für hinreichend große n der Graph einer typischen Trajektorie $Z_n(\omega)$ in einem beliebig kleinen Sektor um die x-Achse (siehe Abbildung 1.3).

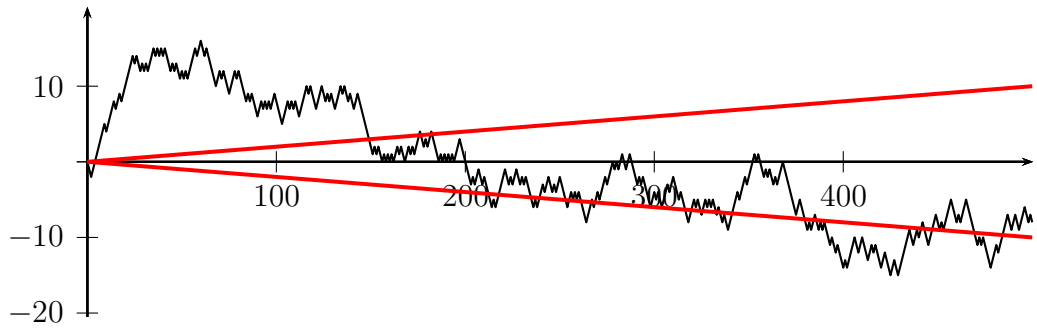


Abbildung 1.3: Random Walk ohne Drift: $p = 0.5, n = 500$

Eine viel präzisere Beschreibung der Asymptotik des Random Walks im Fall $p = 1/2$ liefert der **Satz vom iterierten Logarithmus**:

$$\limsup_{n\to\infty} \frac{Z_n}{\sqrt{n \log\log n}} = +1 \qquad P\text{-fast sicher,}$$

$$\liminf_{n\to\infty} \frac{Z_n}{\sqrt{n \log\log n}} = -1 \qquad P\text{-fast sicher.}$$

Mehr dazu: siehe Vorlesung „Stochastische Prozesse."

1.2 Allgemeine Wahrscheinlichkeitsräume

Bisher haben wir uns noch nicht mit der Frage befasst, ob überhaupt ein Wahrscheinlichkeitsraum existiert, auf dem unendlich viele unabhängige Ereignisse bzw. Zufallsvariablen realisiert werden können. Auch die Realisierung einer auf einem endlichen reellen Intervall gleichverteilten Zufallsvariable auf einem geeigneten Wahrscheinlichkeitsraum haben wir noch nicht gezeigt. Die Existenz solcher Räume wurde stillschweigend vorausgesetzt.

Tatsächlich ist es oft nicht notwendig, den zugrunde liegenden Wahrscheinlichkeitsraum explizit zu kennen - die Kenntnis der gemeinsamen Verteilungen aller relevanten Zufallsvariablen genügt, um Wahrscheinlichkeiten und Erwartungswerte zu berechnen. Dennoch ist es an dieser Stelle hilfreich, die grundlegenden Existenzfragen zu klären, und unsere Modelle auf ein sicheres Fundament zu stellen. Die dabei entwickelten Begriffsbildungen werden sich beim Umgang mit stetigen und allgemeinen Zufallsvariablen als unverzichtbar erweisen.

1.2.1 Beispiele von Wahrscheinlichkeitsräumen

Wir beginnen mit einer Auflistung von verschiedenen Wahrscheinlichkeitsverteilungen. Während wir die ersten beiden Maße direkt auf der Potenzmenge $\mathcal{P}(\Omega)$ realisieren können, erfordert das

Aufstellen eines geeigneten Wahrscheinlichkeitsraums in den nachfolgenden Beispielen zusätzliche Überlegungen.

Dirac-Maße.

Sei Ω beliebig und $a \in \Omega$ ein festes Element. Das **Dirac-Maß** in a ist die durch

$$\delta_a[A] \;:=\; I_A(a) \;=\; \begin{cases} 1 & \text{falls } a \in A, \\ 0 & \text{sonst,} \end{cases}$$

definierte Wahrscheinlichkeitsverteilung $P = \delta_a$ auf der σ-Algebra $\mathcal{A} = \mathcal{P}(\Omega)$. Dirac-Maße sind „deterministische Verteilungen" – es gilt

$$\delta_a[\{a\}] \;=\; 1.$$

Konvexkombinationen von Dirac-Maßen.

Ist C eine abzählbare Teilmenge von Ω, und $p : C \to [0,1]$ eine Gewichtsfunktion mit $\sum_{\omega \in C} p(\omega) = 1$, dann ist durch

$$P[A] \;=\; \sum_{a \in A \cap C} p(a) \;=\; \sum_{a \in C} p(a)\, \delta_a[A] \quad \forall A \subseteq \Omega.$$

eine eindeutige Wahrscheinlichkeitsverteilung P auf der Potenzmenge $\mathcal{A} = \mathcal{P}(\Omega)$ gegeben. Die Verteilung P ist „rein atomar", d.h. die Masse sitzt auf abzählbaren vielen „Atomen" (den Elementen von C). Jede diskrete Wahrscheinlichkeitsverteilung ist von dieser Form.

Unendliches Produktmodell (z.B. Münzwurffolge)

Mehrstufige diskrete Modelle mit endlich vielen Stufen können wir auf der Potenzmenge des Produkts $\Omega = \{(\omega_1, \ldots, \omega_n) : \omega_i \in \Omega_i\} = \Omega_1 \times \ldots \times \Omega_n$ der Grundräume Ω_i realisieren. Es stellt sich die Frage, ob wir auch unendlich viele Zufallsvariablen auf einem ähnlichen Produktraum realisieren können. Im einfachsten Fall möchten wir eine Folge unabhängiger fairer Münzwürfe (0-1-Experimente) auf dem Grundraum

$$\Omega \;=\; \{\omega = (\omega_1, \omega_2, \ldots) : \omega_i \in \{0,1\}\} \;=\; \{0,1\}^{\mathbb{N}}$$

modellieren. Ω ist überabzählbar, denn die Abbildung $X : [0,1) \to \Omega$, die einer reellen Zahl die Ziffernfolge ihrer Binärdarstellung zuordnet, ist injektiv. Diese Abbildung ist explizit gegeben durch $X(u) = (X_1(u),\, X_2(u),\, X_3(u),\, \ldots)$, wobei

$$X_n(u) \;=\; I_{D_n}(u) \qquad \text{mit} \qquad D_n = \bigcup_{i=1}^{2^{n-1}} [(2i-1) \cdot 2^{-n}, 2i \cdot 2^{-n}) \tag{1.2.1}$$

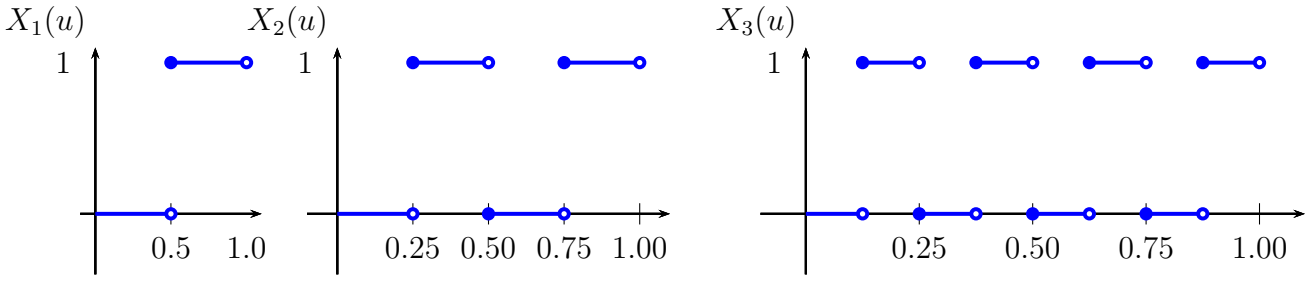


Abbildung 1.4: Binärdarstellung $X(u)$.

Wir suchen eine Wahrscheinlichkeitsverteilung P auf Ω, sodass

$$P[\{\omega \in \Omega : \omega_1 = a_1, \omega_2 = a_2, \ldots, \omega_n = a_n\}] = 2^{-n} \tag{1.2.2}$$

für alle $n \in \mathbb{N}$ und $a_1, \ldots, a_n \in \{0, 1\}$ gilt. Gibt es eine σ-Algebra $\mathcal{A}$, die alle diese Ereignisse enthält, und eine eindeutige Wahrscheinlichkeitsverteilung P auf $\mathcal{A}$ mit (1.2.2) ?

Wir werden in Abschnitt 2.3 zeigen, dass dies der Fall ist; wobei aber

(1). $\mathcal{A} \neq \mathcal{P}(\Omega)$ und

(2). $P[\{\omega\}] = 0$ für alle $\omega \in \Omega$

gelten muss. Das entsprechende Produktmodell unterscheidet sich in dieser Hinsicht grundlegend von diskreten Modellen.

Kontinuierliche Gleichverteilung

Für die Gleichverteilung auf einem endlichen reellen Intervall $\Omega = [a, b]$ oder $\Omega = [a, b]$, $-\infty < a < b < \infty$, sollte gelten:

$$P[(c, d)] = P[[c, d]] = \frac{d - c}{b - a} \quad \forall\, a \leq c < d \leq b. \tag{1.2.3}$$

Gibt es eine σ-Algebra $\mathscr{B}$, die alle Teilintervalle von $[a, b]$ enthält, und eine Wahrscheinlichkeitsverteilung P auf $\mathscr{B}$ mit (1.2.3)?

Wieder ist die Antwort positiv, aber erneut gilt notwendigerweise $\mathscr{B} \neq \mathcal{P}(\Omega)$ und $P[\{\omega\}] = 0$ für alle $\omega \in \Omega$.

Tatsächlich sind die Probleme in den letzten beiden Abschnitten weitgehend äquivalent: die durch die Binärdarstellung (1.2.1) definierte Abbildung X ist eine Bijektion von $[0, 1)$ nach $\{0, 1\}^{\mathbb{N}} \setminus A$, wobei $A = \{\omega \in \Omega : \omega_n = 1 \text{ schließlich}\}$ eine abzählbare Teilmenge ist. Eine Gleichverteilung auf $[0, 1)$ wird durch X auf eine Münzwurffolge auf $\{0, 1\}^{\mathbb{N}}$ abgebildet, und umgekehrt.

Brownsche Bewegung

Simuliert man einen Random Walk, so ergibt sich in einem geeigneten Skalierungslimes mit Schrittweite $\to 0$ anscheinend eine irreguläre, aber stetige zufällige Bewegung in kontinuierlicher Zeit. Der entsprechende, 1923 von N. Wiener konstruierte stochastische Prozess heißt **Brown-**

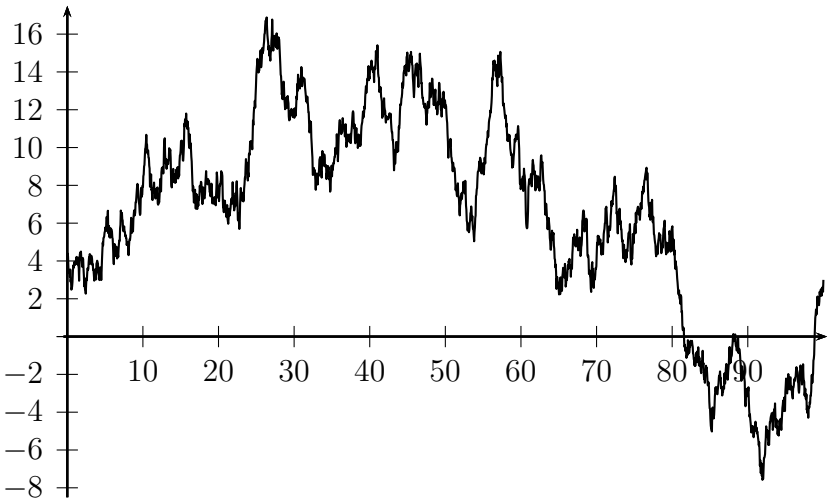


Abbildung 1.5: Graph einer Stichprobe der eindimensionalen Brownschen Bewegung

sche Bewegung, und kann durch eine Wahrscheinlichkeitsverteilung P (das Wienermaß) auf dem Raum

$$\Omega = C([0,1],\mathbb{R}) = \{\omega : [0,1] \to \mathbb{R} \,:\, \omega \text{ stetig}\}$$

beschrieben werden. Für diese, als Modell für Aktienkurse, zufällige Bewegungen, etc. in diversen Anwendungsbereichen fundamentale Wahrscheinlichkeitsverteilung gilt unter anderem

$$P[\{\omega \in \Omega : \omega(t) \in (a,b)\}] = \frac{1}{\sqrt{2\pi t}} \int_a^b e^{-\frac{x^2}{2t}}\, \mathrm{d}x \qquad \text{für alle } t > 0 \text{ und } a \le b, \tag{1.2.4}$$

siehe zum Beispiel die Vorlesung „Stochastische Prozesse“ im Sommersemester. Die Bedingung (1.2.4) legt die Wahrscheinlichkeitsverteilung P allerdings noch nicht eindeutig fest. Um P festzulegen, müssen z.B. die Wahrscheinlichkeiten

$$P[\{\omega \in \Omega : \omega(t_1) \in (a_1,b_1), \ldots, \omega(t_n) \in (a_n,b_n)\}] \tag{1.2.5}$$

für alle $n \in \mathbb{N}$, $t_i \ge 0$ und $a_i \le b_i$ angegeben werden. Im Fall der Brownschen Bewegung ergeben sich diese Wahrscheinlichkeiten aus (1.2.4) sowie der Unabhängigkeit und Stationarität der Inkremente $\omega(t) - \omega(s)$.

Um Wahrscheinlichkeitsverteilungen wie in den letzten drei Beispielen zu konstruieren, benötigen wir zunächst geeignete σ-Algebren, die die relevanten Ereignisse bzw. Intervalle enthalten. Dazu verwenden wir die folgende Konstruktion:

1.2.2 Konstruktion von σ-Algebren

Sei Ω eine beliebige Menge, und $\mathscr{J} \subseteq \mathcal{P}(\Omega)$ eine Kollektion von Ereignissen, die auf jeden Fall in der zu konstruierenden σ-Algebra enthalten sein sollen (z.B. die Mengen in (1.2.2)bei unendlichen Produktmodellen, die reellen Intervalle im Fall kontinuierlicher Gleichverteilungen, oder die Mengen in (1.2.5)auf dem Raum aller stetigen Funktionen.

Definition (Von einem Mengensystem erzeugte σ-Algebra). *Die Kollektion*

$$\sigma(\mathscr{J}) := \bigcap_{\substack{\mathscr{F} \supseteq \mathscr{J} \\ \mathscr{F} \text{ } \sigma\text{-Algebra auf } \Omega}} \mathscr{F}$$

*von Teilmengen von Ω heißt **die von $\mathscr{J}$-erzeugte σ-Algebra**.*

Bemerkung. Wie man leicht nachprüft (Übung), ist $\sigma(\mathscr{J})$ tatsächlich eine σ-Algebra, und damit die kleinste σ-Algebra, die $\mathscr{J}$ enthält.

Beispiel (Borel'sche σ-Algebra auf $\mathbb{R}$). Sei $\Omega = \mathbb{R}$ und $\mathscr{J} = \{(s,t) : -\infty \leq s \leq t \leq \infty\}$ die Kollektion aller offenen Intervalle. Die von $\mathscr{J}$ erzeugte σ-Algebra

$$\mathcal{B}(\mathbb{R}) := \sigma(\mathscr{J})$$

heißt Borel'sche σ-Algebra auf $\mathbb{R}$. Man prüft leicht nach, dass $\mathcal{B}(\mathbb{R})$ auch alle abgeschlossenen und halboffenen Intervalle enthält. Beispielsweise kann ein abgeschlossenes Intervall als Komplement der Vereinigung zweier offener Intervalle dargestellt werden. Die Borel'sche σ-Algebra wird auch erzeugt von der Kollektion aller abgeschlossenen bzw. aller kompakten Intervalle. Ebenso gilt

$$\mathcal{B}(\mathbb{R}) \;=\; \sigma(\{(-\infty, c] : c \in \mathbb{R}\}).$$

Allgemeiner definieren wir:

Definition (Borelsche σ-Algebra auf einem topologischen Raum). *Sei Ω ein topologischer Raum (also z.B. ein metrischer Raum wie $\mathbb{R}^n, C([0,1],\mathbb{R})$ etc.), und sei τ die Kollektion aller offenen Teilmengen von Ω (die **Topologie**). Die von τ erzeugte σ-Algebra*

$$\mathcal{B}(\Omega) := \sigma(\tau)$$

*heißt **Borel'sche σ-Algebra auf** Ω.*

Wieder verifiziert man, dass $\mathcal{B}(\Omega)$ auch von den abgeschlossenen Teilmengen erzeugt wird. Im Fall $\Omega = \mathbb{R}$ ergibt sich die oben definierte, von den Intervallen erzeugte σ-Algebra.

Bemerkung (Nicht-Borelsche Mengen). Nicht jede Teilmenge von $\mathbb{R}$ ist in der Borelschen σ-Algebra $\mathcal{B}(\mathbb{R})$ enthalten. Es ist allerdings gar nicht so einfach, nicht-Borelsche Mengen anzugeben – ein entsprechendes Beispiel wird in den Übungen betrachtet. Tatsächlich enthält $\mathcal{B}(\mathbb{R})$ so gut wie alle Teilmengen von $\mathbb{R}$, die in Anwendungsproblemen auftreten; z.B. alle offenen und abgeschlossenen Teilmengen von $\mathbb{R}$, sowie alle Mengen, die durch Bildung von abzählbar vielen Vereinigungen, Durchschnitten und Komplementbildungen daraus entstehen.

Beispiel (Produkt σ-Algebra auf $\{0,1\}^{\mathbb{N}}$). Auf dem Folgenraum

$$\Omega = \{0,1\}^{\mathbb{N}} = \{(\omega_1, \omega_2, \ldots) : \omega_i \in \{0,1\}\}$$

betrachten wir Teilmengen A von Ω von der Form

$$A = \{\omega \in \Omega : \omega_1 = a_1, \omega_2 = a_2, \ldots, \omega_n = a_n\}, \quad n \in \mathbb{N},\ a_1, \ldots, a_n \in \{0,1\}.$$

Im Beispiel unendlicher Produktmodelle von oben verwenden wir die von der Kollektion $\mathscr{C}$ aller dieser Mengen erzeugte σ-Algebra $\mathcal{A} = \sigma(\mathscr{C})$ auf $\{0,1\}^{\mathbb{N}}$. $\mathcal{A}$ heißt Produkt-σ-Algebra auf Ω.

Allgemeiner sei I eine beliebige Menge, und $\Omega = \prod_{i \in I} \Omega_i$ eine Produktmenge (mit endlich, abzählbar, oder sogar überabzählbar vielen Faktoren $\Omega_i, i \in I$).

Definition (Produkt-σ-Algebra). *Sind $\mathcal{A}_i, i \in I$ σ-Algebren auf Ω_i, dann ist die **Produkt-σ-Algebra***

$$\mathcal{A} = \bigotimes_{i \in I} \mathcal{A}_i$$

die von der Kollektion $\mathscr{C}$ aller Zylindermengen von der Form

$$\{\omega = (\omega_i)_{i \in I} \in \Omega : \omega_{i_1} \in A_{i_1}, \omega_{i_2} \in A_{i_2}, \ldots, \omega_{i_n} \in A_{i_n}\},$$

mit $n \in \mathbb{N}, i_1, \ldots, i_n \in I$, und $A_{i_1} \in \mathcal{A}_{i_1}, \ldots, A_{i_n} \in \mathcal{A}_{i_n}$, erzeugte σ-Algebra auf $\prod_{i \in I} \Omega_i$. .

Man kann nachprüfen, dass die etwas anders definierte Produkt-σ-Algebra aus dem Beispiel oben ein Spezialfall dieser allgemeinen Konstruktion ist.

1.2.3 Existenz und Eindeutigkeit von Wahrscheinlichkeitsverteilungen

Sei $(\Omega, \mathcal{A})$ nun ein **messbarer Raum**, d.h. Ω ist eine nichtleere Menge und $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ eine σ-Algebra. In der Regel sind auch die Wahrscheinlichkeiten $P[A]$ zunächst für Ereignisse A aus einer Teilmenge $\mathcal{J} \subseteq \mathcal{A}$ mit $\mathcal{A} = \sigma(\mathcal{J})$ gegeben, z.B. für Intervalle bei Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$. Es stellt sich die Frage, ob hierdurch bereits die Wahrscheinlichkeiten aller Ereignisse in $\mathcal{A}$ eindeutig festgelegt sind, und ob sich P zu einer Wahrscheinlichkeitsverteilung auf $\mathcal{A}$ fortsetzen lässt. Diese Fragen beantworten die folgenden beiden fundamentalen Sätze.

Definition (Durchschnittsstabiles Mengensystem; Mengenalgebra).

*(1). Ein Mengensystem $\mathcal{J} \subseteq \mathcal{A}$ heißt **durchschnittsstabil**, falls*

$$A, B \in \mathcal{J} \quad \Rightarrow \quad A \cap B \in \mathcal{J}.$$

*(2). $\mathcal{J}$ heißt **Algebra**, falls*

(a) $\Omega \in \mathcal{J}$

(b) $A \in \mathcal{J} \quad \Rightarrow \quad A^C \in \mathcal{J}$

(c) $A, B \in \mathcal{J} \quad \Rightarrow \quad A \cup B \in \mathcal{J}.$

Eine Algebra ist stabil unter endlichen Mengenoperationen (Bilden von endlichen Vereinigungen, Durchschnitten und Komplementen). Insbesondere ist jede Algebra durchschnittsstabil.

Beispiel. (1). Die Kollektion aller offenen Intervalle ist eine durchschnittsstabile Teilmenge von $\mathcal{B}(\mathbb{R})$, aber keine Algebra. Dasselbe gilt für das Mengensystem $\mathcal{J} = \{(-\infty, c] : c \in \mathbb{R}\}$.

(2). Die Kollektion aller endlichen Vereinigungen von beliebigen Teilintervallen von $\mathbb{R}$ ist eine Algebra.

Satz 1.6 (Eindeutigkeitssatz). *Stimmen zwei Wahrscheinlichkeitsverteilungen P und $\widetilde{P}$ auf $(\Omega, \mathcal{A})$ überein auf einem **durchschnittsstabilen Mengensystem** $\mathcal{J} \subseteq \mathcal{A}$, so auch auf $\sigma(\mathcal{J})$.*

Den Satz werden wir am Ende dieses Abschnittes beweisen.

Beispiel. (1). Eine Wahrscheinlichkeitsverteilung P auf $\mathcal{B}(\mathbb{R})$ ist eindeutig festgelegt durch die Wahrscheinlichkeiten $P[(-\infty, c]],\ c \in \mathbb{R}$.

(2). Die Wahrscheinlichkeitsverteilung P im Modell der unendlich vielen Münzwürfe ist eindeutig festgelegt durch die Wahrscheinlichkeiten der Ausgänge der ersten n Würfe für alle $n \in \mathbb{N}$.

Nach dem Eindeutigkeitssatz 1.6 ist eine Wahrscheinlichkeitsverteilung durch die Wahrscheinlichkeiten der Ereignisse aus einem durchschnittsstabilen Erzeugendensystem festgelegt. Umgekehrt zeigt der folgende Satz, dass sich eine auf einem Erzeugendensystem $\mathcal{J}$ gegebene σ-additive Abbildung zu einem Maß auf der σ-Algebra fortsetzen lässt, falls $\mathcal{J}$ eine Algebra ist.

Satz 1.7 (Fortsetzungssatz von Carathéodory). *Ist $\mathcal{J}$ eine Algebra, und $P : \mathcal{J} \to [0, \infty]$ eine σ-additive Abbildung, dann besitzt P eine Fortsetzung zu einem Maß auf $\sigma(\mathcal{J})$.*

Den Beweis dieses klassischen Resultats findet man in vielen Maßtheorie-, Analysis- bzw. Wahrscheinlichkeitstheorie-Büchern (siehe z. B. Williams: „Probability with martingales“, Appendix A1). Wir verweisen hier auf die Analysisvorlesung, da für die weitere Entwicklung der Wahrscheinlichkeitstheorie in dieser Vorlesung der Existenzsatz zwar fundamental ist, das Beweisverfahren aber keine Rolle mehr spielen wird.

Bemerkung (Eindeutigkeit der Fortsetzung). Ist $P[\Omega] = 1$, bzw. allgemeiner $P[\Omega] < \infty$, dann ist die Maßfortsetzung nach Satz 1.6 eindeutig, denn eine Algebra ist durchschnittsstabil.

Als Konsequenz aus dem Fortsetzungs- und Eindeutigkeitssatz erhält man:

Korollar 1.8 (Existenz und Eindeutigkeit der kontinuierlichen Gleichverteilung). *Es existiert genau eine Wahrscheinlichkeitsverteilung $\mathcal{U}_{(0,1)}$ auf $\mathcal{B}((0,1))$ mit*

$$\mathcal{U}_{(0,1)}[(a,b)] = b - a \qquad \textit{für alle } 0 < a \leq b < 1. \tag{1.2.6}$$

Zum Beweis ist noch zu zeigen, dass die durch (1.2.6) definierte Abbildung $\mathcal{U}_{(0,1)}$ sich zu einer σ-additiven Abbildung auf die von den offenen Intervallen erzeugte Algebra $\mathcal{A}_0$ aller endlichen Vereinigungen von beliebigen (offenen, abgeschlossenen, halboffenen) Teilintervallen von $(0,1)$ fortsetzen lässt. Wie die Fortsetzung von $\mathcal{U}_{(0,1)}$ auf $\mathcal{A}_0$ aussieht, ist offensichtlich - der Beweis der σ-Additivität ist dagegen etwas aufwändiger. Wir verweisen dazu wieder auf die Analysisvorlesung, bzw. den Appendix A1 in Williams: „Probability with martingales.“

Bemerkung (Lebesgue-Maß im $\mathbb{R}^d$). Auf ähnliche Weise folgt die Existenz und Eindeutigkeit des durch

$$\lambda[(a_1, b_1) \times \ldots \times (a_d, b_d)] = \prod_{i=1}^{d} (b_i - a_i) \qquad \text{für alle } a_i, b_i \in \mathbb{R} \text{ mit } a_i \leq b_i$$

eindeutig festgelegten Lebesguemaßes λ auf $\mathcal{B}(\mathbb{R}^d)$, siehe Analysis III. Man beachte, dass wegen $\lambda[\mathbb{R}^d] = \infty$ einige Aussagen, die wir für Wahrscheinlichkeitsverteilungen beweisen werden, nicht für das Lebesguemaß auf $\mathbb{R}^d$ gelten (aber normalerweise schon für Einschränkungen von λ auf beschränkte Teilmengen des $\mathbb{R}^d$).

Auch die Existenz der Wahrscheinlichkeitsverteilungen im Modell für unendlich viele faire Münzwürfe kann man mithilfe des Satzes von Carathéodory zeigen. Wir werden diese Wahrscheinlichkeitsverteilung stattdessen in Abschnitt 2.3 unmittelbar aus der Gleichverteilung $\mathcal{U}_{(0,1)}$ konstruieren.

1.2.4 Beweis des Eindeutigkeitssatzes

Zum Abschluss dieses Abschnitts beweisen wir nun den Eindeutigkeitssatz. Dazu betrachten wir das Mengensystem

$$\mathcal{D} := \{A \in \mathcal{A} \,:\, P[A] = \widetilde{P}[A]\} \supseteq \mathcal{J}.$$

Zu zeigen ist: $\mathcal{D} \supseteq \sigma(\mathcal{J})$.

Dazu stellen wir fest, dass $\mathcal{D}$ folgende Eigenschaften hat:

(i) $\Omega \in \mathcal{D}$

(ii) $A \in \mathcal{D} \Rightarrow A^C \in \mathcal{D}$

(iii) $A_1, A_2, \ldots \in \mathcal{D}$ paarweise disjunkt $\Rightarrow \bigcup A_i \in \mathcal{D}$

Definition (Dynkin-System). *Ein Mengensystem $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ mit den Eigenschaften (i) - (iii) heißt **Dynkin-System**.*

Bemerkung. Für ein Dynkin-System $\mathcal{D}$ gilt auch

$$A, B \in \mathcal{D}\,,\ A \subseteq B \ \Rightarrow\ B \backslash A = B \cap A^C = (B^C \cup A)^C \ \in \mathcal{D}\,,$$

da B^C und A disjunkt sind.

Lemma 1.9. *Jedes $\cap$ - stabile Dynkin-System $\mathcal{D}$ ist eine σ - Algebra.*

Beweis. Ist $\mathcal{D}$ ein $\cap$-stabiles Dynkin-System, dann gilt für $A, B \in \mathcal{D}$:

$$A \cup B \;=\; A \underset{\substack{\uparrow \\ disjunkt}}{\cup} \underbrace{(B \backslash \overbrace{(A \cap B)}^{\in \mathcal{D}\ da\ \cap-stabil})}_{\in \mathcal{D}\ nach\ Bem.} \quad \in \mathcal{D}.$$

Hieraus folgt für $A_1, A_2, \ldots \in \mathcal{D}$ durch Induktion

$$B_n := \bigcup_{i=1}^{n} A_i \;\in \mathcal{D},$$

und damit

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{n=1}^{\infty} B_n = \underset{\substack{\uparrow \\ disjunkt}}{\bigcup_{n=1}^{\infty}} \underbrace{(B_n \backslash B_{n-1})}_{\in \mathcal{D}\ nach\ Bem.} \in \mathcal{D}.$$

□

Lemma 1.10. *Ist $\mathcal{J}$ ein $\cap$ - stabiles Mengensystem , so stimmt* ***das von $\mathcal{J}$ erzeugte Dynkin-System***

$$\mathcal{D}(\mathcal{J}) \quad := \quad \bigcap_{\substack{\mathcal{D} \supseteq \mathcal{J} \\ \mathcal{D}\ Dynkin-System}} \mathcal{D}$$

mit der von $\mathcal{J}$ erzeugten σ - Algebra $\sigma(\mathcal{J})$ überein.

Aus Lemma 1.10 folgt die Aussage des Eindeutigkeitssatzes, denn $\{A \in \mathcal{A} : P[A] = \widetilde{P}[A]\}$ ist ein Dynkin-System, das $\mathcal{J}$ enthält, und somit gilt nach dem Lemma

$$\{A \in \mathcal{A} : P[A] = \widetilde{P}[A]\} \;\supseteq\; \mathcal{D}(\mathcal{J}) \;=\; \sigma(\mathcal{J}),$$

falls $\mathcal{J}$ durchschnittsstabil ist.

Beweis. (von Lemma 1.10)
Jede σ - Algebra ist ein Dynkin-System, also gilt $\mathcal{D}(\mathcal{J}) \subseteq \sigma(\mathcal{J})$.

Es bleibt zu zeigen, dass $\mathcal{D}(\mathcal{J})$ eine σ - Algebra ist (hieraus folgt dann $\mathcal{D}(\mathcal{J}) = \sigma(\mathcal{J})$). Nach dem ersten Lemma ist dies der Fall, wenn $\mathcal{D}(\mathcal{J})$ durchschnittsstabil ist. Dies zeigen wir nun in zwei Schritten:

Schritt 1: $B \in \mathcal{J}, A \in \mathcal{D}(\mathcal{J}) \;\Rightarrow\; A \cap B \in \mathcal{D}(\mathcal{J})$

Beweis: Für $B \in \mathcal{J}$ ist $\mathcal{D}_B := \{ A \in \mathcal{A} : A \cap B \in \mathcal{D}(\mathcal{J})\}$ ein Dynkin-System. Z.B. gilt

$$\begin{aligned} A \in \mathcal{D}_B &\Rightarrow A \cap B \in \mathcal{D}(\mathcal{J}) \\ &\Rightarrow A^C \cap B = \underset{\substack{\uparrow \\ \in \mathcal{D}(\mathcal{J})}}{B} \setminus (\underbrace{A \cap B}_{\in \mathcal{D}(\mathcal{J})}) \overset{Bem.}{\in} \mathcal{D}(\mathcal{J}) \\ &\Rightarrow A^C \in \mathcal{D}_B, \end{aligned}$$

usw. Da $\mathcal{J}$ durchschnittsstabil ist, ist $\mathcal{J}$ in $\mathcal{D}_B$ enthalten. Also folgt auch $\mathcal{D}(\mathcal{J}) \subseteq \mathcal{D}_B$, und damit $A \cap B \in \mathcal{D}(\mathcal{J})$ für alle $A \in \mathcal{D}(\mathcal{J})$.

Schritt 2: $A, B \in \mathcal{D}(\mathcal{J}) \Rightarrow A \cap B \in \mathcal{D}(\mathcal{J})$

Beweis: Für $A \in \mathcal{D}(\mathcal{J})$ ist $\mathcal{D}_A := \{ B \in \mathcal{A} : A \cap B \in \mathcal{D}(\mathcal{J})\}$ nach Schritt 1 ein Mengensystem, das $\mathcal{J}$ enthält. Zudem ist auch $\mathcal{D}_A$ ein Dynkin-System, wie man analog zu Schritt 1 zeigt. Also folgt $\mathcal{D}(\mathcal{J}) \subseteq \mathcal{D}_A$, d.h. $A \cap B \in \mathcal{D}(\mathcal{J})$ für alle $B \in \mathcal{D}(\mathcal{J})$. □

1.3 Zufallsvariablen und ihre Verteilung

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Wir wollen nun Zufallsvariablen $X : \Omega \to S$ mit Werten in einem allgemeinen messbaren Raum $(S, \mathcal{S})$ betrachten. Beispielsweise ist $S = \mathbb{R}$ oder $S = \mathbb{R}^d$ und $\mathcal{S}$ ist die Borelsche σ-Algebra. Oft interessieren uns die Wahrscheinlichkeiten von Ereignissen der Form

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

„Der Wert der Zufallsgröße X liegt in B“

wobei $B \subseteq S$ eine Menge aus der σ-Algebra $\mathcal{S}$ auf dem Bildraum ist, also z.B. ein Intervall oder eine allgemeinere Borelmenge, falls $S = \mathbb{R}$ gilt.

Definition (Von einer Abbildung erzeugte σ-Algebra). *Das Mengensystem*

$$\sigma(X) = \{X^{-1}(B) : B \in \mathcal{S}\} \subseteq \mathcal{P}(\Omega)$$

*heißt die **von der Abbildung** X **erzeugte** σ**-Algebra** auf Ω.*

Man verifiziert leicht, dass $\sigma(X)$ tatsächlich eine σ-Algebra ist.

Wir erweitern nun die zuvor eingeführten Konzepte einer Zufallsvariablen und ihrer Verteilung.

1.3.1 Zufallsvariablen mit allgemeinem Zustandsraum

Definition (Meßbare Abbildung; Zufallsvariable). *Eine Abbildung $X : \Omega \to S$ heißt **messbar bzgl.** $\mathcal{A}/\mathcal{S}$, falls*

$$(M) \qquad X^{-1}(B) \in \mathcal{A} \qquad \text{für alle } B \in \mathcal{S}.$$

*Eine **Zufallsvariable** ist eine auf einem Wahrscheinlichkeitsraum definierte messbare Abbildung.*

Bemerkung. (1). Ist $\mathcal{A} = \mathcal{P}(\Omega)$, dann ist jede Abbildung $X : \Omega \to S$ eine Zufallsvariable.

(2). Allgemein ist eine Abbildung $X : \Omega \to S$ genau dann meßbar bzgl. $\mathcal{A}/\mathcal{S}$, wenn $\mathcal{A}$ die von X erzeugte σ-Algebra enthält. Somit ist $\sigma(X)$ die *kleinste σ-Algebra auf Ω, bzgl. der X meßbar ist.*

(3). Ist S abzählbar und $\mathcal{S} = \mathcal{P}(S)$, dann ist X genau dann eine Zufallsvariable, falls

$$\{X = a\} = X^{-1}(\{a\}) \in \mathcal{A} \qquad \text{für alle } a \in S$$

gilt. Dies ist gerade die Definition einer diskreten Zufallsvariable von oben.

Stimmt die σ-Algebra $\mathcal{S}$ auf dem Bildraum nicht mit der Potenzmenge $\mathcal{P}(S)$ überein, dann ist es meist schwierig, eine Bedingung (M) für **alle** Mengen $B \in \mathcal{S}$ explizit zu zeigen. Die folgenden Aussagen liefern handhabbare Kriterien, mit denen man in fast allen praktisch relevanten Fällen sehr leicht zeigen kann, dass die zugrunde liegenden Abbildungen messbar sind. Wir bemerken zunächst, dass es genügt die Bedingung (M) für alle Mengen aus einem Erzeugendensystem $\mathcal{J}$ der σ-Algebra $\mathcal{S}$ zu überprüfen:

Lemma 1.11. *Sei $\mathcal{J} \subseteq \mathcal{P}(S)$ mit $\mathcal{S} = \sigma(\mathcal{J})$. Dann gilt (M) bereits, falls*

$$X^{-1}(B) \in \mathcal{A} \qquad \text{für alle } B \in \mathcal{J}.$$

Beweis. Das Mengensystem $\{B \in \mathcal{S} | X^{-1}(B) \in \mathcal{A}\}$ ist eine σ-Algebra, wie man leicht nachprüft. Diese enthält $\mathcal{J}$ nach Voraussetzung, also enthält sie auch die von $\mathcal{J}$ erzeugte σ-Algebra $\mathcal{S}$. □

Korollar (Reellwertige Zufallsvariablen). *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Eine Abbildung $X : \Omega \to \mathbb{R}$ ist genau dann eine Zufallsvariable bzgl. der Borelschen σ-Algebra, wenn*

$$\{X \le c\} = \{\omega \in \Omega \mid X(\omega) \le c\} \in \mathcal{A} \quad \forall\, c \in \mathbb{R}, \qquad \text{bzw. wenn}$$
$$\{X < c\} = \{\omega \in \Omega \mid X(\omega) < c\} \in \mathcal{A} \quad \forall\, c \in \mathbb{R}.$$

Beweis. Es gilt $\{X \leq c\} = X^{-1}((-\infty, c])$. Die Intervalle $(-\infty, c], c \in \mathbb{R}$, erzeugen $\mathcal{B}(\mathbb{R})$, also folgt die erste Aussage. Die zweite Aussage zeigt man analog. □

Beispiel (Indikatorfunktionen). Für eine Menge $A \subseteq \Omega$ gilt:

$$I_A \text{ ist Zufallsvariable } \Leftrightarrow \ A \in \mathcal{A},$$

denn

$$\{I_A \leq c\} = \begin{cases} \emptyset & \text{falls } c < 0 \\ \Omega & \text{falls } c \geq 1 \\ A^C & \text{falls } 0 \leq c < 1 \end{cases},$$

und A^C ist genau dann in $\mathcal{A}$ enthalten, wenn A in $\mathcal{A}$ enthalten ist.

Korollar (Stetige Abbildungen sind messbar). *Seien Ω und S topologische Räume, und $\mathcal{A}, \mathcal{S}$ die Borelschen σ-Algebren. Dann gilt:*

$$X : \Omega \to S \textit{ stetig } \Rightarrow \ X \textit{ messbar.}$$

Beweis. Sei $\mathcal{J}$ die Topologie von S, d.h. die Kollektion aller offenen Teilmengen von S. Nach Definition der Borelschen σ-Algebra gilt $\mathcal{S} = \sigma(\mathcal{J})$. Wegen

$$B \in \mathcal{J} \quad \Rightarrow B \text{ offen } \overset{X \text{ stetig}}{\Longrightarrow} \ X^{-1}(B) \text{ offen } \Longrightarrow \ X^{-1}(B) \in \mathcal{A}$$

folgt die Behauptung. □

Kompositionen von messbaren Abbildungen sind wieder messbar:

Lemma 1.12. *Sind $(\Omega_1, \mathcal{A}_1), (\Omega_2, \mathcal{A}_2)$ und $(\Omega_3, \mathcal{A}_3)$ messbare Räume, und ist $X_1 : \Omega_1 \to \Omega_2$ messbar bzgl. $\mathcal{A}_1/\mathcal{A}_2$ und $X_2 : \Omega_2 \to \Omega_3$ messbar bzgl. $\mathcal{A}_2/\mathcal{A}_3$, dann ist $X_2 \circ X_1$ messbar bzgl. $\mathcal{A}_1/\mathcal{A}_3$.*

$$\begin{array}{ccccc} \Omega_1 & \xrightarrow{X_1} & \Omega_2 & \xrightarrow{X_2} & \Omega_3 \\ \mathcal{A}_1 & & \mathcal{A}_2 & & \mathcal{A}_3 \end{array}$$

Beweis. Für $B \in \mathcal{A}_3$ gilt $(X_2 \circ X_1)^{-1}(B) = X_1^{-1}(\underbrace{X_2^{-1}(B)}_{\in \mathcal{A}_2}) \in \mathcal{A}_1$. □

Beispiel. (1). Ist $X : \Omega \to \mathbb{R}$ eine reellwertige Zufallsvariable und $f : \mathbb{R} \to \mathbb{R}$ eine messbare (z.B. stetige) Funktion, dann ist auch

$$f(X) := f \circ X : \Omega \to \mathbb{R}$$

wieder eine reellwertige Zufallsvariable. Beispielsweise sind $|X|, |X|^p, e^X$ usw. Zufallsvariablen.

(2). Sind $X, Y : \Omega \to \mathbb{R}$ reellwertige Zufallsvariablen, dann ist $(X, Y) : \omega \mapsto (X(\omega), Y(\omega))$ eine messbare Abbildung in den $\mathbb{R}^2$ mit Borelscher σ-Algebra.
Da die Abbildung $(x, y) \mapsto x + y$ stetig ist, ist $X + Y$ wieder eine reellwertige Zufallsvariable. Dies sieht man auch direkt wie folgt: Für $c \in \mathbb{R}$ gilt:

$$X + Y < c \iff \exists\, r, s \in \mathbb{Q} : r + s < c, X < r \text{ und } Y < s,$$

also

$$\{X + Y < c\} = \bigcup_{\substack{r,s \in \mathbb{Q} \\ r+s<c}} (\{X < r\} \cap \{Y < s\}) \in \mathcal{A}$$

1.3.2 Verteilungen von Zufallsvariablen

Um Zufallsexperimente zu analysieren, müssen wir wissen, mit welchen Wahrscheinlichkeiten die relevanten Zufallsvariablen Werte in bestimmten Bereichen annehmen. Dies wird durch die Verteilung beschrieben. Seien $(\Omega, \mathcal{A})$ und $(S, \mathcal{S})$ messbare Räume.

Satz 1.13 (Bild einer Wahrscheinlichkeitsverteilung unter einer ZV). *Ist P eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$, und $X : \Omega \to S$ messbar bzgl. $\mathcal{A}/\mathcal{S}$, dann ist durch*

$$\mu_X[B] := P[X \in B] = P[X^{-1}(B)] \qquad (B \in \mathcal{S})$$

eine Wahrscheinlichkeitsverteilung auf $(S, \mathcal{S})$ definiert.

Beweis. (1). $\mu_X(S) = P[X^{-1}(S)] = P[\Omega] = 1$

(2). Sind $B_n \in \mathcal{S}$, $n \in \mathbb{N}$, paarweise disjunkte Mengen, dann sind auch die Urbilder $X^{-1}(B_n)$, $n \in \mathbb{N}$, paarweise disjunkt. Also gilt wegen der σ-Additivität von P:

$$\mu_X\left[\bigcup_n B_n\right] = P\left[X^{-1}\left(\bigcup_n B_n\right)\right] = P\left[\bigcup_n X^{-1}(B_n)\right] = \sum_n P[X^{-1}(B_n)] = \sum_n \mu_X[B_n].$$

□

Definition (Verteilung einer Zufallsvariable). *Die Wahrscheinlichkeitsverteilung μ_X auf $(S, \mathcal{S})$ heißt **Verteilung (law) von X unter P**.*

Für μ_X werden häufig auch die folgenden Notationen verwendet:

$$\mu_X = P \circ X^{-1} = \mathcal{L}_X = P_X = X(P)$$

1.3.3 Reelle Zufallsvariablen; Verteilungsfunktion

Die Verteilung μ_X einer Zufallsvariablen X mit abzählbarem Wertebereich S ist eindeutig durch die Massenfunktion

$$p_X(a) = P[X = a] = \mu_X[\{a\}], \qquad a \in S,$$

festgelegt. Die Verteilung μ_X einer reellwertigen Zufallsvariablen $X : \Omega \to \mathbb{R}$ ist eine Wahrscheinlichkeitsverteilung auf $\mathcal{B}(\mathbb{R})$. Sie ist eindeutig festgelegt durch die Wahrscheinlichkeiten

$$\mu_X[(-\infty, c]] = P[X \leq c], \qquad c \in \mathbb{R},$$

da die Intervalle $(-\infty, c], c \in \mathbb{R}$, ein durchschnittsstabiles Erzeugendensystem der Borelschen σ-Algebra bilden.

Definition (Verteilungsfunktion). *Die Funktion $F_X : \mathbb{R} \to [0, 1]$,*

$$F_X(c) := P[X \leq c] = \mu_X[(-\infty, c]]$$

*heißt **Verteilungsfunktion (distribution function)** der Zufallsvariable $X : \Omega \to \mathbb{R}$ bzw. der Wahrscheinlichkeitsverteilung μ_X auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.*

Der folgende Satz nennt einige grundlegende Eigenschaften der Verteilungsfunktion einer auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definierten Zufallsvariable $X : \Omega \to \mathbb{R}$. Wir werden im nächsten Abschnitt sehen, dass umgekehrt jede Funktion mit den Eigenschaften (1)-(3) aus Satz 1.14 die Verteilungsfunktion einer reellen Zufallsvariable ist.

Satz 1.14 (Eigenschaften der Verteilungsfunktion).
Für die Verteilungsfunktion $F_X : \mathbb{R} \to [0, 1]$ einer reellwertigen Zufallsvariable X gilt:

(1). F_X *ist **monoton wachsend**,*

(2). $\lim_{c \to -\infty} F_X(c) = 0$ *und* $\lim_{c \to \infty} F_X(c) = 1$,

(3). F_X *ist **rechtsstetig**, d.h.* $F_X(c) = \lim_{y \searrow c} F_X(y)$ *für alle* $c \in \mathbb{R}$,

(4). $F_X(c) = \lim_{y \nearrow c} F_X(y) + \mu_X[\{c\}]$.
Insbesondere ist F_X genau dann stetig bei c, wenn $\mu_X[\{c\}] = 0$ gilt.

Beweis. Die Aussagen folgen unmittelbar aus der monotonen Stetigkeit und Normiertheit der zugrundeliegenden Wahrscheinlichkeitsverteilung P. Der Beweis der Eigenschaften (1)-(3) wird dem Leser als Übung überlassen. Zum Beweis von (4) bemerken wir, dass für $y < c$ gilt:

$$F_X(c) - F_X(y) = P[X \leq c] - P[X \leq y] = P[y < X \leq c].$$

Für eine monoton wachsende Folge $y_n \nearrow c$ erhalten wir daher aufgrund der monotonen Stetigkeit von P:

$$\begin{aligned} F_X(c) - \lim_{n\to\infty} F_X(y_n) &= \lim_{n\to\infty} P[y_n < X \le c] = P\left[\bigcap_n \{y_n < X \le c\}\right] \\ &= P[X = c] = \mu_X[\{c\}]. \end{aligned}$$

Da dies für alle Folgen $y_n \nearrow c$ gilt, folgt die Behauptung. □

1.3.4 Diskrete und stetige Verteilungen

Nach Satz 1.14 (4) sind die Unstetigkeitsstellen der Verteilungsfunktion F_X einer reellwertigen Zufallsvariable X gerade die *Atome* der Verteilung, d.h. die $c \in \mathbb{R}$ mit $\mu_X[\{c\}] > 0$. Nimmt X nur endlich viele Werte in einem Intervall I an, dann ist F auf I stückweise konstant, und springt nur bei diesen Werten. Allgemeiner nennen wir Verteilung von X **diskret**, wenn $\mu_X[S] = 1$ für eine abzählbare Menge S gilt. In diesem Fall ist die Verteilungsfunktion gegeben durch

$$F_X(c) \;=\; \mu_X[(-\infty, c]] \;=\; \sum_{\substack{a \in S \\ a \le c}} \mu_X[\{a\}].$$

Hingegen nennen wir die Verteilung von X **stetig**, bzw. **absolutstetig**, falls eine integrierbare Funktion $f_X : \mathbb{R} \to [0, \infty)$ existiert mit

$$F_X(c) \;=\; P[X \le c] \;=\; \mu_X[(-\infty, c]] \;=\; \int_{-\infty}^{c} f_X(x)\, \mathrm{d}x \qquad \text{für alle } c \in \mathbb{R}. \tag{1.3.1}$$

Das Integral ist dabei im Allgemeinen als Lebesgueintegral zu interpretieren. Ist die Funktion f_X stetig, dann stimmt dieses mit dem Riemannintegral überein. Da μ_X eine Wahrscheinlichkeitsverteilung ist, folgt, dass f_X eine **Wahrscheinlichkeitsdichte** ist, d.h. $f_X \ge 0$ und

$$\int_{\mathbb{R}} f_X(x)\, dx \;=\; 1.$$

Definition (Dichtefunktion). *Eine Lebesgue-integrierbare Funktion $f_X : \mathbb{R} \to [0, \infty)$ mit (1.3.1) heißt **Dichtefunktion** der Zufallsvariable X bzw. der Verteilung μ_X.*

Bemerkung. (1). Nach dem Hauptsatz der Differential- und Integralrechnung gilt

$$F_X'(x) \;=\; f_X(x) \tag{1.3.2}$$

für alle $x \in \mathbb{R}$, falls f stetig ist. Im Allgemeinen gilt (1.3.2) für λ-fast alle x, wobei λ das Lebesguemaß auf $\mathbb{R}$ ist.

(2). Aus (1.3.1) folgt aufgrund der Eigenschaften des Lebesgueintegrals (s. Kapitel 3 unten):

$$P[X \in B] = \mu_X[B] = \int_B f_X(x)\, dx, \tag{1.3.3}$$

für alle Mengen $B \in \mathcal{B}(\mathbb{R})$. Zum Beweis zeigt man, dass beide Seiten von (1.3.3) Wahrscheinlichkeitsverteilungen definieren, und wendet den Eindeutigkeitssatz an.

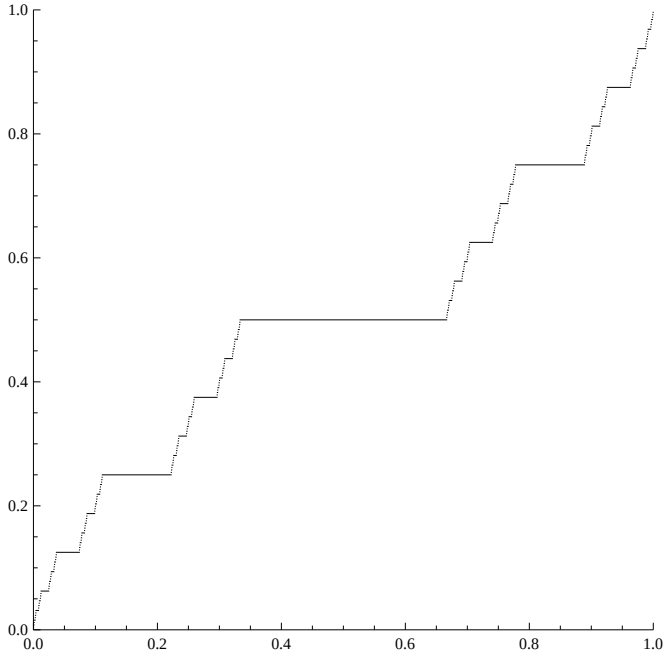
Beispiele (**Diskrete und absolutstetige Verteilungen**). (1). Ist X deterministisch mit $P[X = a] = 1$ für ein $a \in \mathbb{R}$, dann ist die Verteilung von X das Dirac-Maß δ_a. Das Dirac-Maß ist diskret mit Verteilungsfunktion $F_X(c) = I_{[a,\infty)}(c)$.

(2). Die Gleichverteilung $\mathcal{U}_{(0,1)}$ ist eine stetige Verteilung mit Verteilungsfunktion $F(c) = 0$ für $c \leq 0$, $F(c) = c$ für $c \in [0, 1]$, und $F(c) = 1$ für $c \geq 1$.

(3). Die Wahrscheinlichkeitsverteilung $\frac{1}{2}\delta_a + \frac{1}{2}\mathcal{U}_{(0,1)}$ ist weder stetig noch diskret.

Die Verteilung aus dem letzten Beispiel ist zwar weder absolutstetig noch diskret, aber sie kann sofort zerlegt werden in einen absolutstetigen und einen diskreten Anteil. Für die im folgenden Beispiel betrachtete Gleichverteilung auf der Cantor-Menge existiert keine solche Zerlegung:

Beispiel (**Devil's staircase**). Wir betrachten die wie folgt definierte Verteilungsfunktion F einer Wahrscheinlichkeitsverteilung auf dem Intervall $(0, 1)$: $F(c) = 0$ für $c \leq 0$, $F(c) = 1$ für $c \geq 1$, $F(c) = 1/2$ für $c \in [1/3, 2/3]$, $F(c) = 1/4$ für $c \in [1/9, 2/9]$, $F(c) = 3/4$ für $c \in [7/9, 8/9]$, $F(c) = 1/8$ für $c \in [1/27, 2/27]$, usw. Man überzeugt sich leicht, dass auf diese Weise eine eindeutige **stetige** monotone wachsende Funktion $F : \mathbb{R} \to [0, 1]$ definiert ist. Nach Satz 1.14 unten ist F also die Verteilungsfunktion einer Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}$.



Da F stetig ist, hat das Maß μ **keinen diskreten Anteil**, d.h. $\mu[\{a\}] = 0$ für alle $a \in \mathbb{R}$. Da F auf den Intervallen $[1/3, 2/3]$, $[1/9, 2/9]$, $[7/9, 8/9]$ usw. jeweils konstant ist, sind alle diese Intervalle μ-Nullmengen. Die Verteilung μ sitzt also auf dem Komplement

$$C = \left\{ \sum_{i=1}^{\infty} a_i 3^{-i} : a_i \in \{0, 2\} (i \in \mathbb{N}) \right\}$$

der Vereinigung der Intervalle. Die **Cantor-Menge** C ist ein Fraktal, das aus den reellen Zahlen zwischen 0 und 1 besteht, die sich im Dreier-System ohne Verwendung der Ziffer 1 darstellen lassen. Sie ist eine Lebesgue-Nullmenge, aber der Träger des Maßes μ. Da F der Grenzwert der Verteilungsfunktionen von Gleichverteilungen auf den Mengen $C_1 = (0, 1)$, $C_2 = (0, 1/3) \cup (2/3, 1)$, $C_3 = (0, 1/9) \cup (2/9, 1/3) \cup (2/3, 7/9) \cup (8/9, 1)$,... ist, können wir μ als **Gleichverteilung auf der Cantor-Menge** $C = \bigcap C_n$ interpretieren.

Weiterhin ist die Verteilungsfunktion F auf den oben betrachteten Intervallen konstant, und daher Lebesgue-fast überall differenzierbar mit Ableitung $F'(x) = 0$. Die „Teufelsleiter" F wächst also von 0 auf 1, obwohl sie stetig ist und ihre Ableitung fast überall gleich Null ist ! Hieraus folgt, dass die Verteilung μ **nicht absolutstetig** sein kann, denn in diesem Fall wäre F gleich dem Integral der Lebesgue-fast überall definierten Funktion F', also konstant.

1.4 Spezielle Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$

Im Folgenden betrachten wir einige wichtige Beispiele von eindimensionalen Verteilungen und ihren Verteilungsfunktionen. Wir geben zunächst die Verteilungsfunktion für einige elementare diskrete Verteilungen an, und betrachten dann kontinuierliche Analoga dieser diskreten Verteilungen.

1.4.1 Diskrete Verteilungen

Beispiele. (1). GLEICHVERTEILUNG AUF EINER ENDLICHEN MENGE $\{a_1, \ldots, a_n\} \subset \mathbb{R}$:

Ist $S = \{a_1, \ldots, a_n\} \subset \mathbb{R}$ eine n-elementige Teilmenge von $\mathbb{R}$, dann ist die Gleichverteilung μ auf S durch

$$\mu \;=\; \frac{1}{n} \sum_{i=1}^{n} \delta_{a_i} \tag{1.4.1}$$

gegeben. Der Wert der Verteilungsfunktion F von μ springt an jeder der Stellen $a_1, \ldots, a_n$ um $1/n$ nach oben. Die **empirische Verteilung** von $a_1, \ldots, a_n$ ist ebenfalls durch (1.4.1) definiert, wobei hier aber nicht vorausgesetzt wird, dass $a_1, \ldots, a_n$ verschieden sind. Die Sprunghöhen der empirischen Verteilungsfunktion sind dementsprechend Vielfache von $1/n$.

(2). GEOMETRISCHE VERTEILUNG MIT PARAMETER $p \in [0, 1]$:

$$\mu[\{k\}] = (1-p)^{k-1} \cdot p \quad \text{für } k \in \mathbb{N}.$$

Für eine geometrisch verteilte Zufallsvariable T gilt:

$$F(c) = P[T \leq c] = 1 - \underbrace{P[T > c]}_{=P[T > \lfloor c \rfloor]} = 1 - (1-p)^{\lfloor c \rfloor} \quad \text{für } c \geq 0,$$

wobei $\lfloor c \rfloor := \max\{n \in \mathbb{Z} : n \leq c\}$ der ganzzahlige Anteil von c ist.

(3). BINOMIALVERTEILUNG MIT PARAMETERN n UND p:

$$\mu[\{k\}] = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{für} \quad k = 0, 1, \ldots, n.$$

Somit ist die Verteilungsfunktion von $\mathrm{Bin}(n, p)$:

$$F(c) = \sum_{k=0}^{\lfloor c \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$$

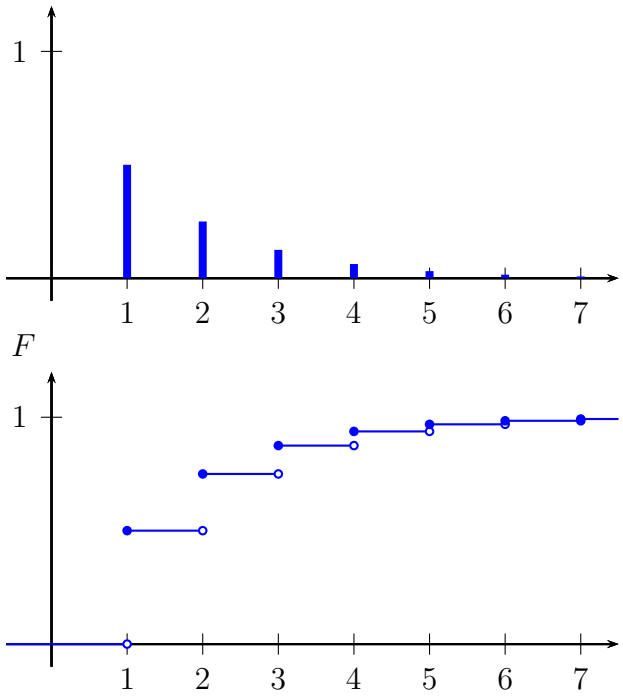


Abbildung 1.6: Massen- und Verteilungsfunktion einer $Geom(\frac{1}{2})$-verteilten Zufallsvariablen.

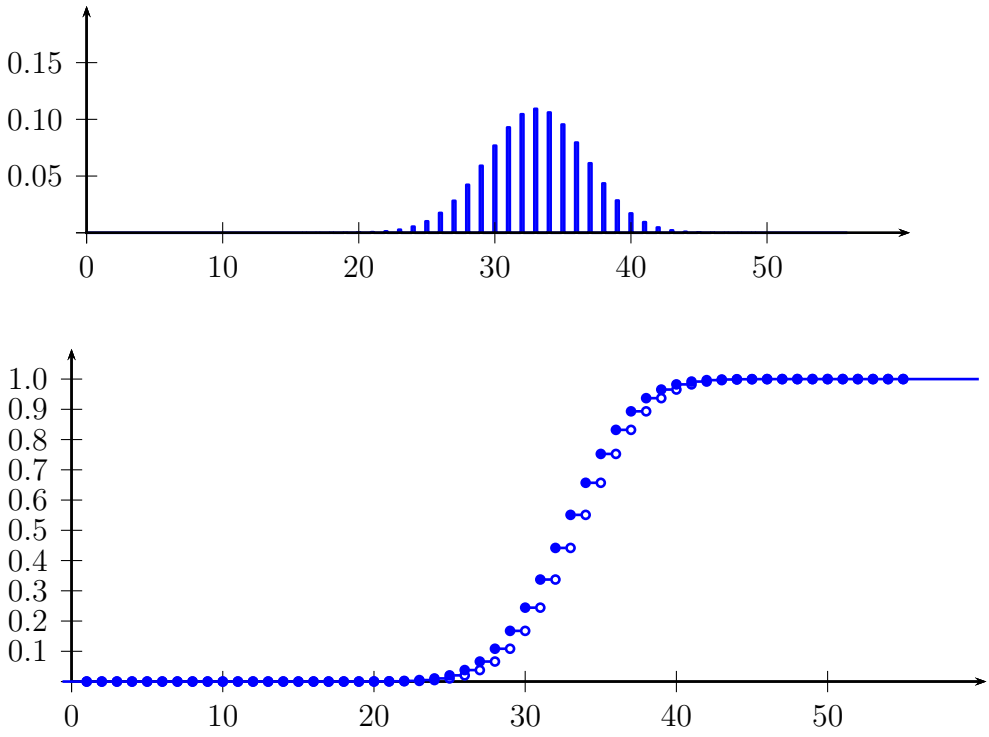


Abbildung 1.7: Massen- und Verteilungsfunktion von $\mathrm{Bin}(55, 0.6)$

1.4.2 Kontinuierliche Gleichverteilung

Seien $a, b \in \mathbb{R}$ mit $a < b$. Eine Zufallsvariable $X : \Omega \to \mathbb{R}$ ist gleichverteilt auf dem Intervall (a, b), falls

$$P[X \leq c] \;=\; \mathcal{U}_{(a,b)}[(a, c)] \;=\; \frac{c-a}{b-a} \qquad \text{für alle } c \in (a, b)$$

gilt. Eine auf $(0, 1)$ gleichverteilte Zufallsvariable ist zum Beispiel die Identität

$$U(\omega) = \omega$$

auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P) = ((0, 1), \mathcal{B}((0, 1)), \mathcal{U}_{(0,1)})$. Ist U gleichverteilt auf $(0, 1)$, dann ist die Zufallsvariable

$$X(\omega) = a + (b - a)U(\omega)$$

gleichverteilt auf (a, b).

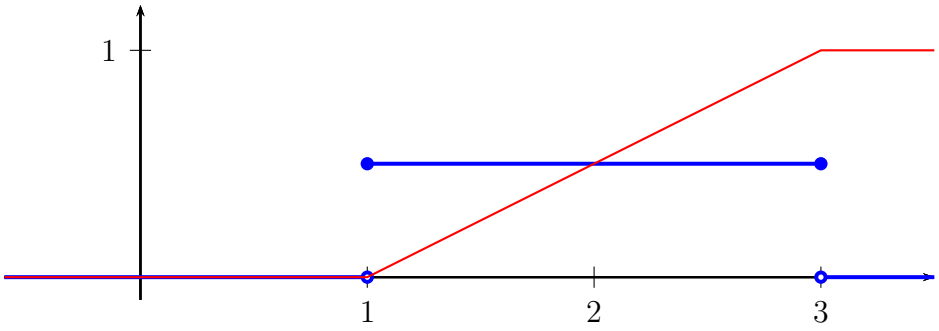


Abbildung 1.8: Dichte $f(x) = I_{[1,3]}(x)/2$ einer auf $[1, 3]$ gleichverteilten Zufallsvariable (blau), und deren Verteilungsfunktion $F(c)$ (rot).

Die Dichte und Verteilungsfunktion der Verteilung $U_{(a,b)}$ sind gegeben durch

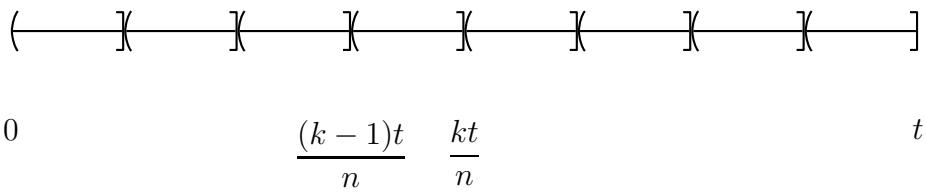
$$f(x) \;=\; \frac{1}{b-a}\, I_{(a,b)}(x), \qquad F(c) \;=\; \begin{cases} 0 & \text{für } c \leq a, \\ \frac{c-a}{b-a} & \text{für } a \leq c \leq b, \\ 1 & \text{für } c \geq b. \end{cases}$$

Affine Funktionen von gleichverteilten Zufallsvariablen sind wieder gleichverteilt.

1.4.3 Exponentialverteilung

Das kontinuierliche Analogon zur geometrischen Verteilung ist die Exponentialverteilung. Angenommen, wir wollen die Wartezeit auf das erste Eintreten eines unvorhersehbaren Ereignisses

(z.B. radioaktiver Zerfall) mithilfe einer Zufallsvariable $T : \Omega \to (0, \infty)$ beschreiben. Wir überlegen uns zunächst, welche Verteilung zur Modellierung einer solchen Situation angemessen sein könnte. Um die Wahrscheinlichkeit $P[T > t]$ zu approximieren, unterteilen wir das Intervall $(0, t]$ in eine große Anzahl $n \in \mathbb{N}$ von gleich großen Intervallen $(\frac{(k-1)t}{n}, \frac{kt}{n}], 1 \leq k \leq n$.



Sei A_k das Ereignis, dass das unvorhersehbare Geschehen im Zeitraum $(\frac{(k-1)t}{n}, \frac{kt}{n}]$ eintritt. Ein nahe liegender Modellierungsansatz ist anzunehmen, dass die Ereignisse A_k unabhängig sind mit Wahrscheinlichkeit

$$P[A_k] \approx \lambda \frac{t}{n},$$

wobei $\lambda > 0$ die „Intensität“, d.h. die mittlere Häufigkeit des Geschehens pro Zeiteinheit, beschreibt, und die Approximation für $n \to \infty$ immer genauer wird. Damit erhalten wir:

$$P[T > t] = P[A_1^C \cap \ldots \cap A_n^C] \approx \left(1 - \frac{\lambda t}{n}\right)^n \quad \text{für großes } n.$$

Für $n \to \infty$ konvergiert die rechte Seite gegen $e^{-\lambda t}$. Daher liegt folgende Definition nahe:

Definition (Exponentialverteilung). *Eine Zufallsvariable $T : \Omega \to [0, \infty)$ heißt **exponentialverteilt zum Parameter** $\lambda > 0$, falls*

$$P[T > t] = e^{-\lambda t} \qquad \textit{für alle } t \geq 0 \textit{ gilt.}$$

*Die **Exponentialverteilung zum Parameter** λ ist dementsprechend die Wahrscheinlichkeitsverteilung $\mu = \mathrm{Exp}(\lambda)$ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit*

$$\mu[(t, \infty)] = e^{-\lambda t} \qquad \textit{für alle } t \geq 0,$$

bzw. mit Verteilungsfunktion

$$F(t) = \mu[(-\infty, t]] = \begin{cases} 1 - e^{-\lambda t} & \textit{für } t \geq 0, \\ 0 & \textit{für } t < 0. \end{cases} \tag{1.4.2}$$

Nach dem Eindeutigkeitssatz ist die $\mathrm{Exp}(\lambda)$-Verteilung durch (1.4.2) eindeutig festgelegt.

Die Dichte der Exponentialverteilung mit Parameter $\lambda > 0$ ist gegeben durch

$$f(t) \;=\; \lambda e^{-\lambda t} I_{(0,\infty)}(t).$$

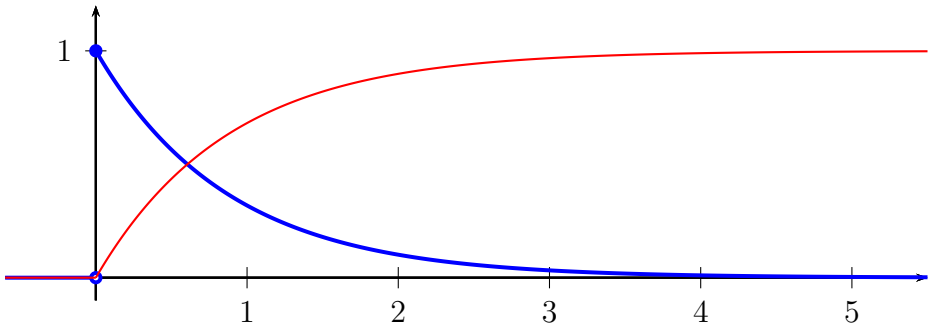


Abbildung 1.9: Dichte $f(t) = I_{[0,\infty)}(t) \cdot e^{-t}$ einer zum Parameter 1 exponentialverteilten Zufallsvariable (blau) und deren Verteilungsfunktion $F(c)$ (rot)

Ist T eine exponentialverteilte Zufallsvariable zum Parameter λ, und $a > 0$, dann ist aT exponentialverteilt zum Parameter $\frac{\lambda}{a}$, denn

$$P[aT > c] \;=\; P[T > c/a] \;=\; \exp\left(-\frac{\lambda}{a}c\right) \qquad \text{für alle } c \geq 0.$$

Eine bemerkenswerte Eigenschaft exponentialverteilter Zufallsvariablen ist die „Gedächtnislosigkeit":

Satz 1.15 (Gedächtnislosigkeit der Exponentialverteilung). *Ist T exponentialverteilt, dann gilt für alle $s, t \geq 0$:*

$$P[T - s > t | T > s] \;=\; P[T > t].$$

Hierbei ist $T - s$ die verbleibende Wartezeit auf das erste Eintreten des Ereignisses. Also: *Auch wenn man schon sehr lange vergeblich gewartet hat, liegt das nächste Ereignis nicht näher als am Anfang!*

Beweis.

$$P[T-s > t | T > s] = \frac{P[T - s > t \text{ und } T > s]}{P[T > s]} = \frac{P[T > s + t]}{T > s} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P[T > t].$$

□

Wir konstruieren nun explizit eine exponentialverteilte Zufallsvariable aus einer gleichverteilten Zufallsvariable. Dazu bemerken wir, dass $T : \Omega \to \mathbb{R}$ genau dann exponentialverteilt mit Parameter λ ist, wenn

$$P[e^{-\lambda T} < u] \;=\; P\left[T > -\frac{1}{\lambda}\log u\right] \;=\; e^{\frac{\lambda}{\lambda}\log u} \;=\; u$$

für alle $u \in (0,1)$ gilt, d.h. wenn $e^{-\lambda T}$ auf $(0,1)$ gleichverteilt ist. Also können wir eine exponentialverteilte Zufallsvariable erhalten, indem wir umgekehrt

$$T \;:=\; -\frac{1}{\lambda}\log U \qquad \text{mit} \qquad U \sim \mathcal{U}_{(0,1)}$$

setzen. Insbesondere ergibt sich die folgende Methode zur Simulation einer exponentialverteilten Zufallsvariable:

Algorithmus 1.16 (Simulation einer exponentialverteilten Stichprobe).

Input: Intensität $\lambda > 0$

Output: Stichprobe t von $\mathrm{Exp}(\lambda)$

(1). Erzeuge $u \sim \mathcal{U}_{(0,1)}$.

(2). Setze $t := -\frac{1}{\lambda}\log u$.

Wir werden in Abschnitt 1.6 zeigen, dass mit einem entsprechenden Verfahren beliebige reelle Zufallsvariablen konstruiert und simuliert werden können.

1.4.4 Normalverteilungen

Wegen $\int_{-\infty}^{\infty} e^{-z^2/2} dz = \sqrt{2\pi}$ ist die „Gaußsche Glockenkurve“

$$f(z) \;=\; \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}, \qquad z \in \mathbb{R},$$

eine Wahrscheinlichkeitsdichte. Eine stetige Zufallsvariable Z mit Dichtefunktion f heißt **standardnormalverteilt**. Die Verteilungsfunktion

$$\Phi(c) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz$$

der Standardnormalverteilung ist im Allgemeinen nicht explizit berechenbar. Ist Z standardnormalverteilt, und

$$X(\omega) = \sigma Z(\omega) + m$$

mit $\sigma > 0, m \in \mathbb{R}$, dann ist X eine Zufallsvariable mit Verteilungsfunktion

$$F_X(c) \;=\; P[X \le c] \;=\; P\left[Z \le \frac{c-m}{\sigma}\right] \;=\; \Phi\left(\frac{c-m}{\sigma}\right).$$

Mithilfe der Substitution $z = \frac{x-m}{\sigma}$ erhalten wir

$$F_X(c) \;=\; \int_{-\infty}^{\frac{c-m}{\sigma}} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz \;=\; \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}\, dx.$$

Definition (Normalverteilung). *Die Wahrscheinlichkeitsverteilung $N(m, \sigma^2)$ auf $\mathbb{R}$ mit Dichtefunktion*

$$f_{m,\sigma}(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

*heißt **Normalverteilung mit Mittel** m **und Varianz** σ^2**.** Die Verteilung $N(0,1)$ heißt **Standardnormalverteilung**.*

Wir werden im nächsten Abschnitt sehen, dass die Binomialverteilung (also die Verteilung der Anzahl der Erfolge bei unabhängigen 0-1-Experimenten mit Erfolgswahrscheinlichkeit p) für große n näherungsweise durch eine Normalverteilung beschrieben werden kann. Entsprechendes gilt viel allgemeiner für die Verteilungen von Summen vieler kleiner unabhängiger Zufallsvariablen (*Zentraler Grenzwertsatz*, s.u.).

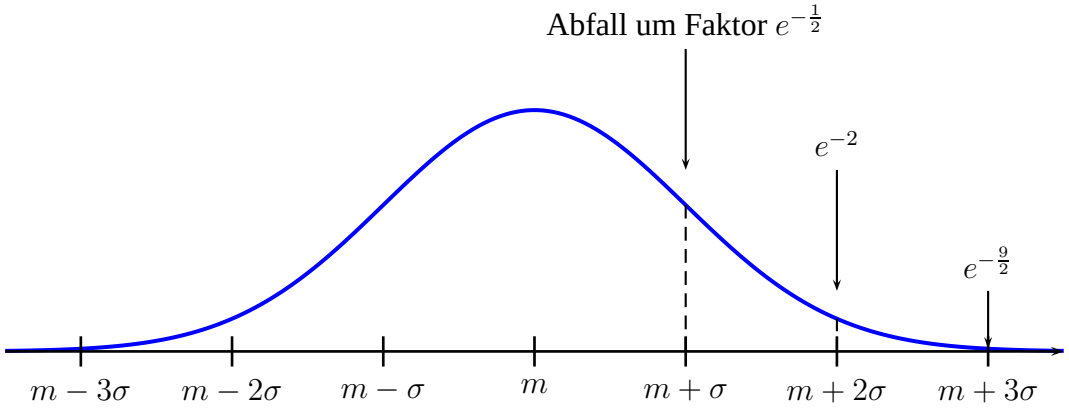


Abbildung 1.10: Dichte der Normalverteilung mit Mittelwert m und Varianz σ^2.

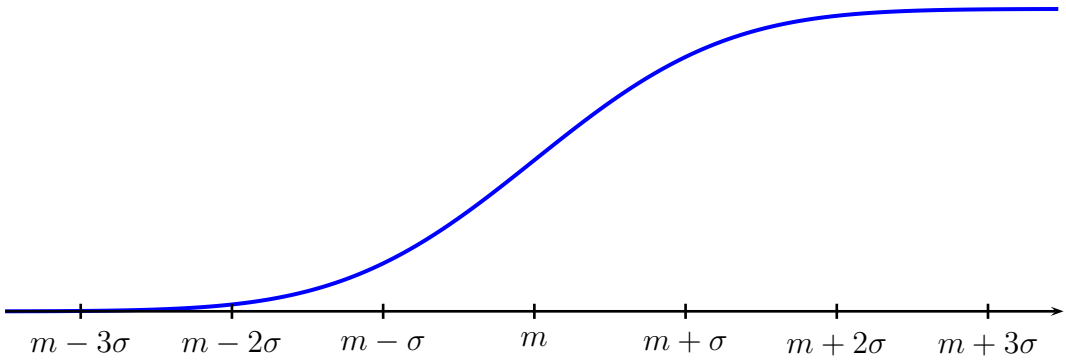


Abbildung 1.11: Verteilungsfunktion der Normalverteilung mit Mittelwert m und Varianz σ^2.

Die Dichte der Normalverteilung ist an der Stelle m maximal, und klingt außerhalb einer σ-Umgebung von m rasch ab. Beispielsweise gilt

$$f_{m,\sigma}(m \pm \sigma) = \frac{f_{m,\sigma}(m)}{\sqrt{e}}, \quad f_{m,\sigma}(m \pm 2\sigma) = \frac{f_{m,\sigma}(m)}{e^2}, \quad f_{m,\sigma}(m \pm 3\sigma) = \frac{f_{m,\sigma}(m)}{e^{9/2}}.$$

Für die Wahrscheinlichkeit, dass eine normalverteilte Zufallsvariable Werte außerhalb der σ-, 2σ- und 3σ-Umgebungen annimmt, erhält man:

$$\begin{aligned} P[|X-m| > k\sigma] &= P\left[\left|\frac{X-m}{\sigma}\right| > k\right] \\ &= P[|Z| > k] \;=\; 2P[Z > k] \;=\; 2(1-\Phi(k)) \\ &= \begin{cases} 31.7\% & \text{für } k=1, \\ 4.6\% & \text{für } k=2, \\ 0.26\% & \text{für } k=3. \end{cases} \end{aligned}$$

Eine Abweichung der Größe σ vom Mittelwert m ist also für eine normalverteilte Zufallsvariable relativ typisch, eine Abweichung der Größe 3σ dagegen schon sehr selten.

Die folgenden expliziten Abschätzungen für die Wahrscheinlichkeiten großer Werte sind oft nützlich:

Lemma 1.17. *Für eine standardnormalverteilte Zufallsvariable Z gilt:*

$$(2\pi)^{-1/2} \cdot \left(\frac{1}{y} - \frac{1}{y^3}\right) \cdot e^{-y^2/2} \quad \leq \quad P[Z \geq y] \quad \leq \quad (2\pi)^{-1/2} \cdot \frac{1}{y} \cdot e^{-y^2/2} \qquad \forall y > 0.$$

Beweis. Es gilt:

$$P[Z \geq y] = (2\pi)^{-1/2} \int_y^\infty e^{-z^2/2}\, dz$$

Um das Integral abzuschätzen, versuchen wir approximative Stammfunktionen zu finden. Zunächst gilt:

$$\frac{d}{dz}\left(-\frac{1}{z}e^{-z^2/2}\right) \;=\; \left(1+\frac{1}{z^2}\right)\cdot e^{-z^2/2} \;\geq\; e^{-z^2/2} \qquad \forall z \geq 0,$$

also

$$\frac{1}{y}e^{-y^2/2} \;=\; \int\limits_y^\infty \frac{d}{dz}\left(-\frac{1}{z}e^{-z^2/2}\right) \;\geq\; \int\limits_y^\infty e^{-z^2/2}\,dz,$$

woraus die obere Schranke für $P[Z \geq y]$ folgt.

Für die untere Schranke approximieren wir die Stammfunktion noch etwas genauer. Es gilt:

$$\frac{d}{dz}\left(\left(-\frac{1}{z}+\frac{1}{z^3}\right)e^{-z^2/2}\right) \;=\; \left(1+\frac{1}{z^2}-\frac{1}{z^2}-\frac{3}{z^4}\right)e^{-z^2/2} \;\leq\; e^{-z^2/2},$$

und damit

$$\left(\frac{1}{y}-\frac{1}{y^3}\right)e^{-y^2/2} \;\leq\; \int\limits_y^\infty e^{-z^2/2}\,dz.$$

□

Für eine $N(m,\sigma^2)$-verteilte Zufallsvariable X mit $\sigma > 0$ ist $Z = \frac{X-m}{\sigma}$ standardnormalverteilt. Also erhalten wir für $y \geq m$:

$$P[X \geq y] \;=\; P\left[\frac{X-m}{\sigma} \;\geq\; \frac{y-m}{\sigma}\right] \;\leq\; \frac{\sigma}{y-m}\cdot(2\pi)^{-1/2}\cdot e^{-\frac{(y-m)^2}{2\sigma^2}}, \tag{1.4.3}$$

sowie eine entsprechende Abschätzung nach unten.

1.5 Normalapproximation der Binomialverteilung

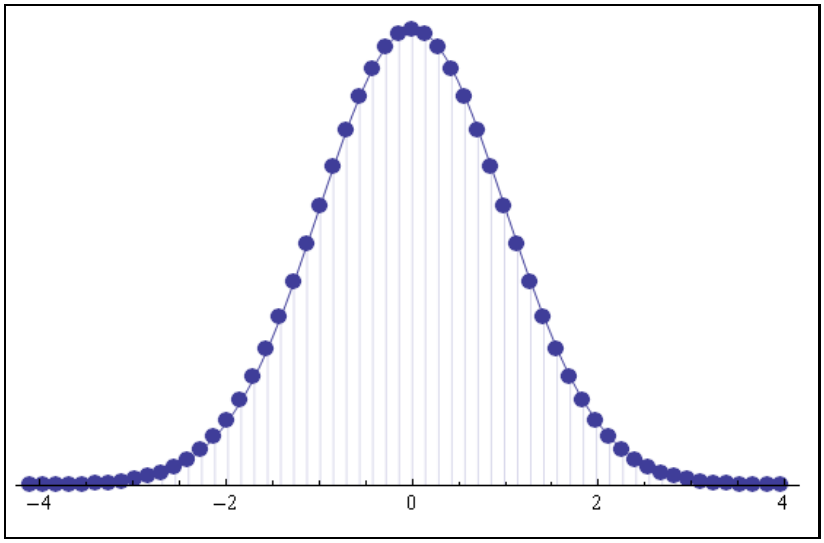
Die Binomialverteilung mit Parametern n und p beschreibt die Verteilung der Anzahl derjenigen unter n unabhängigen Ereignissen mit Wahrscheinlichkeit p, die in einem Zufallsexperiment eintreten. Viele Anwendungsprobleme führen daher auf die Berechnung von Wahrscheinlichkeiten bzgl. der Binomialverteilung. Für große n ist eine exakte Berechnung dieser Wahrscheinlichkeiten aber in der Regel nicht mehr möglich. Bei seltenen Ereignissen kann man die Poissonapproximation zur näherungsweisen Berechnung nutzen:

Konvergiert $n \to \infty$, und konvergiert gleichzeitig der Erwartungswert $n \cdot p_n$ gegen eine positive reelle Zahl $\lambda > 0$, dann nähern sich die Gewichte $b_{n,p_n}(k)$ der Binomialverteilung denen einer Poissonverteilung mit Parameter λ an:

$$b_{n,p_n}(k) \;=\; \binom{n}{k}p_n^k(1-p_n)^{n-k} \;\to\; \frac{\lambda^k}{k}e^{-\lambda} \qquad (k=0,1,2,\ldots),$$

siehe Satz **??**. Geht die Wahrscheinlichkeit p_n für $n \to \infty$ nicht gegen 0, sondern hat zum Beispiel einen festen Wert $p \in (0,1)$, dann kann die Poissonapproximation nicht verwendet werden. Stattdessen scheinen sich die Gewichte der Binomialverteilung einer Gaußschen Glockenkurve anzunähern, wie z.B. die folgende mit Mathematica erstellte Grafik zeigt:

```
Manipulate[
 ListPlot[
  Table[{k, PDF[BinomialDistribution[n, Min[1, lambda /n]], k]}, {k, 0,
     IntegerPart[4 lambda]}],
  Filling -> Axis, PlotRange -> All,
  PlotMarkers -> {Automatic, Medium}, Axes -> {True, False}] , {{n, 10,
     "n"}, 3, 300,1},
  {{lambda, 5, "Erwartungswert: np=Lambda"}, 2, 20}]
```



Wir wollen diese Aussage nun mathematisch präzisieren und beweisen.

1.5.1 Der Grenzwertsatz von De Moivre - Laplace

Wir analysieren zunächst das asymptotische Verhalten von Binomialkoeffizienten mithilfe der Stirlingschen Formel.

Definition (Asymptotische Äquivalenz von Folgen). *Zwei Folgen $a_n, b_n \in \mathbb{R}_+$ ($n \in \mathbb{N}$), heißen* ***asymptotisch äquivalent*** *$(a_n \sim b_n)$, falls*

$$\lim_{n\to\infty} \frac{a_n}{b_n} = 1 \qquad \textit{gilt.}$$

Bemerkung. Für Folgen $a_n, b_n, c_n, d_n \in \mathbb{R}_+$ gilt :

(1). $a_n \sim b_n \iff \exists\, \varepsilon_n \to 0 : a_n = b_n(1+\varepsilon_n) \iff \log a_n - \log b_n \to 0$,

(2). $a_n \sim b_n \iff b_n \sim a_n \iff \frac{1}{a_n} \sim \frac{1}{b_n}$,

(3). $a_n \sim b_n, c_n \sim d_n \implies a_n \cdot c_n \sim b_n \cdot d_n$.

Satz 1.18 (Stirlingsche Formel).

$$n! \;\sim\; \sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n$$

Einen Beweis der Stirlingschen Formel findet man in vielen Analysis-Büchern, siehe z.B. Forster:„Analysis 1". Dass der Quotient der beiden Terme in der Stirlingschen Formel beschränkt ist, sieht man rasch durch eine Integralapproximation von $\log n!$:

$$\begin{aligned}
\log n! \;&=\; \sum_{k=1}^{n} \log k \;=\; \int_0^n \log x\, dx \;+\; \sum_{k=1}^{n} \int_{k-1}^{k} (\log k - \log x)\, dx \\
&=\; n\log n - n \;+\; \frac{1}{2}\sum_{k=1}^{n} \frac{1}{k} \;+\; O(1) \\
&=\; n\log n - n \;+\; \frac{1}{2}\log n \;+\; O(1),
\end{aligned}$$

wobei wir im vorletzten Schritt die Taylor-Approximationen $\log k - \log x = \frac{1}{k}(k-x) + O(k^{-2})$ auf den Intervallen $[k-1, k]$ verwendet haben. Ein alternativer, sehr kurzer vollständiger Beweis der Stirling-Formel mit Identifikation des korrekten asymptotischen Vorfaktors $\sqrt{2\pi}$ wird in J.M. Patin, The American Mathematical Monthly 96 (1989) gegeben.

Mithilfe der Stirlingschen Formel können wir die Gewichte

$$b_{n,p}(k) \;=\; \binom{n}{k} p^k (1-p)^{n-k}$$

der Binomialverteilung für große n und k approximieren. Sei dazu S_n eine $\mathrm{Bin}(n,p)$-verteilte Zufallsvariable auf $(\Omega, \mathcal{A}, P)$. Für den Erwartungswert und die Standardabweichung von S_n gilt:

$$E[S_n] = np \quad \text{und} \quad \sigma[S_n] = \sqrt{\mathrm{Var}[S_n]} = \sqrt{np(1-p)}.$$

Dies deutet darauf hin, dass sich die Masse der Binomialverteilung für große n überwiegend in einer Umgebung der Größenordnung $O(\sqrt{n})$ um np konzentriert.

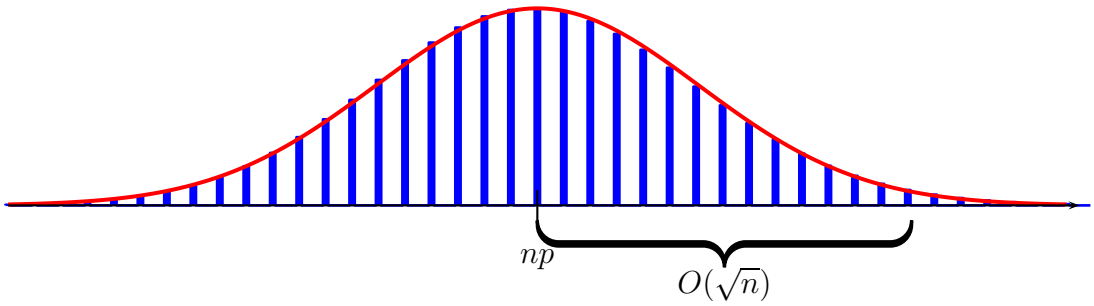


Abbildung 1.12: Die Gewichte der Binomialverteilung liegen für große n näherungsweise auf einer Glockenkurve mit Mittel np und Standardabweichung $\sqrt{np(1-p)}$.

Wir werden nun mithilfe der Stirlingschen Formel die Gewichte

$$b_{n,p}(k) \;=\; P[S_n = k] \;=\; \binom{n}{k} p^k (1-p)^{n-k}$$

der Binomialverteilung für große n und k in Umgebungen der Größenordnung $O(\sqrt{n})$ von np ausgehend von der Stirlingschen Formel approximieren, und die vermutete asymptotische Darstellung präzisieren und beweisen.

Dazu führen wir noch folgende Notation ein: Wir schreiben

$$a_n(k) \approx b_n(k) \qquad (\text{„lokal gleichmäßig asymptotisch äquivalent“}),$$

falls

$$\sup_{k \in U_{n,r}} \left| \frac{a_n(k)}{b_n(k)} - 1 \right| \to 0 \quad \text{für alle } r \in \mathbb{R}_+ \text{ gilt,}$$

wobei

$$U_{n,r} \quad = \quad \{0 \le k \le n : |k - np| \le r \cdot \sqrt{n}\}.$$

Die Aussagen aus der Bemerkung oben gelten analog für diese Art der lokal gleichmäßigen asymptotischen Äquivalenz von $a_n(k)$ und $b_n(k)$.

Satz 1.19 (Satz von de Moivre (1733) und Laplace (1819)). *Sei S_n binomialverteilt mit Parametern $n \in \mathbb{N}$ und $p \in (0,1)$, und sei $\sigma^2 = p(1-p)$. Dann gilt:*

(1). $$P[S_n = k] \quad = \quad b_{n,p}(k) \quad \approx \quad \widetilde{b}_{n,p}(k) \quad := \quad \frac{1}{\sqrt{2\pi n \sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(\frac{k-np}{\sqrt{n}} \right)^2 \right).$$

(2). $$P\left[a \le \frac{S_n - np}{\sqrt{n}} \le b \right] \quad \overset{n \nearrow \infty}{\longrightarrow} \quad \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{x^2}{2\sigma^2}} \, dx \qquad \textit{für alle } a, b \in \mathbb{R} \textit{ mit } a \le b.$$

Beweis. (1). Wir beweisen die Aussage in zwei Schritten:

(a) Wir zeigen zunächst mithilfe der **Stirlingschen Formel**, dass

$$b_{n,p}(k) \quad \approx \quad \overline{b}_{n,p}(k) \quad := \quad \frac{1}{\sqrt{2\pi n \frac{k}{n}(1-\frac{k}{n})}} \cdot \left(\frac{p}{k/n}\right)^k \cdot \left(\frac{1-p}{1-k/n}\right)^{n-k} \tag{1.5.1}$$

gilt. Nach der Stirlingschen Formel ist

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} \quad = \quad 1.$$

Wegen $k \geq np - r \cdot \sqrt{n}$ für $k \in U_{n,r}$ folgt

$$\sup_{k\in U_{n,r}} \left| \frac{k!}{\sqrt{2\pi k}\left(\frac{k}{e}\right)^k} - 1 \right| \quad \longrightarrow \quad 0 \qquad \text{für } n \to \infty,$$

d.h.

$$k! \quad \approx \quad \sqrt{2\pi k}\left(\frac{k}{e}\right)^k.$$

Analog erhält man

$$(n-k)! \quad \approx \quad \sqrt{2\pi(n-k)}\left(\frac{n-k}{e}\right)^{n-k},$$

und damit

$$\begin{aligned} b_{n,p}(k) \quad &= \quad \frac{n!}{k!\cdot(n-k)!}\, p^k (1-p)^{n-k} \quad \approx \quad \frac{\sqrt{2\pi n}\cdot n^n \cdot p^k \cdot (1-p)^{n-k}}{2\pi\sqrt{k(n-k)}\cdot k^k \cdot (n-k)^{n-k}} \\ &= \quad \sqrt{\frac{n}{2\pi k(n-k)}}\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k} \quad = \quad \overline{b}_{n,p}(k). \end{aligned}$$

(b) Wir zeigen nun weiterhin mithilfe einer **Taylorapproximation**, dass

$$\overline{b}_{n,p}(k) \quad \approx \quad \widetilde{b}_{n,p}(k) \tag{1.5.2}$$

gilt. Dazu benutzen wir mehrfach die Abschätzung

$$\left|\frac{k}{n} - p\right| \quad \leq \quad r \cdot n^{-\frac{1}{2}} \qquad \text{für } k \in U_{n,r}.$$

Hieraus folgt zunächst unmittelbar

$$\sqrt{2\pi n \,\frac{k}{n}\left(1-\frac{k}{n}\right)} \quad \approx \quad \sqrt{2\pi np(1-p)} \quad = \quad \sqrt{2\pi n\sigma^2}. \tag{1.5.3}$$

Um die Asymptotik der übrigen Faktoren von $\overline{b}_{n,p}(k)$ zu erhalten, nehmen wir den Logarithmus, und verwenden die **Taylorapproximation**

$$x \log \frac{x}{p} \quad = \quad x - p + \frac{1}{2p}(x-p)^2 + O(|x-p|^3).$$

Wir erhalten

$$\begin{aligned}
\frac{1}{n} \log \left[\left(\frac{p}{k/n} \right)^k \left(\frac{1-p}{1-k/n} \right)^{n-k} \right] &= -\frac{k}{n} \log \left(\frac{\frac{k}{n}}{p} \right) - \left(1 - \frac{k}{n} \right) \log \left(\frac{1-\frac{k}{n}}{1-p} \right) \\
&= -\frac{1}{2p}(\frac{k}{n} - p)^2 - \frac{1}{2(1-p)}(p - \frac{k}{n})^2 + O(|\frac{k}{n} - p|^3) \\
&= -\frac{1}{2p(1-p)}(p - \frac{k}{n})^2 + O(|\frac{k}{n} - p|^3).
\end{aligned}$$

Wegen $\left|\frac{k}{n} - p\right|^3 \quad \leq \quad r^3 \cdot n^{-\frac{3}{2}}$ für $k \in U_{n,r}$ folgt

$$\log \left(\left(\frac{p}{k/n} \right)^k \left(\frac{1-p}{1-k/n} \right)^{n-k} \right) \quad = \quad -\frac{n}{2\sigma^2} \left(\frac{k}{n} - p \right)^2 + R_{k,n},$$

wobei $|R_{k,n}| \leq$ const. $\cdot\, r^3 n^{-\frac{1}{2}}$ für alle $k \in U_{n,r}$, d.h.

$$\left(\frac{p}{k/n} \right)^k \left(\frac{1-p}{1-k/n} \right)^{n-k} \quad \approx \quad \exp \left(-\frac{n}{2\sigma^2} \left(\frac{k}{n} - p \right)^2 \right). \tag{1.5.4}$$

Aussage (1.5.2) folgt dann aus (1.5.3) und (1.5.4).

(c) Aus (a) und (b) folgt nun Behauptung (1).

(2). Aufgrund von (1) erhalten wir für $a, b \in \mathbb{R}$ mit $a < b$:

$$P \left[a \leq \frac{S_n - np}{\sqrt{n}} \leq b \right] \quad = \quad \sum_{\substack{k \in \{0,1,\ldots,n\} \\ a \leq \frac{k-np}{\sqrt{n}} \leq b}} b_{n,p}(k) \quad = \quad \sum_{\substack{k \in \{0,1,\ldots,n\} \\ a \leq \frac{k-np}{\sqrt{n}} \leq b}} \widetilde{b}_{n,p}(k)(1 + \varepsilon_{n,p}(k)),$$

wobei

$$\overline{\varepsilon}_{n,p} \quad := \quad \sup_{a \leq \frac{k-np}{\sqrt{n}} \leq b} |\varepsilon_{n,p}(k)| \quad \longrightarrow \quad 0 \qquad \text{für } n \to \infty. \tag{1.5.5}$$

Wir zeigen nun

$$\lim_{n \to \infty} \sum_{\substack{k \in \{0,1,\ldots,n\} \\ a \leq \frac{k-np}{\sqrt{n}} \leq b}} \widetilde{b}_{n,p}(k) \quad = \quad \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left(-\frac{x^2}{2\sigma^2} \right) dx \tag{1.5.6}$$

Zum Beweis von (1.5.6) bemerken wir, dass

$$\Gamma_n := \left\{ \frac{k-np}{\sqrt{n}} \,\middle|\, k = 0, 1, \ldots, n \right\} \subseteq \mathbb{R}$$

ein äquidistantes Gitter mit Maschenweite $1/\sqrt{n}$ ist. Es gilt

$$\sum_{\substack{k\in\{0,1,\ldots,n\} \\ a\le\frac{k-np}{\sqrt{n}}\le b}} \widetilde{b}_{n,p}(k) \quad = \quad \sum_{\substack{x\in\Gamma_n \\ a\le x\le b}} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{x^2}{2\sigma^2}}\, \frac{1}{\sqrt{n}}\,. \tag{1.5.7}$$

Für $n \to \infty$ folgt (1.5.6), da die rechte Seite in (1.5.7) eine Riemannsummenapproximation des Integrals in (1.5.6) ist, und der Integrand stetig ist.

Die Behauptung folgt nun aus (1.5.5) und (1.5.6). □

Der Satz von de Moivre/Laplace impliziert, dass die Zufallsvariablen $\frac{S_n-np}{\sqrt{n}}$ für $n \to \infty$ *in Verteilung gegen eine $N(0,\sigma^2)$-verteilte Zufallsvariable mit Varianz $\sigma^2 = p(1-p)$ konvergieren*:

$$\frac{S_n - np}{\sqrt{n}} \quad \xrightarrow{\mathcal{D}} \quad \sigma Z \qquad \text{mit } Z \sim N(0,1).$$

Hierbei bedeutet **Konvergenz in Verteilung (convergence in distribution; Notation „$\xrightarrow{\mathcal{D}}$“)**, dass die Verteilungen der Zufallsvariablen **schwach konvergieren**, d.h. für die Verteilungsfunktionen gilt

$$\lim_{n\to\infty} F_{\frac{S_n-np}{\sqrt{n}}}(c) \;=\; \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{x^2}{2\sigma^2}}\, dx \;=\; F_{\sigma Z}(c) \qquad \text{für alle } c \in \mathbb{R}. \tag{1.5.8}$$

Die allgemeine Definition der schwachen Konvergenz einer Folge von Wahrscheinlichkeitsverteilungen wird in Abschnitt 5.1 unten gegeben - für Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$ ist sie äquivalent zur Konvergenz der Verteilungsfunktionen an allen Stellen, an denen die Limes-Verteilungsfunktion stetig ist. Konvergenz in Verteilung ist also nicht wirklich ein Konvergenzbegriff für Zufallsvariablen, sondern „nur” eine Konvergenz der Verteilungen. Für die standardisierten (d.h. auf Erwartungswert 0 und Varianz 1 normierten) Zufallsvariablen gilt entsprechend

$$\frac{S_n - E[S_n]}{\sigma(S_n)} \quad = \quad \frac{S_n - np}{\sigma\sqrt{n}} \quad \xrightarrow{\mathcal{D}} \quad Z. \tag{1.5.9}$$

Bemerkung. (1). Die Aussage (1.5.9) ist ein Spezialfall eines viel allgemeineren zentralen Grenzwertsatzes:

Sind $X_1, X_2, \ldots$ unabhängige, identisch verteilte Zufallsvariablen mit endlicher Varianz,

und ist $S_n = X_1 + \ldots + X_n$, dann konvergieren die Verteilungen der standardisierten Summen

$$\frac{S_n - E[S_n]}{\sigma(S_n)}$$

schwach gegen eine Standardnormalverteilung, s.u.

Die Normalverteilung tritt also als universeller Skalierungslimes von Summen unabhängiger Zufallsvariablen auf.

(2). Heuristisch gilt für große n nach (1.5.9)

$$\text{„} \quad S_n \quad \overset{\mathcal{D}}{\approx} \quad np + \sqrt{np(1-p)} \cdot Z, \quad \text{“} \tag{1.5.10}$$

wobei „$\overset{\mathcal{D}}{\approx}$“ dafür steht, dass sich die Verteilungen der Zufallsvariablen einander in einem gewissen Sinn annähern. In diesem Sinne wäre für große n

$$\text{„}\mathrm{Bin}(n,p) \quad \overset{\mathcal{D}}{\approx} \quad N(np, np(1-p))\text{.“}$$

Entsprechende „Approximationen“ werden häufig in Anwendungen benutzt, sollten aber hinterfragt werden, da beim Übergang von (1.5.9) zu (1.5.10) mit dem divergierende Faktor $\sqrt{n}$ multipliziert wird. Die mathematische Präzisierung entsprechender heuristischer Argumentationen erfolgt üblicherweise über den Satz von de Moivre/Laplace.

Beispiel (Faire Münzwürfe). Seien $X_1, X_2, \ldots$ unabhängige Zufallsvariablen mit $P[X_i = 0] = P[X_i = 1] = \frac{1}{2}$, und sei $S_n = X_1 + \ldots + X_n$ (z.B. Häufigkeit von „Zahl“ bei n fairen Münzwürfen). In diesem Fall ist also $p = \frac{1}{2}$ und $\sigma = \sqrt{p(1-p)} = \frac{1}{2}$.

(1). *100 faire Münzwürfe:* Für die Wahrscheinlichkeit, dass mehr als 60 mal Zahl fällt, gilt

$$P[S_{100} > 60] \quad = \quad P[S_{100} - E[S_{100}] > 10] \quad = \quad P\left[\frac{S_{100} - E[S_{100}]}{\sigma(S_{100})} > \frac{10}{\sigma\sqrt{100}}\right].$$

Da $\frac{S_{100} - E[S_{100}]}{\sigma(S_{100})}$ nach (1.5.9) näherungsweise $N(0,1)$-verteilt ist, und $\frac{10}{\sigma\sqrt{100}} = 2$, folgt

$$P[S_{100} > 60] \quad \approx \quad P[Z > 2] \quad = \quad 1 - \Phi(2) \quad \approx \quad 0.0227 \quad = \quad 2.27\%.$$

(2). *16 faire Münzwürfe*: Für die Wahrscheinlichkeit, dass genau 8 mal Zahl fällt, ergibt sich

$$P[S_{16} = 8] \;=\; P[7.5 \le S_{16} \le 8.5] \;=\; P\left[\left|\frac{S_{16} - E[S_{16}]}{\sigma(S_{16})}\right| \le \frac{0.5}{\sigma\sqrt{16}}\right]$$

Mit $\frac{0.5}{\sigma\sqrt{16}} = \frac{1}{4}$ folgt näherungsweise

$$P[S_{16} = 8] \quad \approx \quad P[|Z| \le 1/4] \quad = \quad 0.1974...$$

Der exakte Wert beträgt $P[S_{16} = 8] = 0.1964...$. Bei geschickter Anwendung ist die Normalapproximation oft schon für eine kleine Anzahl von Summanden relativ genau!

1.5.2 Approximative Konfidenzintervalle

Angenommen, wir wollen den Anteil p der Wähler einer Partei durch Befragung von n Wählern schätzen. Seien $X_1, \ldots, X_n$ unter P_p unabhängige und Bernoulli(p)-verteilte Zufallsvariablen, wobei $X_i = 1$ dafür steht, dass der i-te Wähler für die Partei A stimmen wird. Ein nahe liegender Schätzwert für p ist $\overline{X_n} := \frac{S_n}{n}$. Wie viele Stichproben braucht man, damit der tatsächliche Stimmenanteil mit 95% Wahrscheinlichkeit um höchstens $\varepsilon = 1\%$ von Schätzwert abweicht?

Definition (Konfidenzintervall). *Sei $\alpha \in (0,1)$. Das zufällige Intervall $[\overline{X_n} - \varepsilon, \overline{X_n} + \varepsilon]$ heißt Konfidenzintervall zum Konfidenzniveau $1-\alpha$ (bzw. zum Irrtumsniveau α) für den unbekannten Parameter p, falls*

$$P_p[p \notin [\overline{X_n} - \varepsilon, \overline{X_n} + \varepsilon]] \quad \leq \quad \alpha$$

*für **alle** möglichen Parameterwerte $p \in [0,1]$ gilt.*

Im Prinzip lassen sich Konfidenzintervalle aus den Quantilen der zugrundeliegenden Verteilung gewinnen. In der Situation von oben gilt beispielsweise:

$$\begin{aligned} p \in [\overline{X_n} - \varepsilon, \overline{X_n} + \varepsilon] \quad &\iff \quad |\overline{X_n} - p| \leq \varepsilon \iff \overline{X_n} \in [p-\varepsilon, p+\varepsilon] \\ &\iff \quad S_n \in [n(p-\varepsilon), n(p+\varepsilon)] \end{aligned}$$

Diese Bedingung ist für $p \in [0,1]$ mit Wahrscheinlichkeit $\geq 1-\alpha$ erfüllt, falls z.B. $n(p-\varepsilon)$ oberhalb des $\frac{\alpha}{2}$-Quantils und $n(p+\varepsilon)$ unterhalb des $(1-\frac{\alpha}{2})$-Quantils der Binomialverteilung $\mathrm{Bin}(n,p)$ liegt.

Praktikablere Methoden, um in unserem Modell Konfidenzintervalle zu bestimmen, sind zum Beispiel:

Abschätzung mithilfe der Čebyšev-Ungleichung:

$$P_p\left[\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right] \quad \leq \quad \frac{1}{\varepsilon^2} \cdot Var\left(\frac{S_n}{n}\right) \quad = \quad \frac{p(1-p)}{n\varepsilon^2} \quad \leq \quad \frac{1}{4n\varepsilon^2} \quad \overset{!}{\leq} \quad \alpha \qquad \forall\, p \in [0,1]$$

Dies ist erfüllt für $n \geq \frac{1}{4\varepsilon^2\alpha}$, also im Beispiel für $n \geq 50.000$.

Abschätzung über die exponentielle Ungleichung:

$$P_p\left[\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right] \quad \leq \quad 2 \cdot e^{-2\varepsilon^2 n} \quad \leq \quad \alpha \qquad \forall\, p \in [0,1],$$

ist erfüllt für $n \geq \frac{1}{2\varepsilon^2} \log(\frac{2}{\alpha})$, also im Beispiel für $n \geq 18445$.

Die exponentielle Abschätzung ist genauer - sie zeigt, dass bereits weniger als 20.000 Stichproben genügen. Können wir mit noch weniger Stichproben auskommen ? Dazu berechnen wir die Wahrscheinlichkeit, dass der Parameter im Intervall liegt, näherungsweise mithilfe des zentralen Grenzwertsatzes:

Approximative Berechnung mithilfe der Normalapproximation:

$$\begin{aligned} P_p\left[\left|\frac{S_n}{n} - p\right| \le \varepsilon\right] &= P_p\left[\left|\frac{S_n - np}{\sqrt{np(1-p)}}\right| \le \frac{n\varepsilon}{\sqrt{np(1-p)}}\right] \\ &\approx N(0,1)\left(-\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\,,\ \frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right) \\ &= 2\left(\Phi\left(\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right) - \frac{1}{2}\right) \\ &\overset{p(1-p)\le\frac{1}{4}}{\ge} 2\Phi(2\sqrt{n}\varepsilon) - 1 \quad \ge \quad 1-\alpha \qquad \forall\, p \in [0,1], \end{aligned}$$

falls

$$n \quad \ge \quad \left(\frac{1}{2\varepsilon}\cdot\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2.$$

Im Beispiel gilt $\Phi^{-1}(1-\alpha/2) \approx 1.96$, und die Bedingung ist für $n \ge 9604$ erfüllt. Also sollten bereits ca. 10.000 Stichproben ausreichen! *Exakte* (also ohne Verwendung einer Näherung hergeleitete) Konfidenzintervalle sind in vielen Fällen zu konservativ. In Anwendungen werden daher meistens *approximative* Konfidenzintervalle angegeben, die mithilfe einer Normalapproximation hergeleitet wurden. Dabei ist aber folgendes zu beachten:

Warnung: Mithilfe der Normalapproximation hergeleitete approximative Konfidenzintervalle erfüllen die Niveaubedingung im Allgemeinen nicht (bzw. nur näherungsweise). Da die Qualität der Normalapproximation für $p \to 0$ bzw. $p \to 1$ degeneriert, ist die Niveaubedingung im Allgemeinen selbst für $n \to \infty$ nicht erfüllt. Beispielsweise beträgt das Niveau von approximativen 99% Konfidenzintervallen asymptotisch tatsächlich nur 96.8%!

1.6 Transformationen von reellwertigen Zufallsvariablen

In diesem Abschnitt betrachten wir zunächst Transformationen von absolutstetigen Zufallsvariablen.

Transformationen kann man auch benutzen, um reelle Zufallsvariablen mit einer vorgegebenen Verteilung μ aus gleichverteilten Zufallsvariablen zu erzeugen. Ist die Verteilungsfunktion F :

$\mathbb{R} \to [0,1]$ streng monoton wachsend und stetig, also eine Bijektion von $\mathbb{R}$ nach $(0,1)$, und ist $U : \Omega \to (0,1)$ auf $(0,1)$ gleichverteilt, dann hat die Zufallsvariable $F^{-1}(U)$ die Verteilung μ, denn es gilt

$$P[F^{-1}(U) \le c] \;=\; P[U \le F(c)] \;=\; F(c) \qquad \text{für alle } c \in \mathbb{R}.$$

Das beschriebene Inversionsverfahren werden wir in Satz 1.22 erweitern, um reelle Zufallsvariablen mit beliebigen Verteilungen zu konstruieren. Da die Verteilungsfunktion dann im Allgemeinen keine Bijektion ist, verwenden wir statt der Inversen F^{-1} eine verallgemeinerte (linksstetige) Inverse, die durch die Quantile der zugrundeliegenden Verteilung bestimmt ist.

1.6.1 Transformation von Dichten

Wir haben in Beispielen bereits mehrfach die Verteilung von Funktionen von absolutstetigen Zufallsvariablen berechnet. Sei nun allgemein $I = (a,b) \subseteq \mathbb{R}$ ein offenes Intervall, und $X : \Omega \to I$ eine Zufallsvariable mit stetiger Verteilung.

Satz 1.20 (Eindimensionaler Dichtetransformationssatz). *Ist $\phi : I \to \mathbb{R}$ einmal stetig differenzierbar mit $\phi'(x) \neq 0$ für alle $x \in I$, dann ist die Verteilung von $\phi(X)$ absolutstetig mit Dichte*

$$f_{\phi(X)}(y) \quad = \quad \begin{cases} f_X(\phi^{-1}(y)) \cdot |(\phi^{-1})'(y)| & \text{für } y \in \phi(I), \\ 0 & \text{sonst.} \end{cases} \tag{1.6.1}$$

Beweis. Nach der Voraussetzung gilt entweder $\phi' > 0$ auf I oder $\phi' < 0$ auf I. Wir betrachten nur den ersten Fall. Aus $\phi' > 0$ folgt, dass ϕ streng monoton wachsend ist, also eine Bijektion von I nach $\phi(I)$. Daher erhalten wir mit der Substitution $y = \phi(x)$:

$$\begin{aligned} F_{\phi(X)}(c) \;&=\; P[\phi(X) \le c] \;=\; P[X \le \phi^{-1}(c)] \;=\; F_X(\phi^{-1}(c)) \\ &=\; \int_a^{\phi^{-1}(c)} f_X(x)\,dx \;=\; \int_{\phi(a)}^{c} f_X(\phi^{-1}(y))\,(\phi^{-1})'(y)\,dy \end{aligned}$$

für alle $c \in \phi(I)$. Die Behauptung folgt wegen $P[\phi(X) \notin \phi(I)] = 0$. □

Beispiel (Geometrische Wahrscheinlichkeiten). Sei $\theta : \Omega \to [0, 2\pi)$ ein zufälliger, auf $[0, 2\pi)$ gleichverteilter, Winkel. Wir wollen die Verteilung von $\cos\theta$ berechnen. Da die Kosinusfunktion auf $[0, 2\pi)$ nicht streng monoton ist, ist (1.6.1) nicht direkt anwendbar. Wir können aber das

Intervall $[0, 2\pi)$ in die Teile $[0, \pi)$ und $[\pi, 2\pi)$ zerlegen, und dann die Verteilung ähnlich wie im Beweis von Satz 1.20 berechnen. Wegen

$$\begin{aligned} P[\cos\theta > c] &= P[\cos\theta > c \text{ und } \theta \in [0,\pi)] \;+\; P[\cos\theta > c \text{ und } \theta \in [\pi, 2\pi)] \\ &= P[\theta \in [0, \arccos c)] \;+\; P[\theta \in [2\pi - \arccos c, 2\pi)] \\ &= \frac{2}{2\pi} \cdot \arccos c \end{aligned}$$

erhalten wir, dass $\cos\theta$ eine absolutstetige Verteilung mit Dichte

$$f_{\cos\theta}(x) \quad = \quad F'_{\cos\theta}(x) \quad = \quad \frac{1}{\pi} \cdot \frac{1}{\sqrt{1-x^2}}; \qquad x \in (-1, 1)$$

hat. Anstelle von (1.6.1) gilt in diesem Fall

$$f_{\cos\theta}(x) \quad = \quad f_X(\psi_1(x)) \cdot |\psi_1'(x)| + f_X(\psi_2(x)) \cdot |\psi_2'(x)|,$$

wobei $\psi_1(x) = \arccos x$ und $\psi_2(x) = 2\pi - \arccos x$ die Umkehrfunktionen auf den Teilintervallen sind. Entsprechende Formeln erhält man auch allgemein, wenn die Transformation nur stückweise bijektiv ist.

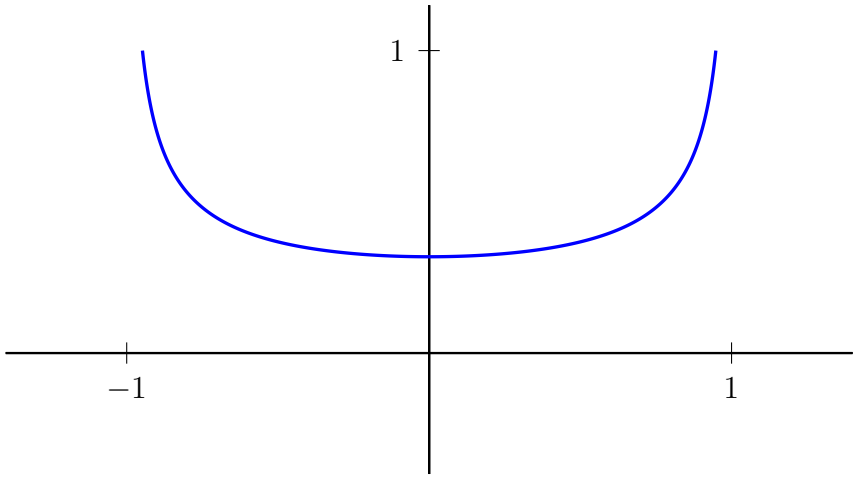


Abbildung 1.13: Graph der Dichtefunktion $f_{\cos\theta}$

Auf ähnliche Weise zeigt man für $a > 0$:

$$f_{a\tan\theta}(x) \quad = \quad \frac{1}{\pi a} \cdot \frac{1}{1 + (x/a)^2}, \qquad x \in \mathbb{R}.$$

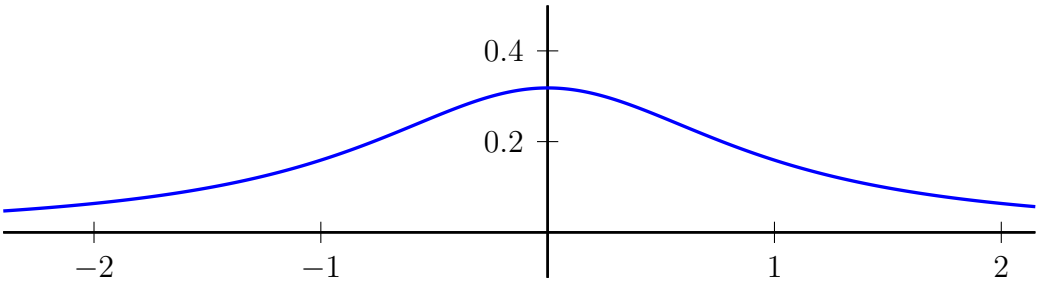
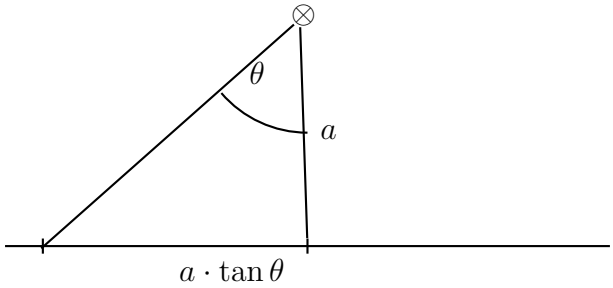


Abbildung 1.14: Graph der Dichtefunktion $f_{\tan\theta}$

Die Verteilung mit dieser Dichte heißt **Cauchyverteilung** zum Parameter a. Sie beschreibt unter anderem die Intensitätsverteilung auf einer Geraden, die von einer in alle Richtungen gleichmäßig strahlenden Lichtquelle im Abstand a bestrahlt wird.



1.6.2 Quantile und Inversion der Verteilungsfunktion

Quantile sind Stellen, an denen die Verteilungsfunktion einen bestimmten Wert überschreitet. Solche Stellen müssen häufig in praktischen Anwendungen (z.B. Qualitätskontrolle) berechnet werden. Mithilfe von Quantilen kann man verallgemeinerte Umkehrfunktionen der im Allgemeinen nicht bijektiven Verteilungsfunktion definieren.

Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit Verteilungsfunktion F.

Definition (Quantile). *Sei $u \in [0,1]$. Dann heißt $q \in \mathbb{R}$ ein u-**Quantil** der Verteilung von X, falls*

$$P[X < q] \leq u \qquad \textit{und} \qquad P[X > q] \leq 1 - u$$

*gilt. Ein $\frac{1}{2}$-Quantil heißt **Median**.*

Ist die Verteilungsfunktion streng monoton wachsend, dann ist $q = F^{-1}(u)$ für $u \in (0,1)$ das einzige u-Quantil. Im Allgemeinen kann es hingegen mehrere u-Quantile zu einem Wert u geben.

Beispiel (Empirische Quantile). Wir betrachten eine Stichprobe, die aus n reellwertigen Daten / Messwerten $x_1, \ldots, x_n$ mit $x_1 \leq x_2 \leq \ldots \leq x_n$ besteht. Die empirische Verteilung der Stichprobe ist die Wahrscheinlichkeitsverteilung

$$\mu = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

auf $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$, d.h. für $B \subseteq \mathbb{R}$ ist

$$\mu[B] = \frac{1}{n} \left| \{x_i \in B, 1 \leq i \leq n\} \right|$$

die relative Häufigkeit des Bereichs B unter den Messwerten x_i. Die empirische Verteilung ergibt sich, wenn wir zufällig ein $i \in \{1, \ldots, n\}$ wählen, und den entsprechenden Messwert betrachten. Die Quantile der empirischen Verteilung bezeichnet man als **Stichprobenquantile**. Ist $n \cdot u$ nicht ganzzahlig, dann ist $x_{\lceil n \cdot u \rceil}$ das eindeutige u-Quantil der Stichprobe. Ist hingegen $u = k/n$ mit $k \in \{1, \ldots, n\}$, dann ist jedes $q \in [x_k, x_{k+1}]$ ein u-Quantil der empirischen Verteilung.

Wir definieren nun zwei verallgemeinerte Inverse einer Verteilungsfunktion F, die ja im Allgemeinen nicht bijektiv ist. Für $u \in (0, 1)$ sei

$$\begin{aligned} \underline{G}(u) &:= \inf\{x \in \mathbb{R} : F(x) \geq u\} = \sup\{x \in \mathbb{R} : F(x) < u\}, \qquad \text{und} \\ \overline{G}(u) &:= \inf\{x \in \mathbb{R} : F(x) > u\} = \sup\{x \in \mathbb{R} : F(x) \leq u\}. \end{aligned}$$

Offensichtlich gilt $\underline{G}(u) \leq \overline{G}(u)$. Die Funktionen $\underline{G}$ bzw. $\overline{G}$ sind links- bzw. rechtsstetig. Ist F stetig und streng monoton wachsend, also eine Bijektion von $\mathbb{R}$ nach $(0, 1)$, dann ist $\underline{G}(u) = \overline{G}(u) = F^{-1}(u)$. Die Funktion $\underline{G}$ heißt daher auch die **linksstetige verallgemeinerte Inverse** von F. Das folgende Lemma zeigt, dass $\underline{G}(u)$ das kleinste und $\overline{G}(u)$ das größte u-Quantil ist:

Lemma 1.21. *Für $u \in (0, 1)$ und $q \in \mathbb{R}$ sind die folgenden Aussagen äquivalent:*

(1). q ist ein u-Quantil.

(2). $F(q-) \leq u \leq F(q)$.

(3). $\underline{G}(u) \leq q \leq \overline{G}(u)$.

Hierbei ist $F(q-) := \lim_{y \nearrow q} F(y)$ der linksseitige Limes von F an der Stelle q.

Beweis. Nach Definition ist q genau dann ein u-Quantil, wenn

$$P[X < q] \leq u \leq 1 - P[X > q] = P[X \leq q]$$

gilt. Hieraus folgt die Äquivalenz von (1) und (2).

Um zu beweisen, dass (3) äquivalent zu diesen Bedingungen ist, müssen wir zeigen, dass $\underline{G}(u)$ das kleinste und $\overline{G}(u)$ das größte u-Quantil ist. Wir bemerken zunächst, dass $\underline{G}(u)$ ein u-Quantil ist, da

$$F(\underline{G}(u)-) = \lim_{x \nearrow \underline{G}(u)} \underbrace{F(x)}_{\substack{<u \\ \text{für } x<\underline{G}(u)}} \leq u,$$

und

$$F(\underline{G}(u)) = \lim_{x \searrow \underline{G}(u)} \underbrace{F(x)}_{\substack{\geq u \\ \text{für } x>\underline{G}(u)}} \geq u.$$

Andererseits gilt für $x < \underline{G}(u)$:

$$F(x) < u,$$

d.h. x ist kein u-Quantil. Somit ist $\underline{G}(u)$ das kleinste u-Quantil. Auf ähnliche Weise folgt, dass $\overline{G}(u)$ das größte u-Quantil ist. □

1.6.3 Konstruktion und Simulation reellwertiger Zufallsvariablen

Wie erzeugt man ausgehend von auf $(0,1)$ gleichverteilten Zufallszahlen Stichproben von anderen Verteilungen μ auf $\mathbb{R}^1$?

Beispiel (Endlicher Fall). Gilt $\mu[S] = 1$ für eine endliche Teilmenge $S \subseteq \mathbb{R}$, dann können wir die Frage leicht beantworten: Sei $S = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}$ mit $n \in \mathbb{N}$ und $x_1 < x_2 < \ldots < x_n$. Die Verteilungsfunktion einer Wahrscheinlichkeitsverteilung μ auf S ist gegeben durch

$$F(c) \;=\; \mu[(-\infty, c]] \;=\; \sum_{i: x_i \leq c} \mu[\{x_i\}].$$

Ist U eine auf $(0,1)$ gleichverteilte Zufallsvariable, dann wird durch

$$X(\omega) := x_k \qquad \text{falls } F(x_{k-1}) < U(\omega) \leq F(x_k), \quad x_0 := -\infty,$$

eine Zufallsvariable mit Verteilung μ definiert, denn für $k = 1, \ldots, n$ gilt

$$P[X = x_k] = F(x_k) - F(x_{k-1}) = \mu[\{x_k\}].$$

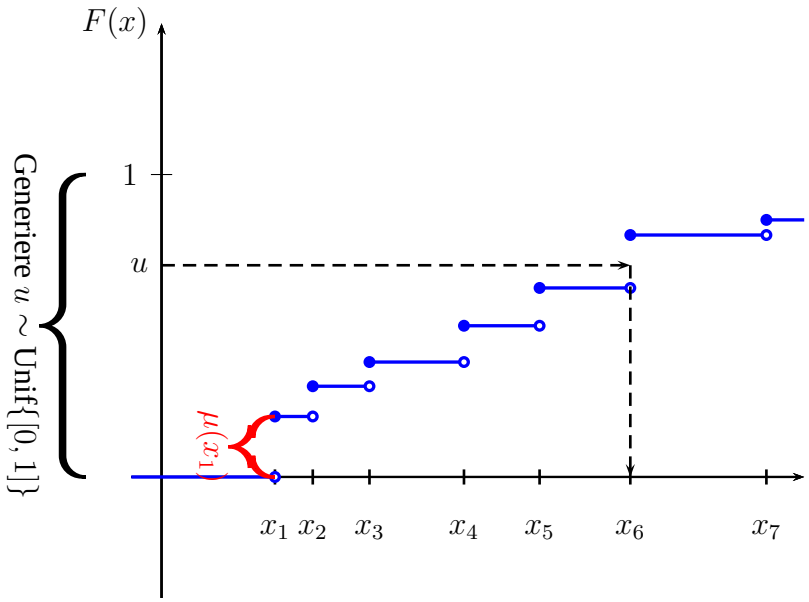


Abbildung 1.15: Wir generieren eine auf $(0,1)$ gleichverteilte Zufallszahl u. Suche nun das minimale $k \in \mathbb{N}$, für das $\sum_{i=1}^{k} \mu(x_i) \geq u$. Dann ist $x = x_k$ eine Stichprobe von der Verteilung μ.

Wir wollen das Vorgehen aus dem Beispiel nun verallgemeinern. Sei $F : \mathbb{R} \to [0,1]$ eine Funktion mit den Eigenschaften

(1). monoton wachsend: $F(x) \leq F(y) \quad \forall\, x \leq y$

(2). rechtsstetig: $\lim_{x \downarrow c} F(x) = F(c) \quad \forall\, c \in \mathbb{R}$

(3). normiert: $\lim_{x \searrow -\infty} F(x) = 0 \quad , \quad \lim_{x \nearrow +\infty} F(x) = 1.$

Das folgende Resultat liefert eine explizite Konstruktion einer Zufallsvariable mit Verteilungsfunktion F:

Satz 1.22 (Quantiltransformation). *Ist $F : \mathbb{R} \to [0,1]$ eine Funktion mit (1)-(3), und*

$$\underline{G}(u) = \inf\{x \in \mathbb{R} : F(x) \geq c\}, \quad u \in (0,1),$$

die linksstetige verallgemeinerte Inverse, dann ist das Bild $\mu := \mathcal{U}_{(0,1)} \circ \underline{G}^{-1}$ der Gleichverteilung auf $(0,1)$ unter $\underline{G}$ ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ mit Verteilungsfunktion F.

Insbesondere gilt: Ist $U : \Omega \to (0,1)$ eine unter P gleichverteilte Zufallsvariable, dann hat die Zufallsvariable

$$X(\omega) \;:=\; \underline{G}(U(\omega))$$

unter P die Verteilungsfunktion F.

Beweis. Da $\underline{G}(u)$ ein u-Quantil ist, gilt $F(\underline{G}(u)) \geq u$, also

$$\underline{G}(u) \;=\; \min\{x \in \mathbb{R} : F(x) \geq u\},$$

und somit für $c \in \mathbb{R}$:

$$\underline{G}(u) \leq c \iff F(x) \geq u \text{ für ein } x \leq c \iff F(c) \geq u.$$

Es folgt:

$$P[\underline{G}(U) \leq c] \;=\; \mathcal{U}_{(0,1)}[\{u \in (0,1) : \underbrace{\underline{G}(u) \leq c}_{\iff F(c) \geq u}\}] \;=\; F(c).$$

Also ist F die Verteilungsfunktion von $G(U)$ bzw. von μ. □

Bemerkung. Nimmt X nur endlich viele Werte $x_1 < x_2 < \ldots < x_n$ an, dann ist F stückweise konstant, und es gilt:

$$\underline{G}(u) = x_k \text{ für } F(x_{k-1}) < u \leq F(x_k), \quad x_0 := -\infty,$$

d.h. $\underline{G}$ ist genau die oben im endlichen Fall verwendete Transformation.

Das Resultat liefert einen

Existenzsatz. Zu jeder Funktion F mit (1)-(3) existiert eine reelle Zufallsvariable X bzw. eine Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}$ mit Verteilungsfunktion F.

Zudem erhalten wir einen expliziten Algorithmus zur Simulation einer Stichprobe von μ:

Algorithmus 1.23 (Inversionsverfahren zur Simulation einer Stichprobe x von μ).

(1). Erzeuge (Pseudo)-Zufallszahl $u \in (0,1)$.

(2). Setze $x := \underline{G}(u)$.

Dieser Algorithmus funktioniert theoretisch immer. Er ist aber oft nicht praktikabel, da man $\underline{G}$ nicht immer berechnen kann, oder da das Anwenden der Transformation $\underline{G}$ (zunächst unwesentliche) Schwachstellen des verwendeten Zufallsgenerators verstärkt. Man greift daher oft selbst im eindimensionalen Fall auf andere Simulationsverfahren wie z.B. „Acceptance Rejection“ Methoden zurück.

Beispiele. (1). BERNOULLI(p)-VERTEILUNG AUF $\{0,1\}$. Hier gilt

$$F = (1-p) \cdot I_{[0,1)} + 1 \cdot I_{[1,\infty)} \qquad \text{und} \qquad \underline{G} = I_{(1-p,1)}.$$

Also ist die Zufallsvariable $\underline{G}(U) = I_{\{U<1-p\}}$ für $U \sim \mathcal{U}_{(0,1)}$ Bernoulli(p)-verteilt.

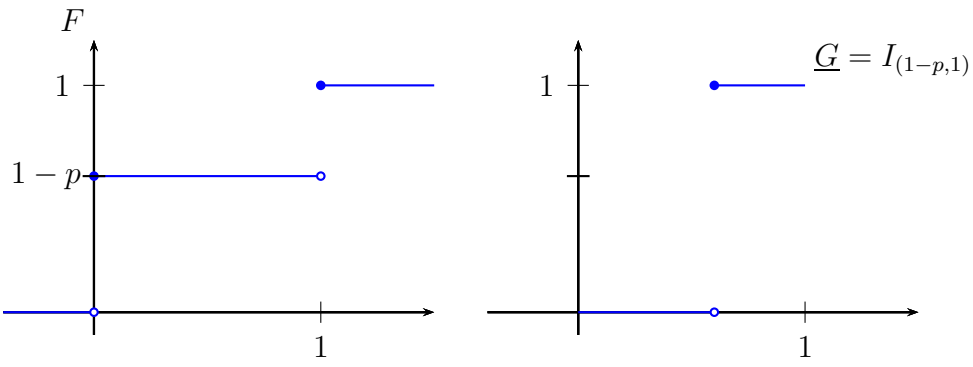


Abbildung 1.16: $\underline{G}(U) = I_{\{U>1-p\}}$ ist Bernoulli(p)-verteilt.

(2). GLEICHVERTEILUNG AUF (a, b). Hier ist $F(c) = \frac{c-a}{b-a}$ für $c \in [a, b]$, und damit

$$\underline{G}(u) = a + (b - a)u,$$

siehe Abbildung 1.17. Also ist $a + (b - a)U$ für $U \sim \mathcal{U}_{(0,1)}$ gleichverteilt auf (a, b).

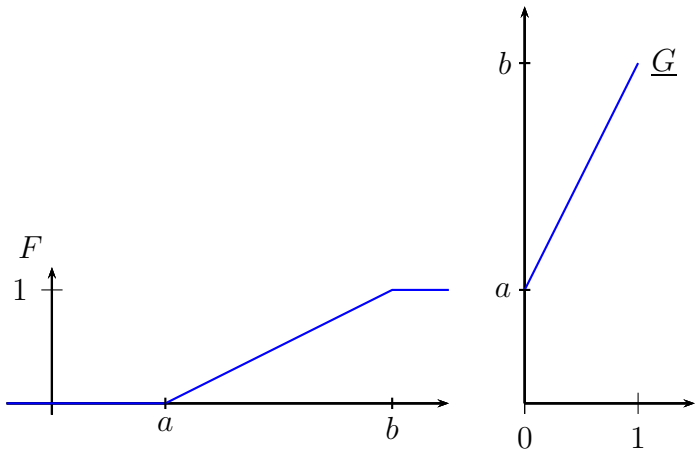


Abbildung 1.17: $\underline{G}(U) = a + (b - a)U$ ist auf (a, b) gleichverteilt.

(3). EXPONENTIALVERTEILUNG MIT PARAMETER $\lambda > 0$:

$$F(c) = 1 - e^{-\lambda c}, \quad G(u) = F^{-1}(u) = -\frac{1}{\lambda}\log(1 - u).$$

Anwenden des negativen Logarithmus transformiert also die gleichverteilte Zufallsvariable $1 - U$ in eine exponentialverteilte Zufallsvariable.

Kapitel 2

Unabhängigkeit und Produktmodelle

2.1 Unabhängigkeit

2.1.1 Unabhängigkeit von Ereignissen

In Abschnitt **??** haben wir einen Unabhängigkeitsbegriff für Ereignisse eingeführt: Eine Kollektion $A_i, i \in I$, von Ereignissen aus derselben σ-Algebra $\mathcal{A}$ heißt **unabhängig** bzgl. einer Wahrscheinlichkeitsverteilung P, falls

$$P[A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_n}] \quad = \quad \prod_{k=1}^{n} P[A_{i_k}] \tag{2.1.1}$$

für alle $n \in \mathbb{N}$ und alle paarweise verschiedenen $i_1, \ldots, i_n \in I$ gilt.

Beispiel. Ein Ereignis A ist genau dann unabhängig von sich selbst, wenn $P[A] = P[A \cap A] = P[A]^2$ gilt, also wenn die Wahrscheinlichkeit von A gleich 0 oder 1 ist. Solche Ereignisse nennt man auch deterministisch.

Wir wollen den obigen Unabhängigkeitsbegriff nun auf Ereignissysteme erweitern.

Definition (Unabhängigkeit von Mengensystemen). *Eine Kollektion $\mathcal{A}_i$ $(i \in I)$ von Mengensystemen $\mathcal{A}_i \subseteq \mathcal{A}$ heißt **unabhängig (bzgl. P)**, falls jede Kollektion A_i $(i \in I)$ von Ereignissen $A_i \in \mathcal{A}_i$ unabhängig ist, d.h.*

$$P[A_{i_1} \cap \ldots \cap A_{i_n}] \quad = \quad \prod_{k=1}^{n} P[A_{i_k}]$$

für alle $n \in \mathbb{N}, i_1, \ldots, i_n \in I$ paarweise verschieden, und $A_{i_k} \in \mathcal{A}_{i_k}$ $(1 \leq k \leq n)$.

Sind zum Beispiel A und B unabhängige Ereignisse, dann sind $\sigma(A) = \{\emptyset, A, A^C, \Omega\}$ und $\sigma(B) = \{\emptyset, B, B^C, \Omega\}$ unabhängige Mengensysteme. Viel allgemeiner gilt:

Satz 2.1 (Kompositions- und Gruppierungssatz für unabhängige Ereignisse). *Seien $\mathcal{A}_i$ $(i \in I)$ unabhängige Mengensysteme. Jedes $\mathcal{A}_i$ sei durchschnittsstabil. Dann folgt:*

(1). Die σ-Algebren $\sigma(\mathcal{A}_i)$ $(i \in I)$ sind unabhängige Mengensysteme.

(2). Ist $I = \bigcup_{k\in K} I_k$ eine disjunkte Zerlegung von I, dann sind auch die σ-Algebren $\sigma(\bigcup_{i\in I_k} \mathcal{A}_i)$ $(k \in K)$ unabhängige Mengensysteme.

Beispiel. Sind $A_1, \ldots, A_n$ unabhängige Ereignisse, dann sind die Mengensysteme $\mathcal{A}_1 = \{A_1\}$, ... ,$\mathcal{A}_n = \{A_n\}$ unabhängig und durchschnittsstabil. Also sind auch die σ-Algebren

$$\sigma(\mathcal{A}_i) \quad = \quad \{\emptyset, A_i, A_i^C, \Omega\} \qquad (i = 1, \ldots, n)$$

unabhängige Mengensysteme, d.h es gilt

$$P[B_1 \cap \ldots \cap B_n] \quad = \quad \prod_{i=1}^{n} P[B_i] \qquad \text{für } B_i \in \{\emptyset, A_i, A_i^C, \Omega\}.$$

Dies kann man auch direkt beweisen, siehe Lemma **??** oben.

Ein Beispiel zum zweiten Teil der Aussage von Satz 2.1 werden wir im Anschluss an den Beweis des Satzes betrachten.

Beweis. (1). Seien $i_1, \ldots, i_n \in I$ paarweise verschieden. Wir müssen zeigen, dass

$$P[B_1 \cap \ldots \cap B_n] \quad = \quad P[B_1] \cdot \ldots \cdot P[B_n] \tag{2.1.2}$$

für alle $B_1 \in \sigma(\mathcal{A}_{i_1}), \ldots, B_n \in \sigma(\mathcal{A}_{i_n})$ gilt. Dazu verfahren wir schrittweise:

(a) Die Aussage (2.1.2) gilt nach Voraussetzung für $B_1 \in \mathcal{A}_{i_1}, \ldots, B_n \in \mathcal{A}_{i_n}$.

(b) Für $B_2 \in \mathcal{A}_{i_2}, \ldots, B_n \in \mathcal{A}_{i_n}$ betrachten wir das Mengensystem $\mathcal{D}$ aller $B_1 \in \mathcal{A}$, für die (2.1.2) gilt. $\mathcal{D}$ ist ein Dynkinsystem, das $\mathcal{A}_{i_1}$ nach (a) enthält. Da $\mathcal{A}_{i_1}$ durchschnittsstabil ist, folgt

$$\mathcal{D} \quad \supseteq \quad \mathcal{D}(\mathcal{A}_{i_1}) \quad = \quad \sigma(\mathcal{A}_{i_1}).$$

Also gilt (2.1.2) für alle $B_1 \in \sigma(\mathcal{A}_{i_1})$.

(c) Für $B_1 \in \sigma(\mathcal{A}_{i_1})$ und $B_3 \in \sigma(\mathcal{A}_{i_3}), \ldots, B_n \in \sigma(\mathcal{A}_{i_n})$ betrachten wir nun das Mengensystem aller $B_2 \in \mathcal{A}$, für die (2.1.2) gilt. Wiederum ist $\mathcal{D}$ ein Dynkinsystem, das $\mathcal{A}_{i_2}$ nach (b) enthält. Wie im letzten Schritt folgt daher

$$\mathcal{D} \quad \supseteq \quad \mathcal{D}(\mathcal{A}_{i_2}) \quad = \quad \sigma(\mathcal{A}_{i_2}),$$

d.h. (2.1.2) ist für alle $B_2 \in \sigma(\mathcal{A}_{i_2})$ erfüllt.

(d) Anschließend verfahren wir auf entsprechende Weise weiter. Nach n-facher Anwendung eines analogen Arguments folgt die Behauptung.

(2). Für $k \in K$ gilt: $\sigma(\bigcup_{i \in I_k} \mathcal{A}_i) = \sigma(\mathcal{C}_k)$ mit

$$\mathcal{C}_k \quad := \quad \{B_1 \cap \ldots \cap B_n \,:\, n \in \mathbb{N}, i_1, \ldots, i_n \in I_k, B_j \in \mathcal{A}_{i_j}\}.$$

Die Mengensysteme $\mathcal{C}_k$, $k \in K$, sind durchschnittsstabil und unabhängig, da jede Kollektion von Ereignissen $A_i \in \mathcal{A}_i$ ($i \in I$), nach Voraussetzung unabhängig ist. Also sind nach Teil (1) der Aussage auch die σ-Algebren $\sigma(\mathcal{C}_k)$, $k \in K$, unabhängig.

□

Beispiel (**Affe tippt Shakespeare**). Wir betrachten unabhängige 0-1-Experimente mit Erfolgswahrscheinlichkeit p. Sei $X_i(\omega) \in \{0, 1\}$ der Ausgang des i-ten Experiments. Für ein binäres Wort $(a_1, \ldots, a_n) \in \{0, 1\}^n$ mit Länge $n \in \mathbb{N}$ gilt:

$$P[X_1 = a_1, \ldots, X_n = a_n] \quad = \quad P\left[\bigcap_{i=1}^{n} \{X_i = a_i\}\right] \quad \overset{\text{unabh.}}{=} \quad p^k \cdot (1-p)^{n-k},$$

wobei $k = a_1 + \ldots + a_n$ die Anzahl der Einsen in dem Wort ist. Wir zeigen nun:

Behauptung: $P[\text{Wort kommt unendlich oft in der Folge } X_1, X_2, \ldots \text{ vor}] = 1$ falls $p \notin \{0, 1\}$.

Zum Beweis bemerken wir, dass die Ereignisse

$$E_m \quad = \quad \{X_{mn+1} = a_1, X_{mn+2} = a_2, \ldots, X_{mn+n} = a_n\}, \qquad m \in \mathbb{N},$$
„Wort steht im m-ten Block“

unabhängig sind. Nach Satz 2.1 sind nämlich die σ-Algebren

$$\sigma(\{\{X_{mn+1} = 1\}, \{X_{mn+2} = 1\}, \ldots, \{X_{mn+n} = 1\}\}), \qquad m \in \mathbb{N},$$

unabhängig, also auch die darin enthaltenen Ereignisse E_m. Für $p \neq 0$ gilt:

$$P[E_m] \;=\; p^k \cdot (1-p)^{n-k} \;>\; 0,$$

also $\sum_{m=1}^{\infty} P[E_m] = \infty$. Damit folgt nach Borel-Cantelli:

$$1 \quad = \quad P[E_m \text{ unendlich oft}] \quad \leq \quad P[\text{Wort kommt unendlich oft vor}].$$

□

2.1.2 Unabhängigkeit von Zufallsvariablen

Wir führen nun einen Unabhängigkeitsbegriff für Zufallsvariablen X_i mit Werten in allgemeinen meßbaren Räumen $(S_i, \mathcal{S}_i)$ ein, der kompatibel mit dem oben definierten Unabhängigkeitsbegriff für Mengensysteme ist.

Definition (Unabhängigkeit von Zufallsvariablen mit allgemeinem Zustandsraum).
Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, und seien $(S_i, \mathcal{S}_i)$ meßbare Räume.

*(1). Eine endliche Kollektion $X_1 : \Omega \to S_1, \ldots, X_n : \Omega \to S_n$ von Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ heißt **unabhängig**, falls*

$$P[X_1 \in B_1, \ldots, X_n \in B_n] = \prod_{i=1}^{n} P[X_i \in B_i] \qquad \forall\, B_i \in \mathcal{S}_i \quad (1 \leq i \leq n). \tag{2.1.3}$$

*(2). Eine beliebige Kollektion $X_i : \Omega \to S_i$ $(i \in I)$ von Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ heißt **unabhängig**, falls jede endliche Teilkollektion $X_{i_1}, \ldots, X_{i_n}$ (mit $i_1, \ldots, i_n \in I$ paarweise verschieden) unabhängig ist.*

Bemerkung. (1). Die Definition ist **konsistent**: Jede endliche Teilkollektion einer unabhängigen endlichen Kollektion von Zufallsvariablen ist wieder unabhängig im Sinne von (2.1.3).

(2). Die Zufallsvariablen $X_i, i \in I$, sind genau dann unabhängig, wenn die σ-Algebren

$$\sigma(X_i) \quad = \quad \{X_i^{-1}(B) : B \in \mathcal{S}_i\}, \qquad i \in I,$$

unabhängige Mengensysteme sind.

Seien $(\widetilde{S}_i, \widetilde{\mathcal{S}}_i)$ weitere meßbare Räume. Eine sehr wichtige Konsequenz von Bemerkung (2) ist:

Satz 2.2 (Funktionen von unabhängigen Zufallsvariablen sind unabhängig).
Sind $X_i : \Omega \to S_i$ $(i \in I)$ unabhängige Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$, und sind $h_i : S_i \to \widetilde{S}_i$ meßbare Abbildungen, dann sind auch die Zufallsvariablen $h_i(X_i)$, $i \in I$, unabhängig bzgl. P.

Beweis. Für $i \in I$ gilt

$$\sigma(h_i(X_i)) \quad = \quad \{(h_i \circ X_i)^{-1}(B) : B \in \widetilde{\mathcal{S}}_i\} \quad = \quad \{X_i^{-1}(h_i^{-1}(B)) : B \in \widetilde{\mathcal{S}}_i\} \quad \subseteq \quad \sigma(X_i).$$

Da die σ-Algebren $\sigma(X_i)$, $i \in I$, unabhängig sind, sind auch die von den Zufallsvariablen $h_i(X_i)$ erzeugten σ-Algebren unabhängige Mengensysteme. □

Allgemeiner können wir Funktionen von disjunkten Gruppen von unabhängigen Zufallsvariablen betrachten, und erhalten wieder unabhängige Zufallsvariablen:

Definition (Von mehreren Abbildungen erzeugte σ-Algebra). *Für $J \subseteq I$ ist die **von den Abbildungen** $X_i : \Omega \to S_i$ $(i \in J)$ **erzeugte** σ**-Algebra** definiert als*

$$\sigma\left(X_i : i \in J\right) \quad := \quad \sigma\left(\{X_i^{-1}(B) : i \in J, B \in \mathcal{S}_i\}\right) \quad = \quad \sigma\left(\bigcup_{i \in J} \sigma(X_i)\right).$$

Offensichtlich ist $\sigma(X_i : i \in J)$ die kleinste σ-Algebra auf Ω, bzgl. der alle Abbildungen X_i, $i \in J$, meßbar sind.

Sei nun $I = \dot{\bigcup}_k I_k$ eine disjunkte Unterteilung der Indexmenge. Dann sind die Abbildungen

$$X_{I_k} := (X_i)_{i \in I_k} : \Omega \to S^{I_k} \qquad (k \in K)$$

Zufallsvariablen bzgl. der Produkt σ-Algebren $\bigotimes_{i \in I_k} \mathcal{S}_i$ auf den Bildräumen $S^{I_k} = \times_{i \in I_k} S_i$. Diese Zufallsvariablen sind wieder unabhängig, denn aufgrund von Satz 2.1 sind die von ihnen erzeugten σ-Algebren

$$\sigma\left(X_i : i \in I_k\right) \quad = \quad \sigma\left(\bigcup_{i \in I_k} \sigma(X_i)\right), \qquad k \in K,$$

unabhängig. Somit sind nach Satz 2.2 auch beliebige (bzgl. der Produkt-σ-Algebra meßbare) Funktionen

$$Y_k \quad = \quad h_k\left(X_{I_k}\right), \qquad k \in K,$$

von den Zufallsvariablen der verschiedenen Gruppen wieder unabhängig.

2.1.3 Maxima von unabhängigen exponentialverteilten Zufallsvariablen

Seien $T_1, T_2, \ldots$ unabhängige $\mathrm{Exp}(1)$-verteilte Zufallsvariablen. Wir wollen uns überlegen, wie sich die Extremwerte (Rekorde)

$$M_n \quad = \quad \max\{T_1, \ldots, T_n\}$$

asymptotisch für $n \to \infty$ verhalten. Dazu gehen wir in mehreren Schritten vor:

(1). Wir zeigen zunächst mithilfe des Borel-Cantelli-Lemmas:

$$\limsup_{n\to\infty} \frac{T_n}{\log n} = 1 \qquad P\text{-fast sicher.} \tag{2.1.4}$$

Zum Beweis berechnen wir für $c \in \mathbb{R}$:

$$P\left[\frac{T_n}{\log n} \geq c\right] \;=\; P[T_n \geq c \cdot \log n] \;=\; e^{-c\log n} \;=\; n^{-c}.$$

Für $c > 1$ gilt $\sum_{n=1}^{\infty} n^{-c} < \infty$. Nach dem 1. Borel-Cantelli-Lemma folgt daher

$$P\left[\limsup_{n\to\infty} \frac{T_n}{\log n} > c\right] \quad \leq \quad P\left[\frac{T_n}{\log n} \geq c \quad \text{unendlich oft}\right] \quad = \quad 0.$$

Für $c \searrow 1$ erhalten wir dann wegen der monotonen Stetigkeit von P:

$$P\left[\limsup_{n\to\infty} \frac{T_n}{\log n} > 1\right] \quad = \quad \lim_{c\searrow 1} P\left[\limsup_{n\to\infty} \frac{T_n}{\log n} > c\right] \quad = \quad 0. \tag{2.1.5}$$

Für $c < 1$ gilt dagegen $\sum_{n=1}^{\infty} n^{-c} = \infty$. Da die Ereignisse $\{T_n \geq c \log n\}, n \in \mathbb{N}$, unabhängig sind, folgt nach dem 2. Borel-Cantelli Lemma:

$$P\left[\limsup_{n\to\infty} \frac{T_n}{\log n} \geq c\right] \quad \geq \quad P\left[\frac{T_n}{\log n} \geq c \quad \text{unendlich oft}\right] \quad = \quad 1.$$

Für $c \nearrow 1$ erhalten wir mithilfe der monotonen Stetigkeit:

$$P\left[\limsup_{n\to\infty} \frac{T_n}{\log n} \geq 1\right] \quad = \quad \lim_{c\nearrow 1} P\left[\limsup_{n\to\infty} \frac{T_n}{\log n} \geq c\right] \quad = \quad 1. \tag{2.1.6}$$

Aus (2.1.5) und (2.1.6) folgt die Behauptung (2.1.4).

(2). Als nächstes folgern wir:

$$M_n \sim \log n, \quad \text{d.h. } \lim_{n\to\infty} \frac{M_n}{\log n} = 1 \qquad P\text{-fast sicher.} \tag{2.1.7}$$

Zum Beweis zeigen wir:

(a) $\limsup_{n\to\infty} \frac{M_n}{\log n} \leq 1$ $\quad P$-f.s., und

(b) $\liminf_{n\to\infty} \frac{M_n}{\log n} \geq 1$ $\quad P$-f.s.

Aussage (a) folgt aus (1), denn für $c \in \mathbb{R}$ gilt:

$$\begin{aligned}
\limsup_{n\to\infty} \frac{M_n}{\log n} \;>\; c \;&\Rightarrow\; \max\{T_1, \ldots, T_n\} \;=\; M_n \;>\; c \cdot \log n \text{ unendlich oft} \\
&\Rightarrow\; T_{k(n)} \;>\; c \cdot \log n \text{ für } k(n) \leq n \text{ für unendlich viele } n \\
&\Rightarrow\; T_k \;>\; c \cdot \log k \quad \text{unendlich oft} \\
&\Rightarrow\; \limsup \frac{T_k}{\log k} \;\geq\; c
\end{aligned}$$

Nach (1) hat das letztere Ereignis für $c > 1$ Wahrscheinlichkeit 0, also gilt wegen der monotonen Stetigkeit von P:

$$P\left[\limsup_{n\to\infty} \frac{M_n}{\log n} > 1\right] \quad = \quad \lim_{c\searrow 1} P\left[\limsup_{n\to\infty} \frac{M_n}{\log n} > c\right] \quad = \quad 0.$$

Zum Beweis von (b) genügt es wegen der monotonen Stetigkeit zu zeigen, dass für $c < 1$

$$P\left[\frac{M_n}{\log n} > c \quad \text{schließlich}\right] \quad = \quad P\left[\frac{M_n}{\log n} \leq c \quad \text{nur endlich oft}\right] \quad = \quad 1$$

gilt. Nach Borel-Cantelli I ist dies der Fall, wenn

$$\sum_{n\in\mathbb{N}} P\left[\frac{M_n}{\log n} \leq c\right] \quad < \quad \infty \tag{2.1.8}$$

gilt. Für $c \in \mathbb{R}$ erhalten wir aber wegen der Unabhängigkeit der T_i:

$$\begin{aligned} P\left[\frac{M_n}{\log n} \leq c\right] \quad &= \quad P\left[T_i \leq c \cdot \log n \quad \forall\, 1 \leq i \leq n\right] \;=\; P\left[T_1 \leq c \cdot \log n\right]^n \\ &= \quad (1 - n^{-c})^n \;\leq\; e^{-n\cdot n^{-c}} \;=\; e^{-n^{1-c}}, \end{aligned}$$

und diese Folge ist für $c < 1$ summierbar. Also gilt (2.1.8) für alle $c < 1$, und damit (b).

(3). Abschließend untersuchen wir die Fluktuationen der Extremwerte M_n um $\log n$ noch genauer. Wir zeigen, dass die Zufallsvariable $M_n - \log n$ in Verteilung konvergiert:

$$P[M_n - \log n \leq c] \quad \overset{n\to\infty}{\longrightarrow} \quad e^{-e^{-c}} \qquad \text{für alle } c \in \mathbb{R}. \tag{2.1.9}$$

Beweis. Wegen

$$P[M_n \leq c] \quad = \quad P[T_i \leq c \quad \forall i = 1, \ldots, n] \overset{\text{i.i.d.}}{=} (1 - e^{-c})^n \qquad \text{für alle } c \in \mathbb{R}$$

folgt

$$P[M_n - \log n \leq c] \quad = \quad P[M_n \leq c + \log n] \quad = \quad (1 - \frac{1}{n} \cdot e^{-c})^n \quad \overset{n\to\infty}{\longrightarrow} e^{-e^{-c}}.$$

□

Aussage (2.1.9) besagt, dass $M_n - \log n$ in Verteilung gegen eine **Gumbel-verteilte** Zufallsvariable X, d.h. eine Zufallsvariable mit Verteilungsfunktion $F_X(c) = e^{-e^{-c}}$ konvergiert. Für große n gilt also näherungsweise

$$M_n \quad \overset{\mathcal{D}}{\approx} \quad \log n + X, \qquad X \sim \text{Gumbel},$$

wobei $\log n$ die Asymptotik und X die Fluktuationen beschreibt.

2.2 Endliche Produktmaße

Um Aussagen über den Zusammenhang mehrerer Zufallsvariablen $X_1, \ldots, X_n$ zu treffen, benötigen wir Kenntnisse über deren gemeinsame Verteilung, d.h. über die Verteilung des Zufallsvektors $X = (X_1, \ldots, X_n)$. Diese ist eine Wahrscheinlichkeitsverteilung auf dem Produkt der Wertebereiche der einzelnen Zufallsvariablen.

2.2.1 Produktmaße und Unabhängigkeit

Seien $(S_i, \mathcal{S}_i)$, $1 \leq i \leq n$, messbare Räume. Die Produkt-σ-Algebra $\mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n$ auf $S_1 \times \ldots \times S_n$ wird von den endlichen Produkten von Mengen aus den σ-Algebren $\mathcal{S}_i$ erzeugt:

$$\mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n \quad = \quad \sigma\left(\{B_1 \times \ldots \times B_n : B_i \in \mathcal{S}_i \quad \forall\, 1 \leq i \leq n\}\right).$$

Bezeichnen wir mit

$$\pi_i : S_1 \times \ldots \times S_n \to S_i, \qquad \pi_i(x_1, \ldots, x_n) := x_i,$$

die kanonische Projektion auf die i-te Komponente, so gilt

$$\mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n \quad = \quad \sigma(\pi_1, \ldots, \pi_n).$$

Beispiel (Borelsche σ-Algebra auf $\mathbb{R}^n$). Die Borelsche σ-Algebra auf $\mathbb{R}^n$ wird zum Beispiel von den offenen Quadern, also den Produkten von offenen Intervallen, erzeugt. Daher gilt

$$\mathcal{B}(\mathbb{R}^n) \quad = \quad \underbrace{\mathcal{B}(\mathbb{R}) \otimes \ldots \otimes \mathcal{B}(\mathbb{R})}_{n \text{ mal}} \quad = \quad \bigotimes_{i=1}^{n} \mathcal{B}(\mathbb{R}).$$

Ein anderes Erzeugendensystem von $\mathcal{B}(\mathbb{R}^n)$ bilden die Produktmengen

$$(-\infty, c_1] \times (-\infty, c_2] \times \ldots \times (-\infty, c_n], \qquad c_1, \ldots, c_n \in \mathbb{R}. \tag{2.2.1}$$

Ist μ eine beliebige Wahrscheinlichkeitsverteilung auf $(S_1 \times \ldots \times S_n, \mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n)$, dann heißen die Wahrscheinlichkeitsverteilungen

$$\mu_{\pi_i} \quad := \quad \mu \circ \pi_i^{-1}, \qquad 1 \leq i \leq n,$$

auf den Komponenten $(S_i, \mathcal{S}_i)$ **(eindimensionale) Randverteilungen (marginals)** von μ. Wir werden in Abschnitt 3.3 allgemeine Wahrscheinlichkeitsverteilungen auf Produkträumen konstruieren. Zunächst betrachten wir hier eine spezielle Klasse solcher Verteilungen: die (endlichen) Produktmaße.

Definition (Endliches Produktmaß). *Seien $(S_i, \mathcal{S}_i, \mu_i)$ Wahrscheinlichkeitsräume, $1 \leq i \leq n$. Eine Wahrscheinlichkeitsverteilung μ auf $(S_1 \times \ldots \times S_n, \mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n)$ heißt Produkt der μ_i, falls*

$$\mu[B_1 \times \ldots \times B_n] \quad = \quad \prod_{i=1}^{n} \mu_i[B_i] \qquad \textit{für alle } B_i \in \mathcal{S}_i \ (1 \leq i \leq n) \tag{2.2.2}$$

gilt.

Bemerkung (Existenz und Eindeutigkeit von Produktmaßen). Das Produktmaß μ ist durch (2.2.2) *eindeutig* festgelegt, denn die Produktmengen bilden einen durchschnittsstabilen Erzeuger der σ-Algebra $\mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n$. Die *Existenz* von Produktmaßen folgt allgemein aus dem Satz von Fubini, den wir in Abschnitt 3.3 beweisen. Für Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$ zeigen wir die Existenz von Produktmaßen im nächsten Abschnitt.

Das nach der Bemerkung eindeutige Produkt der Wahrscheinlichkeitsverteilungen $\mu_1, \ldots, \mu_n$ bezeichnen wir mit $\mu_1 \otimes \ldots \otimes \mu_n$.

Die eindimensionalen Randverteilungen eines Produktmaßes sind gerade die Faktoren μ_i:

Lemma 2.3. *Unter dem Produktmaß $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ sind die Projektionen*

$$\pi_i : S_1 \times \ldots \times S_n \longrightarrow S_i, \quad \pi_i(x_1, \ldots, x_n) = x_i, \qquad 1 \leq i \leq n,$$

unabhängige Zufallsvariablen mit Verteilung μ_i.

Beweis. Für $B_i \in \mathcal{S}_i$ $(1 \leq i \leq n)$ gilt

$$\mu[\pi_1 \in B_1, \ldots, \pi_n \in B_n] \quad = \quad \mu[B_1 \times \ldots \times B_n] \quad = \quad \prod_{i=1}^{n} \mu_i[B_i],$$

also insbesondere $\mu[\pi_i \in B_i] = \mu_i[B_i]$ für $i = 1, \ldots n$. Hieraus folgt die Unabhängigkeit. □

Sind $X_i : \Omega \to S_i$ $(1 \leq i \leq n)$ Zufallsvariablen mit Werten in messbaren Räumen $(S_i, \mathcal{S}_i)$, welche auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definiert sind, dann ist

$$(X_1, \ldots, X_n) : \quad \Omega \quad \longrightarrow \quad S_1 \times \ldots \times S_n$$

eine Zufallsvariable mit Werten im Produktraum $(S_1 \times \ldots \times S_n, \mathcal{S}_1 \otimes \ldots \otimes \mathcal{S}_n)$, denn für beliebige Mengen $B_i \in \mathcal{S}_i$ $(1 \leq i \leq n)$ gilt:

$$\{(X_1, \ldots, X_n) \in B_1 \times \ldots \times B_n\} \quad = \quad \bigcap_{i=1}^{n} \{X_i \in B_i\} \quad \in \quad \mathcal{A}.$$

Wie zuvor im diskreten Fall (s. Abschnitt **??**) definieren wir:

Definition (Gemeinsame Verteilung). *Die Verteilung $\mu_{X_1,\ldots,X_n}$ des Zufallsvektors $(X_1,\ldots,X_n)$ auf $(S_1\times\ldots\times S_n, \mathcal{S}_1\otimes\ldots\otimes\mathcal{S}_n)$ heißt **gemeinsame Verteilung** der Zufallsvariablen $X_1,\ldots,X_n$.*

Der folgende Satz gilt analog zum diskreten Fall:

Satz 2.4 (Unabhängigkeit und Produktmaße). *Die folgenden Aussagen sind äquivalent:*

(1). Die Zufallsvariablen $X_1,\ldots,X_n$ sind unabhängig.

(2). Die gemeinsame Verteilung $\mu_{X_1,\ldots,X_n}$ ist ein Produktmaß.

(3). $\mu_{X_1,\ldots,X_n} = \mu_{X_1}\otimes\ldots\otimes\mu_{X_n}$.

Beweis. „ (1) $\Longrightarrow$ (3) “ folgt direkt aus der Definition der Unabhängigkeit: Sind $X_1,\ldots,X_n$ unabhängige Zufallsvariablen, dann gilt für $B_i\in\mathcal{S}_i$ $(1\le i\le n)$:

$$\begin{aligned}\mu_{X_1,\ldots,X_n}[B_1\times\ldots\times B_n] &= P[(X_1,\ldots,X_n)\in B_1\times\ldots\times B_n]\\ &= P[X_i\in B_i \quad \forall\, 1\le i\le n]\\ &= \prod_{i=1}^n P[X_i\in B_i] \;=\; \prod_{i=1}^n \mu_{X_i}[B_i].\end{aligned}$$

Die Implikation „ (3) $\Longrightarrow$ (2) “: ist offensichtlich, und „ (2) $\Longrightarrow$ (1) “ folgt aus Lemma 2.3: Ist $\mu_{X_1,\ldots,X_n}$ ein Produktmaß, dann sind die kanonischen Projektionen $\pi_1,\ldots,\pi_n$ unabhängig unter $\mu_{X_1,\ldots,X_n}$. Also gilt für $B_i\in\mathcal{S}_i$:

$$\begin{aligned}P[X_1\in B_1,\ldots,X_n\in B_n] &= \mu_{X_1,\ldots,X_n}[B_1\times\ldots\times B_n]\\ &= \mu_{X_1,\ldots,X_n}[\pi_1\in B_1,\ldots,\pi_n\in B_n]\\ &= \prod_{i=1}^n \mu_{X_1,\ldots,X_n}[\pi_i\in B_i] \;=\; \prod_{i=1}^n P[X_i\in B_i].\end{aligned}$$

□

2.2.2 Produktmaße auf $\mathbb{R}^n$

Im Fall $S_1=\ldots=S_n=\mathbb{R}$ mit Borelscher σ-Algebra bilden die Mengen aus (2.2.1) einen durchschnittsstabilen Erzeuger der Produkt-σ-Algebra $\mathcal{B}(\mathbb{R}^n)$. Also ist $\mu=\mu_1\otimes\ldots\otimes\mu_n$ genau dann, wenn

$$\mu[(-\infty,c_1]\times\ldots\times(-\infty,c_n]] \;=\; \prod_{i=1}^n \mu_i[(-\infty,c_i]] \qquad \text{für alle } c_1,\ldots,c_n\in\mathbb{R}$$

gilt. Die gemeinsame Verteilung reellwertiger Zufallsvariablen $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ auf der Produkt-σ-Algebra $\mathcal{B}(\mathbb{R}^n) = \bigotimes_{i=1}^n \mathcal{B}(\mathbb{R})$ ist somit vollständig durch die Werte

$$\begin{aligned} F_{X_1,\ldots,X_n}(c_1,\ldots,c_n) &:= \mu_{X_1,\ldots,X_n}[(-\infty,c_1]\times\ldots\times(-\infty,c_n]] \\ &= P[X_1 \le c_1,\ldots,X_n \le c_n] \end{aligned} \tag{2.2.3}$$

mit $(c_1, \ldots, c_n) \in \mathbb{R}^n$ beschrieben. Insbesondere sind $X_1, \ldots, X_n$ genau dann unabhängig, wenn

$$F_{X_1,\ldots,X_n}(c_1,\ldots,c_n) = \prod_{i=1}^n F_{X_i}(c_i) \qquad \text{für alle } (c_1,\ldots,c_n) \in \mathbb{R}^n \text{ gilt.}$$

Definition (Gemeinsame Verteilungsfunktion). *Die Funktion $F_{X_1,\ldots,X_n} : \mathbb{R}^n \to [0,1]$ heißt **gemeinsame Verteilungsfunktion** der Zufallsvariablen $X_1, \ldots, X_n$ bzw. der multivariaten Wahrscheinlichkeitsverteilung $\mu_{X_1,\ldots,X_n}$.*

Wir betrachten nun den wichtigen Spezialfall, dass die einzelnen Verteilungen absolutstetig sind. Der Satz von Fubini, den wir in Abschnitt 3.3 in größerer Allgemeinheit beweisen werden, besagt unter anderem, dass das n-dimensionale Lebesgue-Integral einer beliebigen Borel-messbaren nicht-negativen Funktion $f : \mathbb{R}^n \to \mathbb{R}$ existiert, und als Hintereinanderausführung von eindimensionalen Integralen nach den Koordinaten $x_1, \ldots, x_n$ berechnet werden kann:

$$\int_{\mathbb{R}^n} f(x)\,dx = \int \cdots \int f(x_1,\ldots,x_n)\,dx_n \cdots dx_1.$$

Hierbei können die eindimensionalen Integrationen in beliebiger Reihenfolge ausgeführt werden. Für den Beweis verweisen wir auf die Analysisvorlesung bzw. auf Abschnitt 3.3 unten.

In Analogie zum eindimensionalen Fall nennen wir ein Wahrscheinlichkeitsmaß μ auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ **stetig** oder **absolutstetig**, falls eine $\mathcal{B}(\mathbb{R}^n)$-messbare *Dichtefunktion* $f : \mathbb{R}^n \to [0,\infty)$ existiert mit

$$\mu[B] = \int_B f(x)\,dx := \int I_B(x) f(x)\,dx$$

für jeden Quader, bzw. allgemeiner für jede Borel-Menge $B \subseteq \mathbb{R}^n$. Wie im eindimensionalen Fall ist die Dichtefunktion bis auf Modifikation auf einer Lebesgue-Nullmenge eindeutig bestimmt.

Beispiel (Gleichverteilung auf einer Teilmenge des $\mathbb{R}^n$). Die Gleichverteilung auf einer meßbaren Teilmenge S des $\mathbb{R}^n$ mit endlichem positivem Lebesgue-Maß $\lambda[S]$ ist definiert durch

$$\mathcal{U}_S[B] := \frac{\lambda[B\cap S]}{\lambda[S]} = \frac{1}{\lambda[S]}\int_B I_S(x)\,dx \qquad \text{für } B \in \mathcal{B}(\mathbb{R}^n).$$

$\mathcal{U}_S$ ist absolutstetig mit Dichte $f(x) = I_S(x)/\lambda[S]$.

Endliche Produkte von eindimensionalen absolutstetigen Wahrscheinlichkeitsverteilungen sind wieder absolutstetig, und die Dichte ist das Produkt der einzelnen Dichten:

Lemma 2.5 (Dichten von Produktmaßen). *Sind $\mu_1, \ldots, \mu_n$ absolutstetige Wahrscheinlichkeitsmaße auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Dichtefunktionen $f_1, \ldots, f_n$, dann ist das Produkt $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ eine absolutstetige Wahrscheinlichkeitsverteilung auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ mit Dichtefunktion*

$$f(x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^{n} f_i(x_i).$$

Hieraus folgt die Behauptung, da die Produktmengen einen durchschnittsstabilen Erzeuger der Borel $sigma$-Algebra auf $\mathbb{R}^n$ bilden.

Beweis. Für jede Produktmenge $B = B_1 \times \ldots \times B_n$ mit Mengen $B_i \in \mathcal{B}(\mathbb{R})$ gilt nach dem Satz von Fubini:

$$\mu[B] \;=\; \prod_{i=1}^{n} \mu_i[B_i] \;=\; \prod_{i=1}^{n} \int\limits_{B_i} f_i(x_i)\, dx_i \;=\; \int \cdots \int I_B(x_1, \ldots, x_n) \prod_{i=1}^{n} f_i(x_i)\, dx_1 \cdots dx_n.$$

□

Beispiel (Gleichverteilung auf n-dimensionalem Quader). Ist $\mu_i = \mathcal{U}_{(a_i, b_i)}$ die Gleichverteilung auf einem endlichen Intervall (a_i, b_i), $-\infty < a_i < b_i < \infty$, dann ist $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ die Gleichverteilung auf dem Quader $S = (a_1, b_1) \times \ldots \times (a_n, b_n)$, denn für die Dichtefunktion gilt:

$$f(x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^{n} \frac{I_{(a_i, b_i)}(x_i)}{b_i - a_i} \;=\; \frac{I_S(x)}{\lambda[S]}.$$

Ein anderes Produktmaß von fundamentaler Bedeutung für die Wahrscheinlichkeitstheorie ist die mehrdimensionale Standardnormalverteilung:

Beispiel (Standardnormalverteilung im $\mathbb{R}^n$). Das Produkt $\mu = \bigotimes_{i=1}^{n} N(0,1)$ von n eindimensionalen Standardnormalverteilungen ist eine absolutstetige Wahrscheinlichkeitsverteilung auf $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ mit Dichte

$$f(x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2} \right) \right) \;=\; (2\pi)^{-n/2} \exp\left(-\frac{\|x\|^2}{2} \right), \qquad x \in \mathbb{R}^n.$$

Das Maß μ heißt **n-dimensionale Standardnormalverteilung**.

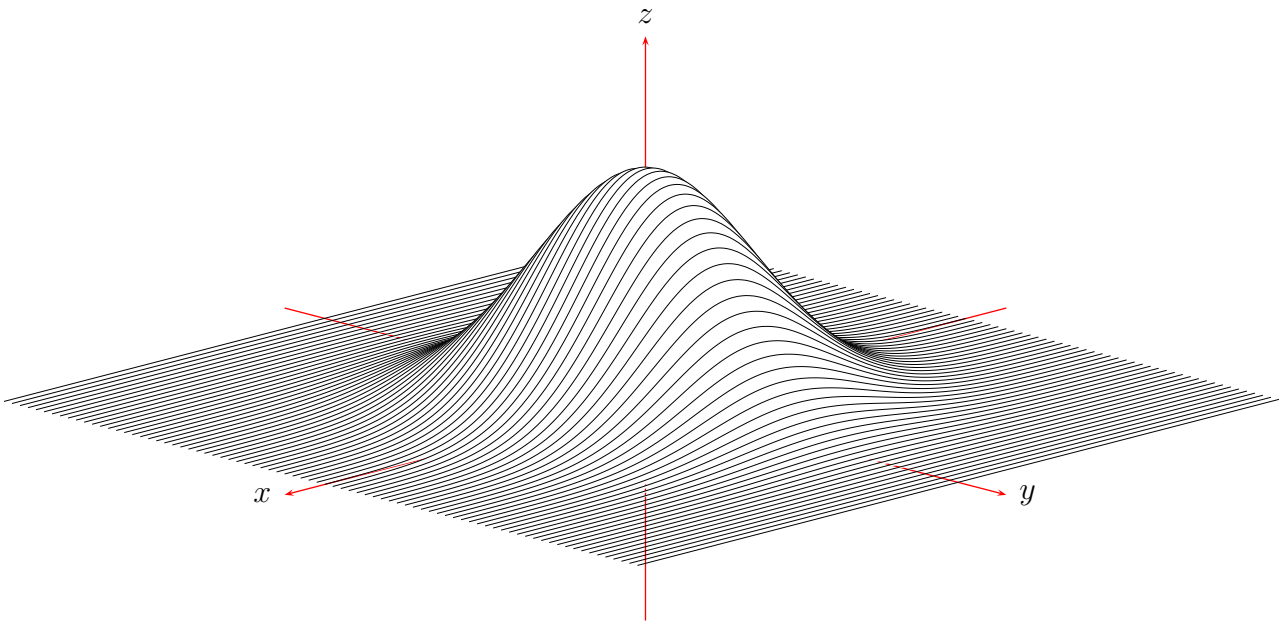


Abbildung 2.1: Dichte der Standardnormalverteilung in $\mathbb{R}^2$.

In Analogie zu der entsprechenden Aussage für diskrete Zufallsvariablen erhalten wir:

Korollar 2.6. *Seien $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ reellwertige Zufallsvariablen.*

(1). Sind $X_1, \ldots, X_n$ unabhängig mit absolutstetigen Verteilungen mit Dichten $f_{X_1}, \ldots, f_{X_n}$, dann ist die gemeinsame Verteilung absolutstetig mit Dichte

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

(2). Umgekehrt gilt: Ist die gemeinsame Verteilung absolutstetig, und hat die Dichte eine Darstellung

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \quad = \quad c \cdot \prod_{i=1}^{n} g_i(x_i)$$

in Produktform mit einer Konstante $c \in \mathbb{R}$ und integrierbaren Funktionen $g_i : \mathbb{R} \to [0, \infty)$, dann sind $X_1, \ldots, X_n$ unabhängig, und die Verteilungen sind absolutstetig mit Dichten

$$f_{X_i}(x_i) \quad = \quad \frac{g_i(x_i)}{\int\limits_{\mathbb{R}} g_i(t)\, dt}.$$

Der Beweis wird dem Leser zur Übung überlassen.

2.2.3 Konfidenzintervalle für Quantile

Sei $(x_1, \ldots, x_n)$ eine n-elementige Stichprobe von einer unbekannten Wahrscheinlichkeitsverteilung μ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Wir nehmen an, dass $x_1, \ldots, x_n$ Realisierungen von unabhängigen Zufallsvariablen mit Verteilung μ sind:

Annahme: $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ unabhängig unter P_μ mit Verteilung μ.

Wir wollen nun die Quantile (z.B. den Median) der zugrundeliegenden Verteilung auf der Basis der Stichprobe schätzen. Eine Funktion $T(X_1, \ldots, X_n)$, $T : \mathbb{R}^n \to \mathbb{R}$ messbar, nennt man in diesem Zusammenhang auch eine **Statistik** der Stichprobe $(X_1, \ldots, X_n)$. Eine Statistik, deren Wert als Schätzwert für eine Kenngröße $q(\mu)$ der unbekannten Verteilung verwendet wird, nennt man auch einen **(Punkt-) Schätzer** für q. Nahe liegende Schätzer für Quantile von μ sind die entsprechenden Stichprobenquantile. Unser Ziel ist es nun, *Konfidenzintervalle* für die Quantile anzugeben, d.h. von den Werten $X_1, \ldots, X_n$ abhängende Intervalle, in denen die Quantile *unabhängig von der tatsächlichen Verteilung* mit hoher Wahrscheinlichkeit enthalten sind. Seien dazu

$$X_{(1)} \quad \leq \quad X_{(2)} \quad \leq \quad \ldots \quad \leq \quad X_{(n)}$$

die der Größe nach geordneten Werte $X_1, \ldots, X_n$ – diese nennt man auch **Ordnungsstatistiken** der Stichprobe. Die Verteilung der Ordnungsstatistiken können wir explizit berechnen:

Satz 2.7 (Verteilung der Ordnungsstatistiken).

(1). Ist F die Verteilungsfunktion von μ, dann hat $X_{(k)}$ die Verteilungsfunktion

$$\begin{aligned} F_{(k)}(c) &= \mathrm{Bin}(n, F(c))[\{k, k+1, \ldots, n\}] \\ &= \sum_{j=k}^{n} \binom{n}{j} F(c)^j \cdot (1 - F(c))^{n-j}. \end{aligned} \tag{2.2.4}$$

(2). Ist F stetig und q ein u-Quantil von μ, dann folgt

$$P_\mu\left[X_{(k)} \leq q\right] = \mathrm{Bin}(n, u)[\{k, k+1, \ldots, n\}].$$

Beweis. Die Ereignisse $\{X_i \leq c\}, 1 \leq i \leq n$, sind unabhängig mit Wahrscheinlichkeit $F(c)$. Also gilt

$$\begin{aligned} F_{(k)}(c) = P_\mu\left[X_{(k)} \leq c\right] &= P_\mu\left[X_i \leq c \text{ für mindestens } k \text{ verschiedene } i \in \{1, \ldots, n\}\right] \\ &= \mathrm{Bin}(n, F(c))[\{k, k+1, \ldots, n\}]. \end{aligned}$$

Ist F stetig und q ein u-Quantil von F, dann gilt $F(q) = u$ nach Lemma 1.21. □

Nach Satz 2.7 ist die Wahrscheinlichkeit, dass der Wert von $X_{(k)}$ unterhalb eines u-Quantils der zugrundeliegenden Verteilung μ liegt, *für alle stetigen Verteilungen gleich!* Damit folgt unmittelbar:

Korollar 2.8 (Ordnungsintervalle). *Sei $u \in (0,1)$ und $1 \leq k < l \leq n$. Dann ist das zufällige Intervall $(X_{(k)}, X_{(l)})$ ein Konfidenzintervall für das u-Quantil der zugrundeliegenden Verteilung μ zum Konfidenzniveau*

$$\beta \quad := \quad \mathrm{Bin}(n,u)[\{k, k+1, \ldots, l-1\}],$$

d.h. für jede Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}$ mit stetiger Verteilungsfunktion, und für jedes u-Quantil q von μ gilt:

$$P_\mu\left[q \in (X_{(k)}, X_{(l)})\right] \quad \geq \quad \beta.$$

Beweis. Nach Satz 2.7 gilt

$$\begin{aligned} P_\mu[X_{(k)} < q < X(l)] &= \mathrm{Bin}(n,u)[\{k, k+1, \ldots, n\}] - \mathrm{Bin}(n,u)[\{l, l+1, \ldots, n\}] \\ &= \mathrm{Bin}(n,u)[\{k, k+1, \ldots, l-1\}]. \end{aligned}$$

□

Für große n kann man die Quantile der Binomialverteilung näherungsweise mithilfe der Normalapproximation berechnen, und erhält daraus entsprechende Konfidenzintervalle für die Quantile von Verteilungen auf $\mathbb{R}$. Bemerkenswert ist, dass diese Konfidenzintervalle nicht nur für Verteilungen aus einer bestimmten parametrischen Familie (z.B. der Familie der Normalverteilungen) gelten, sondern für *alle* Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$ mit stetiger Verteilungsfunktion (*nichtparametrisches statistisches Modell*).

Beispiel (Konfidenzintervalle für den Median). Die Binomialverteilung $\mathrm{Bin}(n, 1/2)$ hat den Mittelwert $m = n/2$ und die Standardabweichung $\sigma = \sqrt{n}/2$. Nach dem Satz von de Moivre/Laplace ist für große n ca. 95 % der Masse in der Menge $\{\lfloor m - 2\sigma \rfloor, \ldots, \lceil m + 2\sigma \rceil\}$ enthalten. Also ist das Intervall $\left(X_{(\lfloor n/2 - \sqrt{n} \rfloor)}, X_{(\lceil n/2 + \sqrt{n} \rceil)}\right)$ ein approximatives 95 % Konfidenzintervall für den Median einer beliebigen Verteilung mit stetiger Verteilungsfunktion. Beispielsweise können wir bei Zufallsstichproben der Größe 100 mit hoher Konfidenz erwarten, dass der Median zwischen dem 40. und 61. Wert liegt.

Ist die zugrundeliegende Verteilung μ absolutstetig mit Dichte f, dann kann man die Dichte der gemeinsamen Verteilung der Ordnungsstatistiken mit einem Symmetrieargument berechnen. Wegen $P[X_i = X_j] = 0$ für $i \neq j$ gilt in diesem Fall

$$\begin{aligned} P[X_{(1)} \leq c_1, \ldots, X_{(n)} \leq c_n] &= n!\, P[X_1 \leq c_1, \ldots, X_n \leq c_n, X_1 < X_2 < \ldots < X_n] \\ &= n! \int_{-\infty}^{c_1} \cdots \int_{-\infty}^{c_n} I_{\{y_1<y_2<\ldots<y_n\}} f(y_1) \cdots f(y_n)\, dy_1 \cdots dy_n. \end{aligned}$$

Hieraus folgt, dass die gemeinsame Verteilung von $X_{(1)}, \ldots, X_{(n)}$ absolutstetig ist mit Dichte

$$f_{X_{(1)},\ldots,X_{(n)}}(y_1, \ldots, y_n) = n! \cdot I_{\{y_1<y_2<\ldots<y_n\}} f(y_1) \cdots f(y_n).$$

Durch Aufintegrieren erhält man daraus mithilfe des Satzes von Fubini und einer erneuten Symmetrieüberlegung die Dichten der Verteilungen der einzelnen Ordnungsstatistiken:

$$f_{X_{(k)}}(y) = \frac{n!}{(k-1)!(n-k)!} F(y)^{k-1} (1 - F(y))^{n-k} f(y).$$

Dasselbe Resultat hätte man auch mit etwas Rechnen aus Satz 2.7 herleiten können.

Beispiel (Beta-Verteilungen). Sind die Zufallsvariablen X_i auf $(0, 1)$ gleichverteilt, dann hat $X_{(k)}$ die Dichte

$$f_{X_{(k)}}(u) = B(k, n-k+1)^{-1} \cdot u^{k-1} \cdot (1-u)^{n-k} \cdot I_{(0,1)}(u)$$

mit Normierungskonstante

$$B(a, b) = \int_0^1 u^{a-1}(1-u)^{b-1}\, du \quad \left(= \frac{(a-1)!(b-1)!}{(a+b-1)!} \qquad \text{für } a, b \in \mathbb{N} \right).$$

Die entsprechende Verteilung heißt **Beta-Verteilung mit Parametern** $a, b > 0$, die Funktion B ist die *Euler'sche Beta-Funktion*.

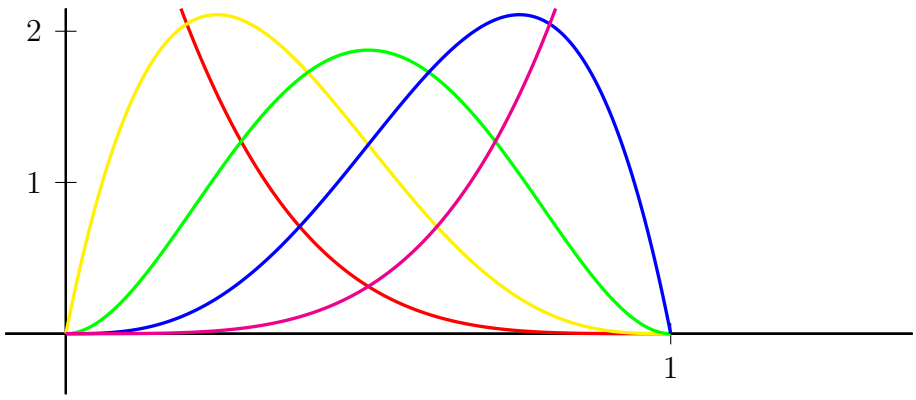


Abbildung 2.2: Abbildung der Dichtefunktionen der Ordnungsstatistiken $X_{(1)}, \ldots, X_{(5)}$ (rot, gelb, grün, blau, magenta) bzgl. der Gleichverteilung auf $(0, 1)$.

2.3 Unendliche Produktmodelle

Seien $\mu_1, \mu_2, \ldots$ vorgegebene Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Wir werden nun explizit unabhängige Zufallsvariablen $X_k, k \in \mathbb{N}$, mit Verteilungen μ_k konstruieren. Als Konsequenz ergibt sich die Existenz des unendlichen Produktmaßes $\bigotimes_{k=1}^{\infty} \mu_k$ als gemeinsame Verteilung der Zufallsvariablen X_k.

2.3.1 Konstruktion von unendlich vielen unabhängigen Zufallsvariablen

Die Zufallsvariablen X_i können wir sogar auf den Raum $\Omega = (0,1)$ mit Gleichverteilung realisieren:

Satz 2.9. *Auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{B}((0,1)), \mathcal{U}_{(0,1)})$ existieren unabhängige Zufallsvariablen $X_k : \Omega \to \mathbb{R}, k \in \mathbb{N}$, mit Verteilungen*

$$P \circ X_k^{-1} \quad = \quad \mu_k \qquad \textit{für alle } 1 \leq i \leq n.$$

Beweis. Wir verfahren in drei Schritten:

(1). Wir konstruieren die Zufallsvariablen im Fall

$$\mu_k \quad = \quad \text{Bernoulli}\left(\frac{1}{2}\right) \quad = \qquad \forall\, k \in \mathbb{N},$$

d.h. im fairen Münzwurfmodell. Dazu verwenden wir die schon in Abschnitt 1.2 eingeführte Transformation $X : (0,1) \to \{0,1\}^{\mathbb{N}}$, die einer reellen Zahl die Ziffernfolge ihrer Binärdarstellung zuordnet, d.h. wir setzen

$$X_k(\omega) \quad = \quad I_{D_k}(\omega), \qquad D_k \quad = \quad \bigcup_{i=1}^{2^{k-1}} [(2i-1)\cdot 2^{-k}, 2i \cdot 2^{-k}),$$

siehe Abbildung 1.4. Die Abbildungen $X_k : (0,1) \to \{0,1\}$ sind messbar, und es gilt

$$P[X_1 = a_1, \ldots, X_n = a_n] \quad = \quad 2^{-n} \qquad \forall n \in \mathbb{N}, a_1, \ldots, a_n \in \{0,1\}, \tag{2.3.1}$$

da die Menge $\{\omega \in \Omega : X_1(\omega) = a_1, \ldots, X_n(\omega) = a_n\}$ gerade aus den Zahlen in $(0,1)$ besteht, deren Binärdarstellung mit den Ziffern $a_1, \ldots, a_n$ beginnt, und damit ein Intervall der Länge 2^{-n} ist. Nach (2.3.1) sind $X_1, \ldots, X_n$ für alle $X_k, k \in \mathbb{N}$, unabhängig mit Verteilung μ_k.

(2). Wir konstruieren die Zufallsvariablen im Fall

$$\mu_k \quad = \quad \mathcal{U}_{(0,1)} \qquad \forall\, k \in \mathbb{N}.$$

Dazu zerlegen wir die gerade konstruierte Folge $X_k(\omega) \in \{0,1\}, k \in \mathbb{N}$, in unendlich viele Teilfolgen, und konstruieren aus jeder Teilfolge wieder eine Zahl aus $[0,1]$ mit den entsprechenden Binärziffern. Genauer setzen wir in Binärdarstellung:

$$\begin{aligned} U_1 &:= 0.X_1X_3X_5X_7\cdots, \\ U_2 &:= 0.X_2X_6X_{10}X_{14}\cdots, \\ U_3 &:= 0.X_4X_{12}X_{20}X_{28}\cdots, \qquad \text{usw.}, \end{aligned}$$

also allgemein für $k \in \mathbb{N}$:

$$U_k(\omega) \quad := \quad \sum_{i=1}^{\infty} X_{k,i}(\omega)\cdot 2^{-i} \qquad \text{mit} \quad X_{k,i} := X_{(2i-1)\cdot 2^{k-1}}.$$

Da die Zufallsvariablen $X_{k,i}, i,k \in \mathbb{N}$, unabhängig sind, sind nach dem Zerlegungssatz auch die σ-Algebren

$$\mathcal{A}_k \quad = \quad \sigma(X_{k,i}|i \in \mathbb{N}), \qquad k \in \mathbb{N},$$

unabhängig, und damit auch die $\mathcal{A}_k$-messbaren Zufallsvariablen U_k, $k \in \mathbb{N}$. Zudem gilt für $n \in \mathbb{N}$ und

$$r = \sum_{i=1}^{n} a_i \cdot 2^{i-1} \quad \in \quad \{0,1,\ldots,2^n-1\}:$$

$$P[U_k \in (r\cdot 2^{-n}, (r+1)\cdot 2^{-n})] \quad = \quad P[X_{k,1}=a_1,\ldots,X_{k,n}=a_n] \quad = \quad 2^{-n}.$$

Da die dyadischen Intervalle ein durchschnittsstabiles Erzeugendensystem der Borelschen σ-Algebra bilden, folgt, dass die Zufallsvariablen U_k auf $[0,1]$ gleichverteilt sind.

(3). Im allgemeinen Fall konstruieren wir die Zufallsvariablen aus den gerade konstruierten unabhängigen gleichverteilten Zufallsvariablen $U_k, k \in \mathbb{N}$, mithilfe des Inversionsverfahrens aus Satz 1.21: Sind $\mu_k, k \in \mathbb{N}$, beliebige Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, und

$$\underline{G}_k(u) \quad = \quad \inf\{x \in \mathbb{R} : F_k(x) \geq u\}$$

die linksstetigen verallgemeinerten Inversen der Verteilungsfunktionen

$$F_k(c) \quad = \quad \mu_k[(-\infty, c]],$$

dann setzen wir

$$Y_k(\omega) \quad := \quad \underline{G}_k(U_k(\omega)), \qquad k \in \mathbb{N}, \omega \in \Omega.$$

Da die Zufallsvariablen $U_k, k \in \mathbb{N}$, unabhängig sind, sind nach Satz 2.2 auch die $Y_k, k \in \mathbb{N}$, wieder unabhängig. Zudem gilt nach Satz 1.21:

$$P \circ Y_k^{-1} \quad = \quad \mu_k \qquad \text{für alle } k \in \mathbb{N}.$$

□

Bemerkung. (1). Der Beweis von Satz 2.9 ist konstruktiv. Für numerische Anwendungen ist allerdings zumindest der erste Schritt des beschriebenen Konstruktionsverfahrens ungeeignet, da Defizite des verwendeten Zufallszahlengenerators und die Darstellungsungenauigkeit im Rechner durch die Transformation verstärkt werden.

(2). Mithilfe des Satzes kann man auch die Existenz einer Folge unabhängiger Zufallsvariablen $X_k, k \in \mathbb{N}$, mit Werten im $\mathbb{R}^d$, oder allgemeiner in vollständigen, separablen, metrischen Räumen $S_k, k \in \mathbb{N}$, und vorgegebenen Verteilungen μ_k auf den Borelschen σ-Algebren $\mathcal{B}(S_k)$ zeigen. Sind beispielsweise $\phi_k : \mathbb{R} \to S_k$ Bijektionen, sodass ϕ_k und ϕ_k^{-1} messbar sind, und sind $\widetilde{X}_k : \Omega \to \mathbb{R}$ unabhängige reellwertige Zufallsvariablen mit Verteilungen $P[\widetilde{X}_k \in B] = \mu_K[\phi_k(B)]$, dann sind die transformierten Zufallsvariablen

$$X_k \quad = \quad \phi_k(\widetilde{X}_k) : \Omega \to S_k, \qquad \forall k \in \mathbb{N},$$

unabhängig mit Verteilungen μ_k.

2.3.2 Random Walks im $\mathbb{R}^d$

Sei μ eine Wahrscheinlichkeitsverteilung auf $(\mathbb{R}^d, \mathcal{B}(R^d))$, und seien $X_i, i \in \mathbb{N}$, unabhängige Zufallsvariablen mit identischer Verteilung $X_i \sim \mu$. Der durch

$$S_n \quad = \quad a + \sum_{i=1}^{n} X_i, \qquad n = 0, 1, 2, \ldots,$$

definierte stochastische Prozess heißt *Random Walk mit Startwert $a \in \mathbb{R}^d$ und Inkrementverteilung μ*.

Im Fall $d = 1$ können wir Stichproben von den Zufallsvariablen X_i, und damit vom Random Walk, beispielsweise mithilfe der Inversionsmethode, simulieren.

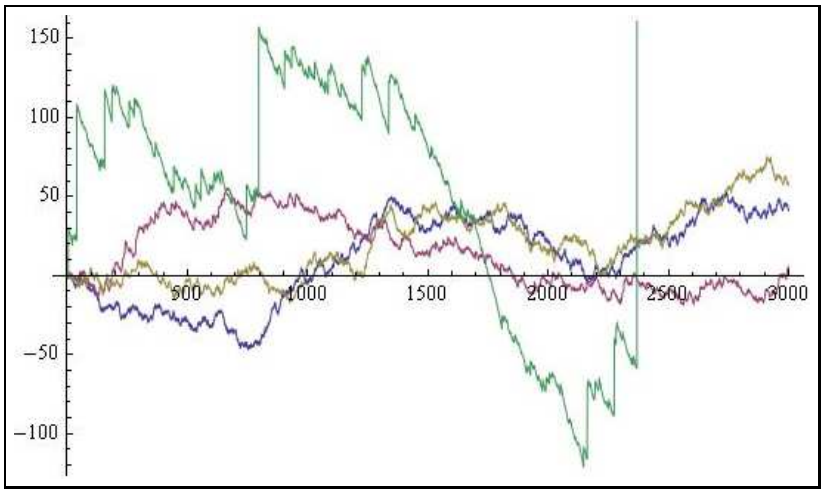


Abbildung 2.3: Grafiken von Trajektorien des Random Walks mit verschiedenen Inkrementverteilungen.

Abbildung 2.3 zeigt Grafiken von Trajektorien des Random Walks mit den Inkrementverteilungen

$$\mu \quad = \quad \frac{1}{2}(\delta_1 + \delta_{-1}) \qquad \text{(klassischer Random Walk (SSRW))},$$

$$\mu \quad = \quad N(0,1) \qquad \text{(diskrete Brownsche Bewegung)},$$

μ mit Dichte

$$f(x) \quad = \quad e^{-(x+1)} I_{(-1,\infty)(x)} \qquad \text{(zentrierte } \mathrm{Exp}(1)\text{-Verteilung)}$$

und μ mit Dichte

$$f(x) \quad = \quad 3{\cdot}2^{-5/2}{\cdot}(x{+}\frac{3}{2})^{-5/2}{\cdot} I_{(\frac{1}{2},\infty)}(x{+}\frac{3}{2}) \qquad \text{(zentrierte Pareto}(\alpha-1,\alpha)\text{-Verteilung mit } \alpha = \frac{3}{2}).$$

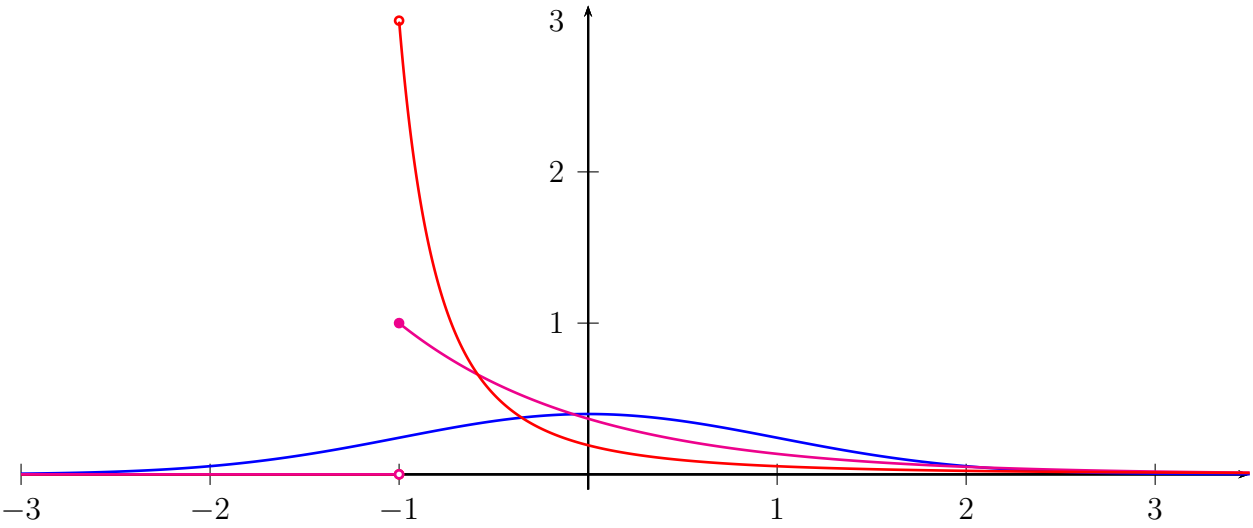


Abbildung 2.4: Dichten der drei stetigen Verteilungen aus Abbildung 2.3: $f_{N(0,1)}$ in Blau, $f_{\text{Exp}(1)-1}$ in Magenta und $f_{\text{Pareto}(\alpha-1,\alpha)}$ in Rot.

Im Gegensatz zu den anderen Verteilungen fällt die Dichte der Pareto-Verteilung für $x \to \infty$ nur sehr langsam ab („heavy tails“). Insbesondere hat die Verteilung unendliche Varianz. Die Trajektorien der Random Walks werden mit der folgenden Mathematica-Routine simuliert:

```
nmax = 10000; )
x = RandomChoice[{−1, 1}, nmax];
z = RandomReal[NormalDistribution[0, 1], nmax];
u = RandomReal[{0, 1}, nmax]; y = −Log[u] − 1;
$\alpha$ = 3/2; x0 = $\alpha$ − 1; p =
 RandomReal[ParetoDistribution[x0, $\alpha$], nmax];
 m = Mean[ParetoDistribution[x0, $\alpha$]]; q = p − m;

rwsimple = Accumulate[x]; rwexp = Accumulate[y];
rwnormal = Accumulate[z]; rwpareto = Accumulate[q];

ListLinePlot[{rwsimple[[1 ;; 3000]], rwexp[[1 ;; 3000]],
rwnormal[[1 ;; 3000]], rwpareto[[1 ;; 3000]]}]
```

Die Trajektorien des klassischen Random Walks, und der Random Walks mit exponential- und normalverteilten Inkrementen sehen in größeren Zeiträumen ähnlich aus. Die Trajektorien des Pareto-Random Walks (grün) verhalten sich dagegen anders, und werden auch in längeren Zeiträumen von einzelnen großen Sprüngen beeinflusst. Tatsächlich kann man zeigen, dass alle obigen Random Walks mit Ausnahme des Pareto-Random Walks in einem geeigneten Skalierungslimes mit Schrittweite gegen 0 in Verteilung gegen eine Brownsche Bewegung konvergieren (funktionaler zentraler Grenzwertsatz).

2.3.3 Unendliche Produktmaße

Als Konsequenz aus dem Satz können wir die Existenz von unendlichen Produktmaßen als gemeinsame Verteilung von unendlich vielen unabhängigen Zufallsvariablen zeigen. Dazu versehen wir den Folgenraum

$$\mathbb{R}^{\mathbb{N}} \quad = \quad \{(x_1, x_2, \ldots) | x_k \in \mathbb{R}, \forall\, k \in \mathbb{N}\}$$

mit der Produkt-σ-Algebra

$$\bigotimes_{k\in\mathbb{N}} \mathcal{B}(\mathbb{R}) \quad = \quad \sigma(\mathcal{C}) \quad = \quad \sigma(\pi_k | k \in \mathbb{N}),$$

die von der Kollektion $\mathcal{C}$ aller Zylindermengen

$$\{\pi_1 \in B_1, \ldots, \pi_n \in B_n\} \quad = \quad \{x = (x_k) \in \mathbb{R}^{\mathbb{N}} | x_1 \in B_1, \ldots, x_n \in B_n\},$$

$n \in \mathbb{N}, B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$, von den Koordinatenabbildungen $\pi_k : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}, \pi_k(x) = x_k$.

Korollar 2.10 (Existenz von unendlichen Produktmaßen). *Zu beliebigen Wahrscheinlichkeitsverteilungen μ_k auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ existiert eine eindeutige Wahrscheinlichkeitsverteilung $\mu = \bigotimes_{k\in\mathbb{N}} \mu_k$ auf $(\mathbb{R}^{\mathbb{N}}, \bigotimes_{k\in\mathbb{N}} \mathcal{B}(\mathbb{R}))$ mit*

$$\mu[\pi_1 \in B_1, \ldots, \pi_n \in B_n] \quad = \quad \mu[B_1] \cdot \ldots \cdot \mu_n[B_n] \tag{2.3.2}$$

für alle $n \in \mathbb{N}$ und $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$.

Definition. *Die Wahrscheinlichkeitsverteilung μ mit (2.3.2) heißt **Produkt der Wahrscheinlichkeitsverteilungen** $\mu_k, k \in \mathbb{N}$.*

Beweis. Die Eindeutigkeit folgt, da die Zylindermengen aus $\mathcal{C}$ ein $\cap$-stabiles Erzeugendensystem der Produkt-σ-Algebra bilden.

Zum Beweis der Existenz: betrachten wir die Abbildung $X : \Omega \to \mathbb{R}^{\mathbb{N}}$ mit

$$X(\omega) \quad = \quad (X_1(\omega), X_2(\omega), \ldots),$$

wobei X_k unabhängige Zufallsvariablen mit Verteilung μ_k sind. X ist messbar bzgl. $\bigotimes_{k\in\mathbb{N}} \mathcal{B}(\mathbb{R})$, denn

$$X^{-1}[\{x \in \mathbb{R}^{\mathbb{N}} | (x_1, \ldots, x_n) \in B\}] \quad = \quad \{\omega \in \Omega | (X_1(\omega), \ldots, X_n(\omega)) \in B\} \quad \in \quad \mathcal{A}$$

für alle $n \in \mathbb{N}$ und $B \in \mathcal{B}(\mathbb{R}^n)$. Sei $\mu = P \circ X^{-1}$ die Verteilung von X auf $\mathbb{R}^{\mathbb{N}}$. Dann gilt

$$\begin{aligned}\mu[\pi_1 \in B_1, \ldots, \pi_m \in B_n] &= \mu[\{x \in \mathbb{R}^{\mathbb{N}} | x_1 \in B_1, \ldots, x_n \in B_n\}] \\ &= P[X_1 \in B_1, \ldots, X_n \in B_n] \\ &= \prod_{k=1}^{n} \mu_k[B_k]\end{aligned}$$

für alle $n \in \mathbb{N}$ und $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$. Also ist μ das gesuchte Produktmaß. □

Bemerkung. Auf analoge Weise folgt nach Bemerkung 2. von oben die Existenz des Produktmaßes $\bigotimes_{k\in\mathbb{N}} \mu_k$ von beliebigen Wahrscheinlichkeitsverteilungen $\mu_k, k \in \mathbb{N}$, auf vollständigen, separablen, messbaren Räumen S_k mit Borelschen σ-Algebren $\mathcal{S}_k$. Das Produktmaß sitzt auf dem Produktraum

$$\left(\bigtimes_{k\in\mathbb{N}} S_k, \bigotimes_{k\in\mathbb{N}} \mathcal{S}_k \right).$$

Der Satz von Carathéodory impliziert sogar die Existenz von beliebigen (auch überabzählbaren) Produkten von allgemeinen Wahrscheinlichkeitsräumen $(S_i, \mathcal{S}_i, \mu_i), i \in I$.

Sind $(S_i, \mathcal{S}_i, \mu_i)$ beliebige Wahrscheinlichkeitsräume, dann sind die Koordinatenabbildungen $\pi_k : \bigtimes_{i\in\mathbb{N}} S_i \to S_k$ unter dem Produktmaß $\bigotimes_{i\in I} \mu_i$ unabhängig und μ_k-verteilt. Man nennt den Produktraum

$$(\Omega, \mathcal{A}, P) \quad = \quad \left(\bigtimes S_i, \bigotimes \mathcal{S}_i, \bigotimes \mu_i \right)$$

daher auch das **kanonische Modell** für unabhängige μ_i-verteilte Zufallsvariablen.

2.4 Das 0-1-Gesetz von Kolmogorov

2.4.1 Asymptotische Ereignisse

Sei X_i $(i \in I)$ eine unendliche Kollektion von Zufallsvariablen, die auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definiert sind.

Definition. *Ein Ereignis $A \in \sigma(X_i \,|i \in I)$ heißt **asymptotisches Ereignis (tail event)**, falls*

$$A \in \sigma(X_i \mid i \in I \backslash I_0) \qquad \textit{für jede } \textbf{endliche} \textit{ Teilmenge } I_0 \subseteq I \textit{ gilt.}$$

Die Menge

$$\tau \quad = \quad \bigcap_{I_0 \subseteq I \text{ endlich}} \sigma(X_i \mid i \in I \backslash I_0)$$

*aller asymptotischen Ereignisse ist eine σ-Algebra. τ heißt **asymptotische σ-Algebra (tail field)**.*

Beispiel. (1). DYNAMISCH: Ist $X_n, n \in \mathbb{N}$ eine Folge von Zufallsvariablen (welche beispielsweise eine zufällige zeitliche Entwicklung beschreibt), dann gilt für ein Ereignis $A \in \sigma(X_n, n \in \mathbb{N})$:

$$A \text{ asymptotisch} \quad \Leftrightarrow \quad A \in \underbrace{\sigma(X_{n+1}, X_{n+2}, \ldots)}_{\text{Zukunft ab } n} \qquad \text{für alle } n.$$

Beispiele für asymptotische Ereignisse von reellwertigen Zufallsvariablen sind

$$\{X_n > 5n \quad \text{unendlich oft}\}, \qquad \left\{\limsup_{n\to\infty} X_n < c\right\}, \qquad \{\exists \lim_{n\to\infty} X_n\}, \qquad \left\{\exists \lim \frac{1}{n} S_n = m\right\},$$

wobei $S_n = X_1 + \ldots + X_n$. Die Ereignisse

$$\{\sup_{n\in\mathbb{N}} X_n = 3\} \qquad \text{und} \qquad \{\lim S_n = 5\}$$

sind dagegen *nicht* asymptotisch.

(2). STATISCH: Eine Kollektion $X_i, i \in \mathbb{Z}^d$, von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ heißt **stochastisches Feld** (random field). Beispielsweise basieren verschiedene grundlegende Modelle der statistischen Mechanik auf stochastischen Feldern $X_i : \Omega \to \{0, 1\}$, wobei $X_i = 1$ dafür steht, dass

- sich ein Teilchen am Gitterpunkt i befindet,
- ein Atom am Gitterpunkt i angeregt ist,
- der Gitterpunkt i durchlässig ist (Perkolationsmodell),
- etc.

Asymptotische Ereignisse beschreiben in diesem Fall „makroskopische“ Effekte.

Satz 2.11 (0-1-Gesetz von Kolmogorov). *Sind X_i $(i \in I)$ unabhängige Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$, dann gilt*

$$P[A] \in \{0, 1\} \textbf{ für alle } A \in \tau.$$

„Asymptotische Ereignisse sind deterministisch.“

Beweis. Der Übersichtlichkeit halber führen wir den Beweis im Fall $I = \mathbb{N}$ - der Beweis im allgemeinen Fall verläuft ähnlich. Es gilt: $X_1, X_2, ...$ unabhängige Zufallsvariablen
$\Longrightarrow \sigma(X_1), \sigma(X_2), ..., \sigma(X_n), \sigma(X_{n+1}), \sigma(X_{n+2}), ...$ unabhängige Mengensysteme
$\Longrightarrow \sigma(X_1, ..., X_n), \sigma(X_{n+1}, X_{n+2}, ...)$ sind unabhängig für alle $n \in \mathbb{N}$

$\Longrightarrow \sigma(X_1, ..., X_n)$ und τ sind unabhängig für alle $n \in \mathbb{N}$
$\Longrightarrow \tau$ unabhängig von $\sigma(X_1, X_2, ...) \supseteq \tau$
$\Longrightarrow$ Ereignisse $A \in \tau$ sind unabhängig von sich selbst
$\Longrightarrow P[A] \in \{0, 1\} \qquad \forall\, A \in \tau$.
Hierbei gilt die zweite Implikation nach Satz 2.1 (2), und die vierte nach Satz 2.1 (1) □

2.4.2 Asymptotische Zufallsvariablen

Definition. *Eine Zufallsvariable $Y : \Omega \to [-\infty, \infty]$ heißt **asymptotisch**, wenn die bzgl. der asymptotischen σ-Algebra τ messbar ist.*

Korollar 2.12. *Sind X_i $(i \in I)$ unabhängige Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, dann ist jede asymptotische Zufallsvariable $Y : \Omega \to [-\infty, \infty]$ P - fast sicher konstant, d.h.*

$$\exists\, c_0 \in [-\infty, \infty] : \quad P[Y = c_0] = 1 .$$

Beweis. Ist Y τ - messbar, dann sind die Ereignisse $\{Y \leq c\}, \quad c \in \mathbb{R}$, in τ enthalten. Aus dem Kolmogorovschen 0-1-Gesetz folgt:

$$F_Y(c) \quad = \quad P[Y \leq c] \quad \in \quad \{0, 1\} \qquad \forall\, c \in \mathbb{R}.$$

Da die Verteilungsfunktion monoton wachsend ist, existiert ein $c_0 \in [-\infty, \infty]$ mit

$$P[Y \leq c] \quad = \quad \begin{cases} 0 & \text{für} \quad c < c_0 \\ 1 & \text{für} \quad c > c_0 \end{cases},$$

und damit $P[Y = c_0] \quad = \quad \lim\limits_{\varepsilon \downarrow 0}(F_Y(c_0) - F_Y(c_0 - \varepsilon)) \quad = \quad 1. = 1.$ □

Beispiele für asymptotische Zufallsvariablen im Fall $I = \mathbb{N}$ sind etwa

$$\underline{\lim_{n\to\infty}} X_n, \quad \overline{\lim_{n\to\infty}} X_n, \quad \underline{\lim_{n\to\infty}} \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{sowie} \quad \overline{\lim_{n\to\infty}} \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Insbesondere sind für unabhängige Zufallsvariablen $X_1, X_2, ... : \Omega \to \mathbb{R}$ sowohl

$$\underline{\lim}\, \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{als auch} \quad \overline{\lim}\, \frac{1}{n}\sum_{i=1}^{n} X_i \quad P \text{ - f.s. konstant.}$$

Hieraus ergibt sich die folgende *Dichotomie*: Sind $X_i, i \in \mathbb{N}$, unabhängige reellwertige Zufallsvariablen, dann gilt *entweder* ein Gesetz großer Zahlen, d.h.

$$\frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{konvergiert } P \text{ - f.s., und der Limes ist } P \text{ - f.s. konstant}$$

(falls der Limes inferior und Limes superior P-fast sicher übereinstimmen), *oder*

$$P\left[\frac{1}{n}\sum_{i=1}^{n} X_i \text{ konvergiert}\right] \;=\; 0.$$

Es ist bemerkenswert, dass für die Gültigkeit der Dichotomie keine Annahmen über die Verteilung der X_i benötigt werden. Insbesondere müssen die X_i nicht identisch verteilt sein!

2.4.3 Anwendungen auf Random Walks und Perkolationsmodelle

Beispiel (Rückkehr zum Startpunkt von Random Walks, Rekurrenz). Wir betrachten einen eindimensionalen klassischen Random Walk mit Startpunkt $a \in \mathbb{Z}$ und unabhängigen Inkrementen X_i mit Verteilung

$$P[X_i = 1] \quad = \quad p, \qquad P[X_i = -1] \quad = \quad 1 - p.$$

Für $n \in \mathbb{N}$ erhält man die Rückkehrwahrscheinlichkeiten

$$P[S_{2n+1} = a] = 0$$
$$P[S_{2n} = a] = \binom{2n}{n} \cdot p^n \cdot (1-p)^n = \frac{(2n)!}{(n!)^2} \cdot p^n \cdot (1-p)^n.$$

Wir betrachten nun die Asymptotik für $n \to \infty$ dieser Wahrscheinlichkeiten. Aus der **Stirlingschen Formel**

$$n! \quad \sim \quad \sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n$$

folgt

$$P[S_{2n} = a] \quad \sim \quad \frac{\sqrt{4\pi n}}{2\pi n} \cdot \frac{(\frac{2n}{e})^{2n}}{(\frac{n}{e})^{2n}} \cdot p^n \cdot (1-p)^n \quad = \quad \frac{1}{\sqrt{\pi n}}(4p(1-p))^n \quad \text{für } n \to \infty.$$

Für $p \neq \frac{1}{2}$ fallen die Wahrscheinlichkeiten also exponentiell schnell ab. Insbesondere gilt dann

$$\sum_{m=0}^{\infty} P[S_m = a] \quad = \quad \sum_{n=0}^{\infty} P[S_{2n} = a] < \infty,$$

d.h. der asymmetrische Random Walk kehrt nach dem 1. Borel-Cantelli Lemma mit Wahrscheinlichkeit 1 nur endlich oft zum Startpunkt zurück (TRANSIENZ). Nach dem starken Gesetz großer Zahl gilt sogar

$$S_n \quad \sim \quad (2p-1)n \qquad P\text{-fast sicher.}$$

Für $p = \frac{1}{2}$ gilt dagegen $P[S_{2n} = a] \sim 1/\sqrt{\pi n}$, und damit

$$\sum_{m=0}^{\infty} P[S_m = a] \quad = \quad \sum_{n=0}^{\infty} P[S_{2n} = a] \quad = \quad \infty.$$

Dies legt nahe, dass der Startpunkt mit Wahrscheinlichkeit 1 unendlich oft besucht wird.

Ein Beweis dieser Aussage über das Borel-Cantelli-Lemma ist aber nicht direkt möglich, da die Ereignisse $\{S_{2n} = 0\}$ nicht unabhängig sind. Wir beweisen nun eine stärkere Aussage mithilfe des Kolmogorovschen 0-1-Gesetzes:

Satz 2.13 (Rekurrenz und unbeschränkte Oszillationen des symmetrischen Random Walks). *Für $p = \frac{1}{2}$ gilt*

$$P[\overline{\lim}\, S_n = +\infty \text{ und } \underline{\lim}\, S_n = -\infty] \quad = \quad 1.$$

*Insbesondere ist der eindimensionale Random Walk **rekurrent**, d.h.*

$$P[S_n = a \text{ unendlich oft}] = 1.$$

Tatsächlich wird nach dem Satz mit Wahrscheinlichkeit 1 sogar jeder Punkt $\lambda \in \mathbb{Z}$ unendlich oft getroffen.

Beweis. Für alle $k \in \mathbb{N}$ gilt:

$$P[S_{n+k} - S_n = k \text{ unendlich oft}] = 1,$$

denn nach dem Beispiel zu Satz 2.1 („Affe tippt Shakespeare“) gibt es P-fast sicher unendlich viele Blöcke der Länge k mit $X_{n+1} = X_{n+2} = ... = X_{n+k} = 1$. Es folgt

$$P[\overline{\lim}\, S_n - \underline{\lim}\, S_n = \infty] \quad \geq \quad P\left[\bigcap_k \bigcup_n \{S_{n+k} - S_n = k\}\right] \quad = \quad 1,$$

und damit

$$1 \quad = \quad P[\overline{\lim}\, S_n = +\infty \text{ oder } \underline{\lim}\, S_n = -\infty] \quad \leq \quad P[\overline{\lim}\, S_n = +\infty] + P[\underline{\lim}\, S_n = -\infty].$$

Also ist eine der beiden Wahrscheinlichkeiten auf der rechten Seite größer als $\frac{1}{2}$, und damit nach dem Kolmogorovschen 0-1-Gesetz gleich 1. Aus Symmetriegründen folgt

$$P[\underline{\lim}\, S_n = -\infty] \quad = \quad P[\overline{\lim}\, S_n = +\infty] \quad = \quad 1.$$

□

Das vorangehende Beispiel zeigt eine typische Anwendung des Kolmogorovschen 0-1-Gesetzes auf stochastische Prozesse. Um die Anwendbarkeit in räumlichen Modellen zu demonstrieren, betrachten wir ein einfaches Perkolationsmodell:

Beispiel (Perkolation im $\mathbb{Z}^d$). Sei $p \in (0,1)$ fest, und seien X_i $(i \in \mathbb{Z}^d)$ unabhängige Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit

$$P[X_i = 1] = p \quad , \qquad P[X_i = 0] = 1 - p \, .$$

Ein Gitterpunkt $i \in \mathbb{Z}^d$ heißt **durchlässig**, falls $X_i = 1$ gilt. Wir verbinden Gitterpunkte $i, j \in \mathbb{Z}^d$ mit $|i - j| = 1$ durch eine Kante. Sei A das Ereignis, dass bzgl. dieser Graphenstruktur eine unendliche Zusammenhangskomponente (Cluster) aus durchlässigen Gitterpunkten existiert (Eine Flüssigkeit könnte in diesem Fall durch ein makroskopisches Modellstück, das aus mikroskopischen Gitterpunkten aufgebaut ist, durchsickern - daher der Name „Perkolation“). A ist asymptotisch, also gilt nach dem Satz von Kolmogorov

$$P[A] \in \{0, 1\}.$$

Hingegen ist es im Allgemeinen nicht trivial, zu entscheiden, welcher der beiden Fälle eintritt. Im Fall $d = 1$ zeigt man leicht (Übung):

$$P[A] = 0 \quad \text{für alle } p < 1.$$

Für $d = 2$ gilt:

$$P[A] = 1 \quad \Longleftrightarrow \quad p > \frac{1}{2} \, ,$$

s. z.B. die Monografie *„Percolation“* von *Grimmett*. Für $d \geq 3$ ist nur bekannt, dass ein kritischer Parameter $p_c \in (0,1)$ existiert mit

$$P[A] = \begin{cases} 1 & \text{für } p > p_c. \\ 0 & \text{für } p < p_c. \end{cases}$$

Man kann obere und untere Schranken für p_c herleiten (z.B. gilt $\frac{1}{2d-1} \leq p_c \leq \frac{2}{3}$), aber der genaue Wert ist nicht bekannt. Man vermutet, dass $P[A] = 0$ für $p = p_c$ gilt, aber auch diese Aussage konnte bisher nur in Dimension $d \geq 19$ (sowie für $d = 2$) bewiesen werden, siehe das Buch von Grimmett.

Das Perkolationsmodell ist ein Beispiel für ein sehr einfach formulierbares stochastisches Modell, das zu tiefgehenden mathematischen Problemstellungen führt. Es ist von großer Bedeutung, da ein enger Zusammenhang zu anderen Modellen der statistischen Mechanik und dabei auftretenden Phasenübergängen besteht. Einige elementare Aussagen über Perkolationsmodelle werden in den Wahrscheinlichkeitstheorie-Lehrbüchern von *Y. Sinai* und *A. Klenke* hergeleitet.

Kapitel 3

Integration bzgl. Wahrscheinlichkeitsmaßen

In diesem Kapitel definieren wir den Erwartungswert, die Varianz und die Kovarianz allgemeiner reellwertiger Zufallsvariablen, und beweisen grundlegende Eigenschaften. Einen weiteren Schwerpunkt bildet das Rechnen mit Dichten.

Da wir auch Grenzübergänge durchführen wollen, erweist es sich als günstig, die Werte $+\infty$ und $-\infty$ zuzulassen. Wir setzen daher $\overline{\mathbb{R}} = [-\infty, \infty]$. Der Raum $\overline{\mathbb{R}}$ ist ein topologischer Raum bzgl. des üblichen Konvergenzbegriffs. Die Borelsche σ-Algebra auf $\overline{\mathbb{R}}$ wird u.a. erzeugt von den Intervallen $[-\infty, c], c \in \mathbb{R}$. Die meisten Aussagen über reellwertige Zufallsvariablen aus den vorangegangenen Abschnitten übertragen sich unmittelbar auf Zufallsvariablen $X : \Omega \to \overline{\mathbb{R}}$, wenn wir die Verteilungsfunktion $F_X : \mathbb{R} \to [0, 1]$ definieren durch

$$F_X(c) \quad = \quad \mu_X[[-\infty, c]] \quad = \quad P[X \leq c].$$

3.1 Erwartungswert als Lebesgue-Integral

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \overline{\mathbb{R}}$ eine Zufallsvariable. Wir wollen den Erwartungswert von X bezüglich der Wahrscheinlichkeitsverteilung P in sinnvoller Weise definieren. Dazu gehen wir schrittweise vor:

3.1.1 Definition des Erwartungswerts

Indikatorfunktionen

Der Erwartungswert der Indikatorfunktion I_A einer meßbaren Menge $A \in \mathcal{A}$ bzgl. P ist definiert durch

$$E[I_A] = P[A].$$

Elementare Zufallsvariablen

Nimmt X nur endlich viele Werte $c_1, ..., c_n \in \mathbb{R}$ an, dann können wir X als Linearkombination von Indikatorfunktionen darstellen. Da der Erwartungswert linear von der Zufallsvariable abhängen sollte, definieren wir $E[X]$ dann als die entsprechende Linearkombination der Erwartungswerte der Indikatorfunktionen:

Definition (Erwartungswert von elementaren Zufallsvariablen). *Eine Zufallsvariable von der Form*

$$X = \sum_{i=1}^{n} c_i I_{A_i} \qquad \textit{mit } n \in \mathbb{N},\ c_i \in \mathbb{R},\ \textit{und } A_i \in \mathcal{A}$$

heißt ***elementar****. Ihr* ***Erwartungswert*** *bzgl. P ist*

$$E[X] = \sum_{i=1}^{n} c_i \cdot P[A_i]. \tag{3.1.1}$$

Wir müssen zeigen, dass der Erwartungswert $E[X]$ durch (3.1.1) *wohldefiniert*, d.h. unabhängig von der gewählten Darstellung der Zufallsvariable X als Linearkombination von Indikatorfunktionen ist. Dazu bemerken wir, dass $E[X]$ unabhängig von der Darstellung mit dem in Kapitel **??** für diskrete Zufallsvariablen definierten Erwartungswert

$$\widetilde{E}[X] \quad = \quad \sum_{i=1}^{n} c_i \cdot P[X = c_i]$$

übereinstimmt. In der Tat folgt nämlich aus der Linearität von $\widetilde{E}[\cdot]$:

$$E[X] = \sum_{i=1}^{n} c_i \cdot E[I_{A_i}] = \sum_{i=1}^{n} c_i \cdot \widetilde{E}[I_{A_i}] = \widetilde{E}[X].$$

Die Abbildung $X \mapsto E[X]$ ist offensichtlich *linear* und *monoton*:

$$E[aX + bY] \quad = \quad a \cdot E[X] + b \cdot E[Y] \qquad \text{für alle } a, b \in \mathbb{R},$$

$$X \leq Y \quad \Longrightarrow \quad E[X] \leq E[Y].$$

Die Definition des Erwartungswerts einer elementaren Zufallsvariable stimmt genau mit der des Lebesgue-Integrals der Elementarfunktion X bzgl. des Maßes P überein:

$$E[X] \quad = \quad \int X \, dP \quad = \quad \int X(\omega) \, P(d\omega)$$

Für allgemeine Zufallsvariablen liegt es nahe, den Erwartungswert ebenfalls als Lebesgueintegral bzgl. des Maßes P zu definieren. Wir skizzieren hier die weiteren Schritte zur Konstruktion des Lebesgue-Integrals bzw. des Erwartungswerts einer allgemeinen Zufallsvariable, siehe auch die Analysisvorlesung.

Nichtnegative Zufallsvariablen

Die Definition des Erwartungswerts bzw. Lebesgue-Integrals einer nichtnegativen Zufallsvariable bzgl. P beruht auf der monotonen Approximation durch elementare Zufallsvariablen:

Lemma 3.1 (Monotone Approximation durch elementare Zufallsvariablen). *Sei $X : \Omega \to [0, \infty]$ eine nichtnegative Zufallsvariable auf $(\Omega, \mathcal{A}, P)$. Dann existiert eine monoton wachsende Folge elementarer Zufallsvariablen $0 \leq X_1 \leq X_2 \leq \ldots$ mit*

$$X(\omega) \quad = \quad \lim_{n\to\infty} X_n(\omega) \quad = \quad \sup_{n\in\mathbb{N}} X_n(\omega) \qquad \textit{für alle } \omega \in \Omega.$$

Beweis. Für $n \in \mathbb{N}$ sei

$$X_n(\omega) := \begin{cases} (k-1)\cdot 2^{-n} & \text{falls } (k-1)\cdot 2^{-n} \leq X(\omega) < k \cdot 2^{-n} \text{ für ein } k = 1, 2, \ldots, n\cdot 2^n, \\ n & \text{falls } X(\omega) \geq n. \end{cases}$$

Dann ist X_n eine elementare Zufallsvariable, denn es gilt

$$X_n \quad = \quad \sum_{k=1}^{n2^n} \frac{k-1}{2^n} I_{\left\{\frac{k-1}{2^n} \leq X < \frac{k}{2^n}\right\}} + n I_{\{X \geq n\}}.$$

Die Folge $X_n(\omega)$ ist für jedes ω monoton wachsend, da die Unterteilung immer feiner wird, und

$$\sup_{n\in\mathbb{N}} X_n(\omega) \quad = \quad \lim_{n\to\infty} X_n(\omega) \quad = \quad X(\omega) \qquad \text{für alle } \omega \in \Omega.$$

□

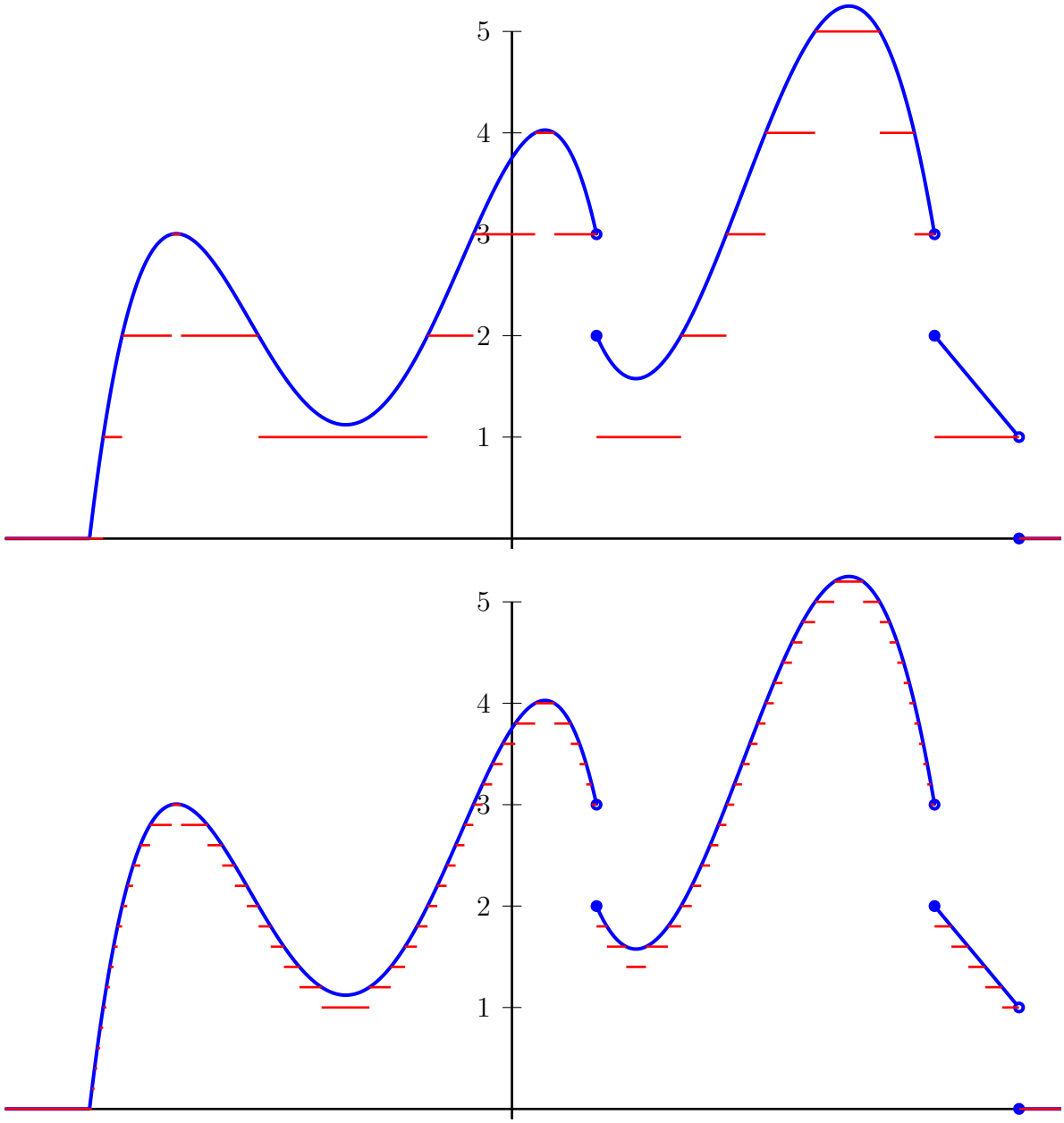


Abbildung 3.1: Approximation durch Elementarfunktionen. Hier ist die Annäherung in rot in zwei verschiedenen Feinheiten dargestellt.

Definition (Erwartungswert einer nicht-negativen Zufallsvariable). *Der **Erwartungswert** (bzw. das **Lebesgueintegral**) einer Zufallsvariable $X : \Omega \to [0, \infty]$ bzgl. P ist definiert als*

$$E[X] := \lim_{n\to\infty} E[X_n] = \sup_{n\to\infty} E[X_n] \in [0, \infty], \tag{3.1.2}$$

wobei $(X_n)_{n\in\mathbb{N}}$ eine beliebige monoton wachsende Folge von nichtnegativen elementaren Zufallsvariablen mit $X = \sup X_n$ ist.

Auch in diesem Fall ist der Erwartungswert wohldefiniert (in $[0, \infty]$):

Lemma 3.2 (**Wohldefiniertheit**). *Die Definition ist unabhängig von der Wahl einer monoton wachsenden Folge (X_n) von nichtnegativen elementaren Zufallsvariablen mit $X = \sup_{n\in\mathbb{N}} X_n$.*

Für den Beweis verweisen wir auf die Analysisvorlesung oder auf die Literatur, siehe z.B. Appendix 5 in WILLIAMS „Probability with martingales“.

Bemerkung (**Monotone Stetigkeit und monotone Konvergenz**). Sind $X_n = I_{A_n}$ und $X = I_A$ Indikatorfunktionen, dann folgt (3.1.2) aus der monotonen Stetigkeit von P. In diesem Fall gilt nämlich:

$$X_n \nearrow X \qquad \Longleftrightarrow \qquad A_n \nearrow A \qquad (\text{d.h. } A_n \text{ monoton wachsend und } A = \bigcup A_n).$$

Aus der monotonen Stetigkeit von P folgt dann $E[X] = P[A] = \lim P[A_n] = \lim E[X_n]$.

Die Monotonie des Erwartungswerts überträgt sich von elementaren auf nichtnegative Zufallsvariablen:

Lemma 3.3 (**Monotonie des Erwartungswerts**). *Für nichtnegative Zufallsvariablen X, Y mit $X \le Y$ gilt $E[X] \le E[Y]$.*

Beweis. Ist $X \le Y$, dann gilt auch $X_n \le Y_n$ für die im Beweis von Lemma 3.1 konstruierten, approximierenden elementaren Zufallsvariablen, also

$$E[X] \quad = \quad \sup_{n\in\mathbb{N}} E[X_n] \quad \le \quad \sup_{n\in\mathbb{N}} E[Y_n] \quad = \quad E[Y].$$

□

Allgemeine Zufallsvariablen

Eine allgemeine Zufallsvariable $X : \Omega \to \overline{\mathbb{R}}$ können wir in ihren positiven und negativen Anteil zerlegen:

$$X \quad = \quad X^+ - X^- \qquad \text{mit} \qquad X^+ \quad := \quad \max(X, 0), \quad X^- \quad := \quad -\min(X, 0).$$

X^+ und X^- sind beides nichtnegative Zufallsvariablen. Ist mindestens einer der beiden Erwartungswerte $E[X^+]$ bzw. $E[X^-]$ endlich, dann können wir (ähnlich wie schon in Kapitel **??** für diskrete Zufallsvariablen) definieren:

Definition (**Erwartungswert einer allgemeinen Zufallsvariable**). *Für eine Zufallsvariable $X : \Omega \to \overline{\mathbb{R}}$ mit $E[X^+] < \infty$ oder $E[X^-] < \infty$ ist der Erwartungswert (bzw. das Lebesgue-Integral) bzgl. P definiert als*

$$E[X] \quad := \quad E[X^+] - E[X^-] \quad \in \quad [-\infty, \infty].$$

Notation: Da der Erwartungswert $E[X]$ das Lebesgue-Integral der meßbaren Funktion $X : \Omega \to \overline{\mathbb{R}}$ bzgl. des Maßes P ist, verwenden wir auch folgende Notation:

$$E[X] \quad = \quad \int X \, dP \quad = \quad \int X(\omega) \, P(d\omega).$$

Bemerkung (Integration bzgl. σ-endlicher Maße). Die Definition des Lebesgue-Integrals kann auf σ-endliche Maße erweitert werden. Ein Maß Q auf $(\Omega, \mathcal{A})$ heißt σ-endlich, falls meßbare Mengen $B_1, B_2, \ldots \subseteq \Omega$ existieren mit $Q[B_n] < \infty$ für alle $n \in \mathbb{N}$ und $\Omega = \bigcup B_n$.

3.1.2 Eigenschaften des Erwartungswerts

Nachdem wir den Erwartungswert einer allgemeinen Zufallsvariable $X : \Omega \to \overline{\mathbb{R}}$ definiert haben, fassen wir nun einige elementare Eigenschaften zusammen. Dazu bezeichnen wir mit

$$\mathcal{L}^1 \; = \; \mathcal{L}^1(P) \; = \; \mathcal{L}^1(\Omega, \mathcal{A}, P) \; := \; \{X : \Omega \to \overline{\mathbb{R}} \text{ Zufallsvariable} : E[|X|] < \infty\}$$

die Menge aller **bzgl. P integrierbaren** Zufallsvariablen. Für eine $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ ist nach Lemma 3.3 sowohl $E[X^+]$ als auch $E[X^-]$ endlich. Also ist der Erwartungswert $E[X]$ definiert und endlich.

Satz 3.4 (Linearität und Monotonie des Erwartungswerts). *Für Zufallsvariablen $X, Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ und $a, b \in \mathbb{R}$ gilt:*

(1). $X \geq 0$ P-fast sicher $\Longrightarrow E[X] \geq 0$.

(2). Die Zufallsvariable $aX + bY$ ist bzgl. P integrierbar, und

$$E[aX + bY] \quad = \quad a \cdot E[X] + b \cdot E[Y].$$

Insbesondere ist der Erwartungswert monoton:

(3). $X \leq Y$ P-fast sicher $\Longrightarrow E[X] \leq E[Y]$.

Zum Beweis der Eigenschaften (1) und (2) verweisen wir auf die Analysisvorlesung oder die Literatur. Eigenschaft (3) folgt unmittelbar aus (1) und (2).

Nach Aussage (2) des Satzes ist $\mathcal{L}^1(\Omega, \mathcal{A}, P)$ ein Vektorraum. Durch

$$X \sim Y \quad :\Longleftrightarrow \quad P[X \neq Y] = 0$$

wird eine Äquivalenzrelation auf diesem Raum definiert. Eine Konsequenz von Aussage (3) des Satzes ist, dass zwei äquivalente (also P-fast sicher identische) Zufallsvariablen denselben Erwartungswert haben:

$$X \sim Y \qquad \Longrightarrow \qquad E[X] = E[Y].$$

Daher ist der Erwartungswert einer Äquivalenzklasse von P-fast sicher gleichen Zufallsvariablen eindeutig definiert. In Zukunft verwenden wir häufig dieselbe Notation für die Äquivalenzklassen und Repräsentanten aus den Äquivalenzklassen. Satz 3.4 besagt, dass der Erwartungswert ein *positives lineares Funktional* auf dem Raum

$$L^1(\Omega, \mathcal{A}, P) \quad := \quad \mathcal{L}^1(\Omega, \mathcal{A}, P)/\sim$$

aller Äquivalenzklassen von integrierbaren Zufallsvariablen definiert. Aus dem Satz folgt zudem:

Korollar 3.5. *Durch*

$$\|X\|_{L^1(\Omega,\mathcal{A},P)} \quad = \quad E[|X|]$$

wird eine Norm auf $L^1(\Omega, \mathcal{A}, P)$ definiert. Insbesondere gilt für Zufallsvariablen $X : \Omega \to \overline{\mathbb{R}}$:

$$E[|X|] = 0 \qquad \Longrightarrow \qquad X = 0 \quad P\text{-fast sicher.}$$

Beweis. Für eine Zufallsvariable $X : \Omega \to \overline{\mathbb{R}}$ mit $E[|X|] = 0$ und $\varepsilon > 0$ gilt wegen der Monotonie und Linearität des Erwartungswerts:

$$P[|X| \geq \varepsilon] \;=\; E\left[I_{\{|X|\geq\varepsilon\}}\right] \quad \leq \quad E\left[\frac{|X|}{\varepsilon}\right] \quad = \quad \frac{1}{\varepsilon}E[|X|] \quad = \quad 0.$$

Für $\varepsilon \searrow 0$ erhalten wir

$$P[|X| > 0] \;=\; \lim_{\varepsilon\searrow 0} P[|X| \geq \varepsilon] \;=\; 0,$$

also $X = 0$ P-fast sicher.
Zudem folgt aus der Monotonie und Linearität des Erwartungswerts die Dreiecksungleichung:

$$E[|X+Y|] \quad \leq \quad E[|X| + |Y|] \quad = \quad E[|X|] + E[|Y|].$$

□

In der Analysis wird gezeigt, dass der Raum $L^1(\Omega, \mathcal{A}, P)$ bzgl. der im Korollar definierten Norm ein Banachraum ist.

3.1.3 Konvergenzsätze

Ein entscheidender Vorteil des Lebesgue-Integrals gegenüber anderen Integrationsbegriffen ist die Gültigkeit von sehr allgemeinen Konvergenzsätzen (Vertauschen von Grenzwerten und Integralen bzw. Erwartungswerten). Dass solche Aussagen sehr wichtig sind, und keineswegs immer gelten, demonstriert das folgende Beispiel:

Beispiel (Setzen mit Verdoppeln). Wir betrachten Setzen mit Verdoppeln auf »Zahl« für eine Folge von fairen Münzwürfen. Bei Anfangseinsatz 1 beträgt das Kapital des Spielers nach n Würfen

$$X_n \quad = \quad 2^n \cdot I_{\{n<T\}} \, ,$$

wobei T die Wartezeit auf den ersten »Kopf« ist. Es folgt

$$E[X_n] \; = \; 2^n P[T > n] \; = \; 2^n \cdot 2^{-n} \; = \; 1 \quad \text{für alle } n \in \mathbb{N},$$

das Spiel ist also fair. Andererseits fällt aber P-fast sicher irgendwann einmal »Kopf«, d.h.

$$\lim_{n\to\infty} X_n \; = \; 0 \quad P\text{-fast sicher.}$$

In dieser Situation ist also

$$\lim E[X_n] \; = \; 1 \; \neq \; 0 \; = \; E[\lim X_n] \qquad !!!$$

Die Voraussetzungen der Konvergenzsätze 3.6 und 3.9 sind hier nicht erfüllt.

Die im folgenden betrachteten Konvergenzsätze lassen sich alle zurückführen auf einen fundamentalen Konvergenzsatz. Dieser ergibt sich aus der oben skizzierten Konstruktion des Lebesgue-Integrals, und charakterisiert dieses zusammen mit der Linearität und Monotonie:

Satz 3.6 (Satz von der monotonen Konvergenz, B. Levi). *Ist $(X_n)_{n\in\mathbb{N}}$ eine monoton wachsende Folge von Zufallsvariablen mit $E[X_1^-] < \infty$ (z.B. $X_1 \geq 0$), dann gilt:*

$$E[\sup_{n\in\mathbb{N}} X_n] \quad = \quad E[\lim_{n\to\infty} X_n] \quad = \quad \lim_{n\to\infty} E[X_n] \quad = \quad \sup_{n\in\mathbb{N}} E[X_n].$$

Der Beweis findet sich in zahlreichen Lehrbüchern der Integrations- oder Warscheinlichkeitstheorie, siehe z.B. WILLIAMS: PROBABILITY WITH MARTINGALES, APPENDIX 5.

Eine erste wichtige Konsequenz des Satzes von der monotonen Konvergenz ist:

Korollar 3.7. *Für nichtnegative Zufallsvariablen $X_i, i \in \mathbb{N}$, gilt:*

$$E\left[\sum_{i=1}^{\infty} X_i\right] \quad = \quad \sum_{i=1}^{\infty} E[X_i].$$

Beweis. Durch Anwenden von Satz 3.6 auf die Partialsummen erhalten wir:

$$\begin{aligned} E\left[\sum_{i=1}^{\infty} X_i\right] &= E\left[\lim_{n\to\infty} \sum_{i=1}^{n} X_i\right] \\ &= \lim_{n\to\infty} E\left[\sum_{i=1}^{n} X_i\right] && \text{(wegen monotoner Konvergenz)} \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} E[X_i] && \text{(wegen Linearität)} \\ &= \sum_{i=1}^{\infty} E[X_i]. \end{aligned}$$

□

Bemerkung (**Abzählbare Wahrscheinlichkeitsräume, Summation als Spezialfall von Integration**)**.** Falls Ω abzählbar ist, können wir jede Zufallsvariable $X : \Omega \to \mathbb{R}$ auf die folgende Weise als abzählbare Linearkombination von Indikatorfunktionen darstellen:

$$X = \sum_{\omega\in\Omega} X(\omega) \cdot I_{\{\omega\}}.$$

Ist $X \geq 0$, dann gilt nach Korollar 3.7:

$$E[X] = \sum_{\omega\in\Omega} X(\omega) \cdot P[\{\omega\}].$$

Dieselbe Darstellung des Erwartungswerts gilt auch für allgemeine reellwertige Zufallsvariablen auf Ω, falls der Erwartungswert definiert ist, d.h. falls $E[X^+]$ oder $E[X^-]$ endlich ist. Damit sehen wir, dass *Summation ein Spezialfall von Integration* ist: Ist Ω abzählbar und $p(\omega) \geq 0$ für alle $\omega \in \Omega$, dann gilt

$$\sum_{\omega\in\Omega} X(\omega) \cdot p(\omega) = \int X \, dP,$$

wobei P das Maß mit Massenfunktion p ist. Beispielsweise gilt auch

$$\sum_{\omega\in\Omega} X(\omega) = \int X \, d\nu,$$

wobei ν das durch $\nu[A] = |A|$, $A \subseteq \Omega$, definierte **Zählmaß** ist. Insbesondere bemerken wir:

Konvergenzsätze wie der Satz von der monotonen Konvergenz lassen sich auch auf abzählbare Summen anwenden!

Wir beweisen nun noch zwei wichtige Konvergenzsätze, die sich aus dem Satz von der monotonen Konvergenz ergeben:

Korollar 3.8 (Lemma von Fatou). *Seien $X_1, X_2, \cdots : \Omega \to \overline{\mathbb{R}}$ Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ und sei $Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ (z.B. $Y \equiv 0$).*

(1). Gilt $X_n \geq Y$ für alle $n \in \mathbb{N}$, dann folgt

$$E[\liminf X_n] \leq \liminf\ E[X_n].$$

(2). Gilt $X_n \leq Y$ für alle $n \in \mathbb{N}$, dann folgt

$$E[\limsup X_n] \geq \limsup\ E[X_n].$$

Beweis. Die Aussagen folgen aus dem Satz über monotone Konvergenz. Beispielsweise gilt:

$$\begin{aligned} E\left[\liminf\ X_n\right] &= E\left[\lim_{n\to\infty} \inf_{k\geq n} X_k\right] = \lim_{n\to\infty} E\left[\inf_{k\geq n} X_k\right] \\ &\leq \lim_{n\to\infty} \inf_{k\geq n} E[X_k] = \liminf_{n\to\infty} E[X_n], \end{aligned}$$

da die Folge der Infima monoton wachsend ist und durch die integrierbare Zufallsvariable Y nach unten beschränkt ist. Die zweite Aussage zeigt man analog. □

Korollar 3.9 (Satz von der majorisierten Konvergenz, Lebesgue). *Sei $X_n : \Omega \to \overline{\mathbb{R}}$ $(n \in \mathbb{N})$ eine P-fast sicher konvergente Folge von Zufallsvariablen. Existiert eine integrierbare Majorante, d.h. eine Zufallsvariable $Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ mit $|X_n| \leq Y$ für alle $n \in \mathbb{N}$, dann gilt*

$$E[\lim X_n] \quad = \quad \lim\ E[X_n]. \tag{3.1.3}$$

Beweis. Aus dem Lemma von Fatou folgt

$$E[\liminf X_n]\ \leq\ \liminf\ E[X_n]\ \leq\ \limsup\ E[X_n]\ \leq\ E[\limsup X_n],$$

da $X_n \geq -Y \in \mathcal{L}^1$ und $X_n \leq Y \in \mathcal{L}^1$ für alle $n \in \mathbb{N}$ gilt. Konvergiert X_n P-fast sicher, dann stimmen die linke und rechte Seite der obigen Ungleichungskette überein. □

3.2 Berechnung von Erwartungswerten; Dichten

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. In diesem Abschnitt zeigen wir, wie man in verschiedenen Fällen den Erwartungswert einer Zufallsvariable $X : \Omega \to [0, \infty]$ aus der Verteilung von X berechnen kann.

3.2.1 Diskrete Zufallsvariablen

Falls X nur abzählbar viele Werte annimmt, können wir die Zufallsvariable X auf folgende Weise als abzählbare Linearkombination von Indikatorfunktionen darstellen:

$$X = \sum_{a \in X(\Omega)} a \cdot I_{\{X=a\}}.$$

Nach Korollar 3.7 folgt

$$E[X] \;=\; \sum_{a \in X(\Omega)} E[a \cdot I_{\{X=a\}}] \;=\; \sum_{a \in X(\Omega)} a \cdot P[X=a].$$

Dieselbe Aussage gilt allgemeiner für diskrete reellwertige Zufallsvariablen X mit $E[X^+] < \infty$ oder $E[X^-] < \infty$.

Für Zufallsvariablen $X : \Omega \to S$, mit Werten in einer beliebigen abzählbaren Menge S, und eine Borel-messbare Funktion $h : S \to \overline{\mathbb{R}}$ erhalten wir entsprechend

$$E[h(X)] \;=\; \sum_{a \in X(\Omega)} h(a) \cdot P[X=a], \tag{3.2.1}$$

falls $E[h(X)]$ definiert ist, also z.B. falls $h \geq 0$ oder $h(X) \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ gilt.

Die allgemeine Definition des Erwartungswerts als Lebesgueintegral stimmt also für diskrete Zufallsvariablen mit der in Kapitel **??** gegebenen Definition überein.

3.2.2 Allgemeine Zufallsvariablen

Die Berechnungsformel (3.2.1) für den Erwartungswert diskreter Zufallsvariablen lässt sich auf Zufallsvariablen mit beliebigen Verteilungen erweitern. Sei dazu $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $(S, \mathcal{S})$ ein messbarer Raum, $X : \Omega \to S$ eine Zufallsvariable, und $h : S \to [0, \infty]$ eine messbare Abbildung.

Satz 3.10 (Transformationssatz). *Unter den obigen Voraussetzungen gilt:*

$$E_P[h(X)] \quad = \quad \int h(X(\omega))\, P(d\omega) \quad = \quad \int h(x)\, \mu(dx) \quad = E_\mu[h],$$

wobei $\mu = P \circ X^{-1}$ die Verteilung von X unter P ist, und E_P bzw. E_μ den Erwartungswert unter P bzw. μ bezeichnet.

Die Erwartungswerte hängen somit nur von der Verteilung von X ab!

Beweis. Der Beweis erfolgt in drei Schritten:

(1). Ist $h = I_B$ die Indikatorfunktion einer messbaren Menge $B \in \mathcal{S}$, dann erhalten wir

$$E[h(X)] \;=\; \int I_B(X(\omega))\, P(d\omega) \;=\; P[X^{-1}(B)] \;=\; \mu[B] \;=\; \int I_B \, d\mu,$$

da $I_B(X(\omega)) = I_{X^{-1}(B)}(\omega)$ für alle $\omega \in \Omega$ gilt.

(2). Für Linearkombinationen $h = \sum_{i=1}^{n} c_i I_{B_i}$ von Indikatorfunktionen mit $n \in \mathbb{N}, c_i \in \mathbb{R}$, und $B_i \in \mathcal{S}$ gilt die Aussage auch, da das Lebesgueintegral linear vom Integranden abhängt.

(3). Für eine allgemeine messbare Funktion $h \geq 0$ existiert schließlich eine monoton wachsende Folge h_n von Elementarfunktionen mit $h_n(x) \nearrow h(x)$ für alle $x \in S$. Durch zweimalige Anwendung des Satzes von der monotonen Konvergenz erhalten wir erneut:

$$E[h(X)] \;=\; E[\lim h_n(X)] \;=\; \lim E[h_n(X)] \;=\; \lim \int h_n \, d\mu \;=\; \int h \, d\mu.$$

□

Bemerkung. Das hier verwendete ***Beweisverfahren der »maßtheoretischen Induktion«*** wird uns noch sehr häufig begegnen: Wir zeigen eine Aussage

(1). für Indikatorfunktionen,

(2). für Elementarfunktionen,

(3). für nichtnegative meßbare Funktionen,

(4). für allgemeine integrierbare Funktionen.

Für eine integrierbare reellwertige Zufallsvariable X definieren wir genau wie für diskrete Zufallsvariablen (s. Abschnitt **??**) die **Varianz** $\mathrm{Var}[X]$ und die **Standardabweichung** $\sigma[X]$ durch

$$\mathrm{Var}[X] \quad := \quad E[(X - E[X])^2], \qquad \sigma[X] \quad := \quad \sqrt{\mathrm{Var}[X]}\,.$$

Auch in diesem Fall folgen aus der Linearität des Erwartungswerts die Rechenregeln

$$\mathrm{Var}[X] \;=\; E[X^2] - E[X]^2, \qquad \text{und} \tag{3.2.2}$$

$$\mathrm{Var}[aX + b] \;=\; \mathrm{Var}[aX] \;=\; a^2 \cdot \mathrm{Var}[X] \qquad \text{für alle } a, b \in \mathbb{R}. \tag{3.2.3}$$

Insbesondere ist die Varianz genau dann endlich, wenn $E[X^2]$ endlich ist. Nach Korollar 3.5 gilt zudem genau dann $\mathrm{Var}[X] = 0$, wenn X P-fast sicher konstant gleich $E[X]$ ist. Aufgrund des Transformationssatzes können wir den Erwartungswert und die Varianz auch allgemein aus der Verteilung $\mu_X = P \circ X^{-1}$ berechnen:

Korollar 3.11 (Berechnung von Erwartungswert und Varianz aus der Verteilung). *Der Erwartungswert $E[X]$ und die Varianz $\mathrm{Var}[X]$ einer reellen Zufallsvariable X hängen nur von der Verteilung $\mu_X = P \circ X^{-1}$ ab:*

$$\mathrm{Var}[X] \quad = \quad \int (x - \overline{x})^2 \, \mu_X(dx) \qquad \textit{mit} \qquad \overline{x} \; = \; E[X] \; = \; \int x \, \mu_X(dx).$$

Beweis. Durch Anwenden von Satz 3.10 auf die nicht-negativen Zufallsvariablen X^+ und X^- erhalten wir

$$E[X] \; = \; E[X^+] - E[X^-] \; = \; \int x^+ \mu_X(dx) - \int x^- \mu_X(dx) \; = \; \int x \mu_X(dx), \qquad \text{und}$$

$$\mathrm{Var}[X] \quad = \quad E[(X - E[X])^2] \quad = \quad \int (x - E[X])^2 \mu_X(dx).$$

□

Der nächste Satz zeigt unter anderem, wie man den Erwartungswert einer nicht-negativen Zufallsvariable T konkret aus der Verteilungsfunktion

$$F_T(t) \; = \; P[T \leq t] \; = \; \mu_T[[0,t]], \qquad t \in \mathbb{R}, \qquad \text{berechnet:}$$

Satz 3.12 (Berechnung von Erwartungswerten aus der Verteilungsfunktion). *Für eine Zufallsvariable $T : \Omega \to \mathbb{R}_+$ und eine stetig differenzierbare Funktion $h : \mathbb{R}_+ \to \mathbb{R}_+$ gilt*

$$E[h(T)] \; = \; \int_0^\infty P[T > t] \, h'(t) \, dt \; = \; \int_0^\infty (1 - F_T(t)) \, h'(t) \, dt.$$

Beweis. Wegen

$$h(T(\omega)) \; = \; \int_0^{T(\omega)} h'(t) \, dt \; = \; \int_0^\infty I_{\{T>t\}}(\omega) \, h'(t) \; dt$$

erhalten wir

$$E[h(T)] \; = \; E\left[\int_0^\infty I_{\{T>t\}} \, h'(t) \, dt\right] \; = \; \int_0^\infty E\left[I_{\{T>t\}} h'(t)\right] \; dt \; = \; \int_0^\infty P[T > t] \, h'(t) \, dt.$$

Hierbei haben wir im Vorgriff auf Abschnitt 3.3 den *Satz von Fubini* benutzt, der gewährleistet, dass man zwei Lebesgueintegrale (das Integral über t und den Erwartungswert) unter geeigneten Voraussetzungen (Produktmeßbarkeit) vertauschen kann, siehe Satz 3.16. □

Bemerkung (Lebesgue-Stieltjes-Integral). Das Lebesgue-Stieltjes-Integral $\int g \, dF$ einer meßbaren Funktion $g : \mathbb{R} \to [0, \infty]$ bzgl. der Verteilungsfunktion F einer Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}$ ist definiert als das Lebesgue-Integral von g bzgl. μ:

$$\int g(t) \, dF(t) \quad := \quad \int g(t) \, \mu(dt).$$

Ist g stetig, dann lässt sich das Integral als Limes von Riemann-Summen darstellen. Nach dem Transformationssatz gilt für eine Zufallsvariable $T : \Omega \to \mathbb{R}_+$ und $h \in C(\mathbb{R}_+, \mathbb{R}_+)$:

$$E[h(T)] \quad = \quad \int h(t)\, \mu_T(dt) \quad = \quad \int h(t)\, dF_T(t).$$

Die Aussage von Satz 3.12 ergibt sich hieraus formal durch partielle Integration.

Beispiel (Exponentialverteilung). Für eine zum Parameter $\lambda > 0$ exponentialverteilte Zufallsvariable T erhalten wir

$$E[T] \;=\; \int_0^\infty P[T > t]\, dt \quad = \quad \int_0^\infty e^{-\lambda t}\, dt \quad = \quad \frac{1}{\lambda}.$$

Die mittlere Wartezeit ist also der Kehrwert der Intensität. Mit partieller Integration folgt zudem

$$\begin{aligned} E[T^2] \;&=\; \int_0^\infty P[T > t]\, 2t\, dt \;=\; \int_0^\infty 2te^{-\lambda t}\, dt \;=\; \frac{2}{\lambda} \int_0^\infty e^{-\lambda t}\, dt \;=\; \frac{2}{\lambda^2}, \qquad \text{also} \\ \sigma[T] \;&=\; \sqrt{\mathrm{Var}[T]} \quad = \quad (E[T^2] - E[T]^2)^{1/2} \quad = \quad \frac{1}{\lambda}. \end{aligned}$$

Die Standardabweichung ist somit genauso groß wie der Erwartungswert!

Beispiel (Heavy tails). Sei $\alpha > 0$. Für eine Zufallsvariable $T : \Omega \to [0, \infty)$ mit

$$P[T > t] \quad \sim \quad t^{-\alpha} \qquad \text{für } t \to \infty$$

ist der Erwartungswert $E[T] = \int_0^\infty P[T > t]\, dt$ genau dann endlich, wenn $\alpha > 1$ ist. Allgemeiner ist das p-te Moment

$$E\left[T^p\right] \quad = \quad \int_0^\infty P\left[T^p > t\right]\, dt \quad = \quad \int_0^\infty \underbrace{P\left[T > t^{1/p}\right]}_{\sim t^{-\alpha/p}}\, dt$$

genau für $p < \alpha$ endlich. Beispielsweise ist T für $\alpha \in (1, 2]$ integrierbar, aber die Varianz ist in diesem Fall unendlich.

3.2.3 Zufallsvariablen mit Dichten

Die Verteilungen vieler Zufallsvariablen haben eine Dichte bzgl. des Lebesguemaßes, oder bzgl. eines anderen geeigneten Referenzmaßes. In diesem Fall können wir Erwartungswerte durch eine Integration bzgl. des Referenzmaßes berechnen. Sei $(S, \mathcal{S})$ ein messbarer Raum und ν ein Maß auf $(S, \mathcal{S})$ (z.B. das Lebesguemaß oder eine Wahrscheinlichkeitsverteilung).

Definition (Wahrscheinlichkeitsdichte). *Eine **Wahrscheinlichkeitsdichte** auf $(S, \mathcal{S}, \nu)$ ist eine meßbare Funktion $\varrho : S \to [0, \infty]$ mit*

$$\int_S \varrho(x)\, \nu(dx) \;=\; 1.$$

Satz 3.13 (Integration bzgl. Wahrscheinlichkeitsmaßen mit Dichten).

(1). Ist ϱ eine Wahrscheinlichkeitsdichte auf $(S, \mathcal{S}, \nu)$, dann wird durch

$$\mu[B] \quad := \quad \int_B \varrho(x)\, \nu(dx) \quad = \quad \int I_B(x)\varrho(x)\, \nu(dx) \tag{3.2.4}$$

eine Wahrscheinlichkeitsverteilung μ auf $(S, \mathcal{S})$ definiert.

(2). Für jede meßbare Funktion $h : S \to [0, \infty]$ gilt

$$\int h(x)\, \mu(dx) \quad = \quad \int h(x)\varrho(x)\, \nu(dx). \tag{3.2.5}$$

Insbesondere folgt nach dem Transformationssatz:

$$E[h(X)] \quad = \quad \int h(x)\varrho(x)\, \nu(dx)$$

für jede Zufallsvariable X mit Verteilung μ.

Beweis. Wir zeigen zunächst, dass μ eine Wahrscheinlichkeitsverteilung auf $(S, \mathcal{S})$ ist: Sind $B_1, B_2, ... \in \mathcal{S}$ disjunkte Mengen, so folgt wegen $\varrho \geq 0$ nach Korollar 3.7:

$$\begin{aligned} \mu\left[\bigcup B_i\right] \;&=\; \int I_{\bigcup B_i}(x) \cdot \varrho(x)\, \nu(dx) \;=\; \int \sum I_{B_i}(x) \cdot \varrho(x)\, \nu(dx) \\ &=\; \sum \int I_{B_i}(x) \cdot \varrho(x)\, \nu(dx) \quad = \quad \sum \mu[B_i]. \end{aligned}$$

Zudem gilt $\mu[S] = \int \varrho\, d\nu = 1$, da ϱ eine Wahrscheinlichkeitsdichte ist.

Die Aussage (3.2.5) über den Erwartungswert beweisen wir durch maßtheoretische Induktion:

(1). Die Aussage folgt unmittelbar, wenn $h = I_B$ für $B \in \mathcal{S}$ gilt.

(2). Für Linearkombinationen $h = \sum_{i=1}^n c_i I_{B_i}$ folgt die Aussage aus der Linearität beider Seiten von (3.2.5) in h.

(3). Für allgemeine $h \geq 0$ existiert eine Teilfolge h_n aus Elementarfunktionen mit $h_n \nearrow h$. Mit monotoner Konvergenz folgt dann

$$\int h\, d\mu \quad = \quad \lim \int h_n\, d\mu \quad = \quad \lim \int h_n \varrho\, d\nu \quad = \quad \int h\varrho\, d\nu.$$

□

Bemerkung (Eindeutigkeit der Dichte). Durch (3.2.4) wird die Dichte $\varrho(x)$ der Wahrscheinlichkeitsverteilung μ bzgl. des Referenzmaßes ν für ν-fast alle x eindeutig festgelegt: Existiert eine weitere Funktion $\widetilde{\varrho} \in \mathcal{L}^1(S, \mathcal{S}, \nu)$ mit

$$\int_B \widetilde{\varrho}\, d\nu \quad = \quad \mu[B] \quad = \quad \int_B \varrho\, d\nu \qquad \text{für alle } B \in \mathcal{S},$$

dann folgt

$$\begin{aligned} \int (\varrho - \widetilde{\varrho})^+ \, d\nu &= \int_{\{\varrho > \widetilde{\varrho}\}} (\varrho - \widetilde{\varrho})\, d\nu \;=\; 0, \qquad \text{und entsprechend} \\ \int (\varrho - \widetilde{\varrho})^- \, d\nu &= \int_{\{\varrho < \widetilde{\varrho}\}} (\widetilde{\varrho} - \varrho)\, d\nu \;=\; 0. \end{aligned}$$

Somit erhalten wir $(\varrho - \widetilde{\varrho})^+ = (\varrho - \widetilde{\varrho})^- = 0$ ν-fast überall, also $\varrho = \widetilde{\varrho}$ ν-fast überall.

Notation. Die Aussage (3.2.5) rechtfertigt in gewissem Sinn die folgenden Notationen für eine Wahrscheinlichkeitsverteilung μ mit Dichte ϱ bzgl. ν:

$$\mu(dx) \quad = \quad \varrho(x)\, \nu(dx) \qquad \text{bzw.} \qquad d\mu \quad = \quad \varrho\, d\nu \qquad \text{bzw.} \qquad \mu \quad = \quad \varrho \cdot \nu.$$

Für die nach der Bemerkung ν-fast überall eindeutig bestimmte Dichte von μ bzgl. ν verwenden wir dementsprechend auch die Notation

$$\varrho(x) \quad = \quad \frac{d\mu}{d\nu}(x).$$

Wichtige Spezialfälle.

(1). MASSENFUNKTION ALS DICHTE BZGL. DES ZÄHLMASSES.
Das Zählmaß auf einer abzählbaren Menge S ist durch

$$\nu[B] \quad = \quad |B| \qquad \text{für } B \subseteq S$$

definiert. Die Massenfunktion $x \mapsto \mu[\{x\}]$ einer Wahrscheinlichkeitsverteilung μ auf S ist die Dichte von μ bzgl. des Zählmaßes ν. Insbesondere ist die Massenfunktion einer Zufallsvariable $X : \Omega \to S$ die Dichte der Verteilung von X bzgl. ν:

$$\mu_X[B] \quad = \quad P[X \in B] \quad = \quad \sum_{a \in B} p_X(a) \quad = \quad \int_B p_X(a) \nu(da) \qquad \text{für alle } B \subseteq S.$$

Die Berechnungsformel für den Erwartungswert diskreter Zufallsvariablen ergibt sich damit als Spezialfall von Satz 3.13:

$$E[h(X)] \overset{3.13}{=} \int h(a)p_X(a)\,\nu(da) = \sum_{a\in S} h(a)p_X(a) \qquad \text{für alle } h : S \to [0,\infty].$$

(2). DICHTEN BZGL. DES LEBESGUE-MASSES.
Eine Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}^d$ mit Borelscher σ-Algebra hat genau dann eine Dichte ϱ bzgl. des Lebesgue-Maßes λ, wenn

$$\mu[(-\infty, c_1] \times \ldots \times (-\infty, c_d]] = \int_{-\infty}^{c_1} \cdots \int_{-\infty}^{c_d} \varrho(x_1, \ldots, x_d)\, dx_d \cdots dx_1$$

für alle $(c_1, \ldots, c_d) \in \mathbb{R}^d$ gilt. Insbesondere hat die Verteilung einer reellwertigen Zufallsvariable X genau dann die Dichte f_X bzgl. λ, wenn

$$F_X(c) = \mu_X[(-\infty, c]] = \int_{-\infty}^{c} f_X(x)\, dx \qquad \text{für alle } c \in \mathbb{R}$$

gilt. Die Verteilungsfunktion ist in diesem Fall eine Stammfunktion der Dichte, und damit λ-fast überall differenzierbar mit Ableitung

$$F_X'(x) = f_X(x) \qquad \text{für fast alle } x \in \mathbb{R}.$$

Für den Erwartungswert ergibt sich

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

für alle meßbaren Funktionen $h : \mathbb{R} \to \mathbb{R}$ mit $h \geq 0$ oder $h \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$.

Beispiel (Normalverteilungen). Die Dichte der Standardnormalverteilung bzgl. des Lebesgue-Maßes ist $\varrho(x) = (2\pi)^{-1/2} e^{-x^2/2}$. Damit erhalten wir für den Erwartungswert und die Varianz einer Zufallsvariable $Z \sim N(0,1)$:

$$E[Z] = \int_{-\infty}^{\infty} x \cdot (2\pi)^{-1/2} e^{-x^2/2}\, dx = 0,$$

und, mit partieller Integration,

$$\mathrm{Var}[Z] = E[Z^2] = \int_{-\infty}^{\infty} x^2 \cdot (2\pi)^{-1/2} e^{-x^2/2}\, dx = \int_{-\infty}^{\infty} (2\pi)^{-1/2} \cdot e^{-x^2/2}\, dx = 1.$$

Ist X eine $N(m, \sigma^2)$-verteilte Zufallsvariable, dann ist $Z = \frac{X-m}{\sigma}$ standardnormalverteilt, und es gilt $X = m + \sigma Z$, also

$$E[X] \quad = \quad m + \sigma E[Z] \quad = \quad m, \qquad \text{und}$$

$$\mathrm{Var}[X] \quad = \quad \mathrm{Var}[\sigma Z] \quad = \quad \sigma^2 \, \mathrm{Var}[Z] \quad = \quad \sigma^2.$$

Die Parameter m und σ geben also den Erwartungswert und die Standardabweichung der Normalverteilung an.

(3). RELATIVE DICHTEN. Seien μ und ν zwei Wahrscheinlichkeitsverteilungen auf einem meßbaren Raum $(S, \mathcal{S})$ mit Dichten f bzw. g bezüglich eines Referenzmaßes λ (zum Beispiel bzgl. des Zählmaßes oder des Lebesgue-Maßes). Gilt $g > 0$ λ-fast überall, dann hat μ bzgl. ν die Dichte

$$\frac{d\mu}{d\nu} \quad = \quad \frac{f}{g} \quad = \quad \frac{d\mu/d\lambda}{d\nu/d\lambda},$$

denn nach Satz 3.13 gilt für alle $B \in \mathcal{S}$:

$$\mu[B] \quad = \quad \int_B f \, d\lambda \quad = \quad \int_B \frac{f}{g} \, g \, d\lambda \; = \; \int_B \frac{f}{g} \, d\nu \, .$$

In der Statistik treten relative Dichten als „Likelihoodquotienten“ auf, wobei $f(x)$ bzw. $g(x)$ die „Likelihood“ eines Beobachtungswertes x bzgl. verschiedener möglicher zugrundeliegender Wahrscheinlichkeitsverteilungen beschreibt, s. Abschnitt 6.4.

3.2.4 Existenz von Dichten

Wir geben nun ohne Beweis den Satz von Radon-Nikodym an. Dieser Satz besagt, dass eine Wahrscheinlichkeitsverteilung genau dann eine Dichte bzgl. eines anderen (σ-endlichen) Maßes ν hat, wenn alle ν-Nullmengen auch μ-Nullmengen sind. Hierbei nennen wir ein Maß ν auf einem meßbaren Raum $(S, \mathcal{S})$ σ-**endlich**, wenn eine Folge von meßbaren Mengen $B_n \in \mathcal{S}$ mit $\nu[B_n] < \infty$ und $S = \bigcup_{n \in \mathbb{N}} B_n$ existiert.

Definition (Absolutstetigkeit und Singularität von Maßen).

*(1). Ein Maß μ auf $(S, \mathcal{S})$ heißt **absolutstetig** bzgl. eines anderen Maßes ν auf demselben meßbaren Raum ($\mu \ll \nu$) falls für alle $B \in \mathcal{S}$ gilt:*

$$\nu[B] = 0 \qquad \Longrightarrow \qquad \mu[B] = 0$$

*μ und ν heißen **äquivalent** ($\mu \approx \nu$), falls $\mu \ll \nu$ und $\nu \ll \mu$.*

*(2). Die Maße μ und ν heißen **singulär** ($\mu \perp \nu$), falls eine Menge $B \in \mathcal{S}$ mit $\nu[B] = 0$ und $\mu[B^C] = 0$ existiert.*

Beispiel (Dirac-Maß). Ein Diracmaß δ_x, $x \in \mathbb{R}$, ist singulär zum Lebesguemaß λ auf $\mathbb{R}$.

Satz 3.14 (Radon-Nikodym). *Für ein Wahrscheinlichkeitsmaß μ und ein σ-endliches Maß ν gilt $\mu \ll \nu$ genau dann, wenn eine Dichte $\varrho \in \mathcal{L}^1(S, \mathcal{S}, \nu)$ existiert mit*

$$\mu[B] \quad = \quad \int_B \varrho \, d\nu \qquad \text{für alle } B \in \mathcal{S}.$$

Die eine Richtung des Satzes zeigt man leicht: Hat μ eine Dichte bzgl. ν, und gilt $\nu[B] = 0$, so folgt

$$\mu[B] \quad = \quad \int_B \varrho \, d\nu \quad = \quad \int \varrho \cdot I_B \, d\nu \quad = \quad 0$$

wegen $\varrho \cdot I_B = 0$ ν-fast überall. Der Beweis der Umkehrung ist nicht so einfach, und kann funktionalanalytisch erfolgen, siehe z.B. Klenke: „Wahrscheinlichkeitstheorie“. Ein stochastischer Beweis des Satzes von Radon-Nikodym basierend auf dem Martingal-Konvergenzsatz findet sich z.B. in Williams: „Probability with martingales”.

Beispiel (Absolutstetigkeit von diskreten Wahrscheinlichkeitsverteilungen). Sind μ und ν Maße auf einer abzählbaren Menge S, dann gilt $\mu \ll \nu$ genau dann, wenn $\mu(x) = 0$ für alle $x \in S$ mit $\nu(x) = 0$ gilt. In diesem Fall ist die Dichte von μ bzgl. ν durch

$$\frac{d\mu}{d\nu}(x) \quad = \quad \begin{cases} \dfrac{\mu(x)}{\nu(x)} & \text{falls } \nu(x) \neq 0, \\ \text{beliebig} & \text{sonst,} \end{cases}$$

gegeben. Man beachte, dass die Dichte nur für ν-fast alle x (also für alle x mit $\nu(x) \neq 0$) eindeutig bestimmt ist.

Seien nun μ und ν zwei Wahrscheinlichkeitsmaße auf $(S, \mathcal{S})$. Ist μ absolutstetig bzgl. ν mit Dichte $d\mu/d\nu$, dann gilt nach Satz 3.13:

$$\int f \, d\mu \quad = \quad \int f \cdot \frac{d\mu}{d\nu} \, d\nu \qquad \text{für alle meßbaren Funktionen } f : S \to \mathbb{R}_+. \tag{3.2.6}$$

Die folgenden elementaren Aussagen ergeben sich unmittelbar aus (3.2.6):

Satz 3.15. *(1). Ist μ absolutstetig bzgl. ν mit ν-fast überall strikt positiver relativer Dichte, dann ist auch ν absolutstetig bzgl. μ, und*

$$\frac{d\nu}{d\mu}(x) \quad = \quad \left(\frac{d\mu}{d\nu}(x) \right)^{-1} \qquad \text{für } \mu\text{-fast alle } x \in S.$$

(2). Sind μ und ν beide absolutstetig bzgl. eines Referenzmaßes λ mit Dichten f und g, und gilt $g > 0$ λ-fast überall, dann ist μ absolutstetig bzgl. ν mit relativer Dichte

$$\frac{d\mu}{d\nu}(x) \quad = \quad \frac{f(x)}{g(x)} \qquad \textit{für } \nu\textit{-fast alle } x \in S.$$

(3). Sind $\mu_1, \ldots, \mu_n$ und $\nu_1, \ldots, \nu_n$ Wahrscheinlichkeitsverteilungen auf meßbaren Räumen $(S_1, \mathcal{S}_1), \ldots, (S_n, \mathcal{S}_n)$ mit $\mu_i \ll \nu_i$ für alle $1 \leq i \leq n$, dann ist auch $\mu_1 \otimes \mu_2 \otimes \ldots \otimes \mu_n$ absolutstetig bzgl. $\nu_1 \otimes \nu_2 \otimes \ldots \otimes \nu_n$ mit relativer Dichte

$$\frac{d(\mu_1 \otimes \ldots \mu_n)}{d(\nu_1 \otimes \ldots \otimes \nu_n)}(x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^{n} \frac{d\mu_i}{d\nu_i}(x_i).$$

Die letzte Aussage gilt nicht für unendliche Produkte:

Beispiel (Singularität von unendlichen Produktmaßen). Sind μ und ν zwei unterschiedliche Wahrscheinlichkeitsverteilungen auf einem meßbaren Raum $(S, \mathcal{S})$, dann ist das unendliche Produkt $\mu^\infty := \bigotimes_{i \in \mathbb{N}} \mu$ *nicht absolutstetig* zu $\nu^\infty := \bigotimes_{i \in \mathbb{N}} \nu$. In der Tat gilt nämlich nach dem Gesetz der großen Zahlen:

$$\begin{aligned}
\mu^\infty \left[\left\{ (x_1, x_2, \ldots) \in S^\infty \; : \; \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} I_B(x_i) \; = \; \mu[B] \right\} \right] \quad &= \quad 1 \\
\nu^\infty \left[\left\{ (x_1, x_2, \ldots) \in S^\infty \; : \; \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} I_B(x_i) \; = \; \nu[B] \right\} \right] \quad &= \quad 1
\end{aligned}$$

für alle $B \in \mathcal{S}$. Ist $\mu \neq \nu$, dann existiert eine Menge $B \in \mathcal{S}$ mit $\mu[B] \neq \nu[B]$. Also sind die Wahrscheinlichkeitsverteilungen μ^∞ und ν^∞ in diesem Fall sogar *singulär*. In Satz 6.7 werden wir sehen, dass die relativen Dichten $d\mu^n / d\nu^n$ der endlichen Produktmaße für $\mu \neq \nu$ und $n \to \infty$ exponentiell schnell anwachsen.

3.3 Mehrstufige Modelle und bedingte Dichten

Seien $(S_i, \mathcal{S}_i), 1 \leq i \leq n$, meßbare Räume. Wir wollen allgemeine Wahrscheinlichkeitsverteilungen auf dem Produktraum $S_1 \times ... \times S_n$ konstruieren und effektiv beschreiben. In Analogie zu diskreten, mehrstufigen Modellen versuchen wir diese in der Form

$$P(dx_1 ... dx_n) = \mu(dx_1) p(x_1, dx_2) p((x_1, x_2), dx_3) \cdots p((x_1, ..., x_{n-1}), dx_n)$$

darzustellen.

3.3.1 Stochastische Kerne und der Satz von Fubini

Wir betrachten zunächst den Fall $n = 2$, der allgemeine Fall ergibt sich dann durch Iteration der Konstruktion. Seien also $(S, \mathcal{S})$ und $(T, \mathscr{T})$ meßbare Räume, und sei

$$\Omega := S \times T, \quad \text{und} \quad \mathcal{A} := \mathcal{S} \otimes \mathscr{T} \quad \text{die Produkt-}\sigma\text{-Algebra.}$$

Unser Ziel ist die Konstruktion einer Wahrscheinlichkeitsverteilung P auf $(\Omega, \mathcal{A})$ vom Typ

$$P(dxdy) \;=\; \mu(dx)p(x, dy).$$

Definition (Stochastischer Kern). *Eine Abbildung*

$$p : S \times \mathscr{T} \longrightarrow [0, 1], \qquad (x, C) \mapsto p(x, C),$$

*heißt **stochastischer Kern**, wenn gilt:*

(i) $p(x, \bullet)$ *ist für jedes* $x \in S$ *eine Wahrscheinlichkeitsverteilung auf* $(T, \mathscr{T})$*, und*

(ii) $p(\bullet, C)$ *ist für jedes* $C \in \mathscr{T}$ *eine meßbare Funktion auf* $(S, \mathcal{S})$*.*

Bemerkung (Diskreter Spezialfall). Sind S und T abzählbar mit σ-Algebren $\mathcal{S} = \mathscr{P}(S)$ und $\mathscr{T} = \mathscr{P}(T)$, dann ist p eindeutig festgelegt durch die Matrix mit Komponenten

$$p(x, y) \;:=\; p(x, \{y\}) \qquad (x \in S\ ,\ y \in T).$$

Da p ein stochastischer Kern ist, ist $p(x, y)$ $(x \in S, y \in T)$ eine *stochastische Matrix*.

Der folgende Satz zeigt im allgemeinen Fall die Existenz eines zweistufigen Modells mit μ als Verteilung der ersten Komponente, und $p(x, \bullet)$ als bedingte Verteilung der zweiten Komponente gegeben den Wert x der ersten Komponente. Der Satz zeigt zudem, dass Erwartungswerte im mehrstufigen Modell durch Hintereinanderausführen von Integralen berechnet werden können.

Satz 3.16 (Fubini). *Sei* $\mu(dx)$ *eine Wahrscheinlichkeitsverteilung auf* $(S, \mathcal{S})$ *und* $p(x, dy)$ *ein stochastischer Kern von* $(S, \mathcal{S})$ *nach* $(T, \mathscr{T})$*. Dann existiert eine eindeutige Wahrscheinlichkeitsverteilung* $\mu \otimes p$ *auf* $(\Omega, \mathcal{A})$ *mit*

$$(\mu \otimes p)[B \times C] = \int_B \mu(dx)\; p(x, C) \qquad \textit{für alle } B \in \mathcal{S},\ C \in \mathscr{T}. \tag{3.3.1}$$

Für diese Wahrscheinlichkeitsverteilung gilt:

$$\int f\; d(\mu \otimes p) = \int \left(\int f(x, y)\; p(x, dy) \right) \mu(dx) \qquad \textit{für alle } \mathcal{A}\textit{-messbaren } f : \Omega \to \mathbb{R}_+. \tag{3.3.2}$$

Beweis. (1). *Eindeutigkeit:* Das Mengensystem $\{B \times C \mid B \in \mathcal{S},\ C \in \mathscr{T}\}$ ist ein durchschnittsstabiler Erzeuger der Produkt-σ-Algebra $\mathcal{A}$. Also ist die Wahrscheinlichkeitsverteilung $\mu \otimes \nu$ durch (3.3.1) eindeutig festgelegt.

(2). *Existenz:* Wir wollen die Wahrscheinlichkeitsverteilung $\mu \otimes p$ über (3.3.2) mit $f = I_A,\ A \in \mathcal{A}$, definieren. Dazu müssen wir überprüfen, ob die rechte Seite in diesem Fall definiert ist (d.h. ob die Integranden messbar sind), und ob

$$(\mu \otimes p)[A] \ := \ \int \left(\int I_A(x,y)\, p(x,dy) \right) \mu(dx)$$

eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$ definiert.

Für Produktmengen $A = B \times C$ ($B \in \mathcal{S}, C \in \mathscr{T}$) ist die Funktion $x \mapsto \int I_A(x,y) p(x,dy)$ nach Definition des stochastischen Kerns messbar. Da die Mengen $A \in \mathcal{A}$, für die diese Funktion messbar ist, ein Dynkinsystem bilden, folgt die Messbarkeit für alle $A \in \mathcal{A}$.

$\mu \otimes p$ ist eine Wahrscheinlichkeitsverteilung, denn einerseits folgt

$$(\mu \otimes p)[\Omega] = (\mu \otimes p)[S \times T] = \int \left(\int I_S(x) I_T(y) p(x,dy) \right) \mu(dx) = \mu[S] = 1$$

aus $\int_T p(x,dy) = p(x,T) = 1$; andererseits gilt für disjunkte Mengen A_i ($i \in \mathbb{N}$)

$$I_{\bigcup A_i} = \sum I_{A_i},$$

woraus unter zweimaliger Anwendung des Satzes von der monotonen Konvergenz folgt:

$$\begin{aligned}
(\mu \otimes p) \left[\bigcup_i A_i \right] &= \int \left(\int \sum_i I_{A_i}(x,y)\, p(x,dy) \right) \mu(dx) \\
&= \sum_i \int \left(\int I_{A_i}(x,y)\, p(x,dy) \right) \mu(dx) \\
&= \sum_i (\mu \otimes p)[A_i].
\end{aligned}$$

Durch maßtheoretische Induktion zeigt man nun, dass die Wahrscheinlichkeitsverteilung $\mu \otimes p$ auch (3.3.2) erfüllt.

□

Als nächstes wollen wir die **Randverteilungen** des gerade konstruierten zweistufigen Modells berechnen. Sei also $P := \mu \otimes p$, und seien

$$\begin{array}{rccc} X: & S \times T & \to & S \\ & (x,y) & \mapsto & x \end{array} \quad , \quad \begin{array}{rccc} Y: & S \times T & \to & T \\ & (x,y) & \mapsto & y \end{array},$$

die Projektionen auf die 1. bzw. 2. Komponente. Wegen $p(x,T) = 1$ gilt:

$$P[X \in B] = P[B \times T] = \int_B \mu(dx)\, p(x,T) = \mu[B] \qquad \forall\, B \in \mathcal{S},$$

also ist die Verteilung $P \circ X^{-1}$ der ersten Komponente gleich μ. Für die Verteilung der zweiten Komponente erhalten wir

$$P[Y \in C] = P[S \times C] = \int_S \mu(dx)\, p(x,C) \qquad \forall\, C \in \mathscr{T}.$$

Definition. *Die durch*

$$(\mu p)[C] \quad := \quad \int \mu(dx)\, p(x,C), \qquad C\}in\mathscr{T},$$

*definierte Wahrscheinlichkeitsverteilung auf $(T, \mathscr{T})$ heißt **Mischung** der Wahrscheinlichkeitsverteilungen $p(x,\bullet)$ bezüglich μ.*

Wie gerade gezeigt, ist $\mu p = P \circ Y^{-1}$ die Verteilung der zweiten Komponente im zweistufigen Modell.

Bemerkung. Sind S und T abzählbar, dann sind $\mu \otimes p$ und μp die schon in Abschnitt **??** betrachteten Wahrscheinlichkeitsverteilungen mit Gewichten

$$\begin{aligned} (\mu \otimes p)(x,y) &= \mu(x)\, p(x,y), \\ (\mu p)(y) &= \sum_{x \in S} \mu(x)\, p(x,y). \end{aligned}$$

Die Massenfunktionen von $\mu \otimes p$ und μp sind also das Tensor- bzw. Matrixprodukt des Zeilenvektors μ und der stochastischen Matrix p.

3.3.2 Wichtige Spezialfälle

Produktmaße: Ist $p(x,\bullet) \equiv \nu$ eine feste (von x unabhängige) Wahrscheinlichkeitsverteilung auf $(T, \mathscr{T})$, dann ist $\mu \otimes p$ das Produkt $\mu \otimes \nu$ der Wahrscheinlichkeitsverteilungen μ und ν. Der Satz von Fubini liefert also die Existenz des Produktmaßes, und die schon mehrfach verwendete Berechnungsformel

$$\int f\, d(\mu \otimes \nu) \;=\; \int_S \left(\int_T f(x,y)\, \nu(dy) \right) \mu(dx) \tag{3.3.3}$$

für die Integrale nicht-negativer oder integrierbarer messbarer Funktionen bzgl. des Produktmaßes. Die Integrationsreihenfolge kann man in diesem Fall vertauschen, denn wegen

$$(\mu \otimes \nu)[B \times C] \;=\; \mu[B]\nu[C] \qquad \text{für alle } B \in S,\ C \in \mathscr{T} \tag{3.3.4}$$

gilt $(\nu \otimes \mu) \circ R^{-1} = \mu \otimes \nu$, wobei $R(x,y) = (y,x)$, und damit nach dem Transformationssatz:

$$\begin{aligned} \int \left(\int f(x,y)\, \mu(dx) \right) \nu(dy) &\overset{Fub.}{=} \int f \circ R \;\, d(\nu \otimes \mu) \\ &= \int f \; d(\mu \otimes \nu) \\ &\overset{Fub.}{=} \int \left(\int f(x,y)\, \nu(dy) \right) \mu(dx). \end{aligned}$$

Durch wiederholte Anwendung dieses Arguments erhalten wir zudem:

Korollar 3.17. *Seien $(S_i, \mathcal{S}_i, \mu_i)$ Wahrscheinlichkeitsräume $(1 \le i \le n)$. Dann existiert eine eindeutige Wahrscheinlichkeitsverteilung $\mu_1 \otimes ... \otimes \mu_n$ auf $(S_1 \times ... \times S_n, \mathcal{S}_1 \otimes ... \otimes \mathcal{S}_n)$ mit:*

$$(\mu_1 \otimes ... \otimes \mu_n)\,[B_1 \times ... \times B_n] \;=\; \prod_{i=1}^{n} \mu_i[B_i] \qquad \textit{für alle } B_i \in \mathcal{S}_i \quad (1 \le i \le n).$$

Für alle produktmessbaren Funktionen $f : S_1 \times ... \times S_n \to [0, \infty)$ gilt:

$$\int f \; d(\mu_1 \otimes ... \otimes \mu_n) = \int ... \left(\int f(x_1, ..., x_n) \mu_n(dx_n) \right) ... \mu_1(dx_1),$$

wobei die Integration auch in beliebiger anderer Reihenfolge ausgeführt werden kann.

Beweis. Die Existenz folgt durch wiederholte Anwendung des Satzes von Fubini, die Eindeutigkeit aus dem Eindeutigkeitssatz. Dass die Integrationsreihenfolge vertauscht werden kann, zeigt man ähnlich wie im oben betrachteten Fall $n = 2$. □

Deterministische Kopplung: Gilt $p(x, \bullet) = \delta_{f(x)}$ für eine messbare Funktion $f : S \to T$, dann folgt $(\mu \otimes p)[\{(x,y) \mid y = f(x)\}] = 1$. Die zweite Komponente ist also durch die erste Komponente mit Wahrscheinlichkeit 1 eindeutig festgelegt. Die Verteilung der zweiten Komponente ist in diesem Fall das Bild von μ unter f:

$$\mu p = \mu \circ f^{-1}.$$

Übergangskerne von Markovschen Ketten: Gilt $S = T$, dann können wir $p(x, dy)$ als Übergangswahrscheinlichkeit (Bewegungsgesetz) einer Markovkette auf $(S, \mathcal{S})$ auffassen. In Analogie zum diskreten Fall definieren wir:

Definition. *Eine Wahrscheinlichkeitsverteilung μ auf $(S, \mathcal{S})$ heißt **Gleichgewicht (stationäre oder auch invariante Verteilung)** von p, falls $\mu p = \mu$ gilt, d.h. falls*

$$\int \mu(dx) p(x, B) = \mu[B] \qquad \textit{für alle } \; B \in \mathcal{S}.$$

Beispiel (Autoregressiver Prozess). Der AR(1)-Prozess mit Parametern $\varepsilon, \alpha \in \mathbb{R}$ ist eine Markovkette mit Übergangskern $p(x, \bullet) = N(\alpha x, \varepsilon^2)$. Die Normalverteilung $N\left(0, \frac{\varepsilon^2}{1-\alpha^2}\right)$ ist für $\alpha \in (0, 1)$ ein Gleichgewicht, siehe Lemma 5.16. Für $\alpha \geq 1$ existiert kein Gleichgewicht.

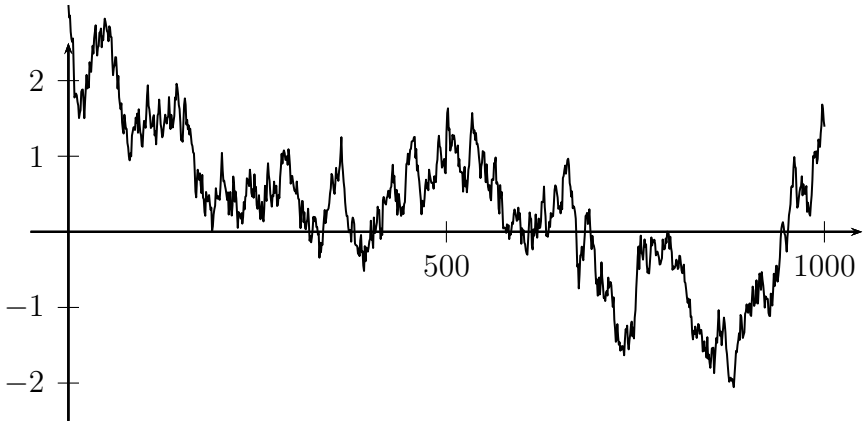


Abbildung 3.2: Simulation einer Trajektorie eines AR(1)-Prozesses mit Parametern $\alpha = 0.8$ und $\varepsilon^2 = 1.5$.

3.3.3 Bedingte Dichten und Bayessche Formel

Wir betrachten nun Situationen mit nichttrivialer Abhängigkeit zwischen den Komponenten im kontinuierlichen Fall. Seien $X : \Omega \to \mathbb{R}^n$ und $Y : \Omega \to \mathbb{R}^m$ Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, deren gemeinsame Verteilung absolutstetig ist mit Dichte $f_{X,Y}$, d.h.

$$P[x \in B, Y \in C] \quad = \quad \int_B \int_C f_{X,Y}(x, y)\, dy\, dx \qquad \text{für alle } B \in \mathcal{B}(\mathbb{R}^n), C \in \mathcal{B}(\mathbb{R}^m).$$

Nach dem Satz von Fubini sind dann auch die Verteilungen von X und Y absolutstetig mit Dichten

$$f_X(x) \quad = \quad \int_{\mathbb{R}^m} f_{X,Y}(x, y)\, dy, \qquad \text{und} \qquad f_Y(y) \quad = \quad \int_{\mathbb{R}^n} f_{X,Y}(x, y)\, dx.$$

Obwohl bedingte Wahrscheinlichkeiten gegeben $X = x$ nicht im herkömmlichen Sinn definiert werden können, da das Ereignis $\{X = x\}$ eine Nullmenge ist, können wir die bedingte Dichte und die bedingte Verteilung von Y gegeben X in diesem Fall sinnvoll definieren. Anschaulich beträgt die Wahrscheinlichkeit, dass der Wert von Y in einem infinitesimal kleinen Volumenelement

dy liegt, gegeben, dass der Wert von X in einem entsprechenden infinitesimalen Volumenelement dx liegt:

$$\begin{aligned} P[Y \in dy \,|\, X \in dx] &= \frac{P[X \in dx, Y \in dy]}{P[X \in dx]} = \frac{f_{X,Y}(x,y)\,dx\,dy}{f_X(x)\,dx} \\ &= \frac{f_{X,Y}(x,y)}{f_X(x)}\,dx \end{aligned}$$

Diese heuristische Überlegung motiviert die folgende Definition:

Definition. *Die Funktion $f_{Y|X} : \mathbb{R}^m \times \mathbb{R}^n \to [0,\infty]$ mit*

$$f_{Y|X}(y|x) = \begin{cases} \dfrac{f_{X,Y}(x,y)}{f_X(x)} & \text{falls } f_X(x) \neq 0 \\ f_Y(y) & \text{falls } f_X(x) = 0, \end{cases}$$

*heißt **bedingte Dichte von Y gegeben X**, und die von x abhängende Wahrscheinlichkeitsverteilung*

$$p(x,C) := \int_C f_{X|Y}(x,y)\,dy, \qquad \text{für } C \in \mathcal{B}(\mathbb{R}^m),$$

*heißt **bedingte Verteilung von Y gegeben X**.*

Bemerkung. (1). Für festes x ist die bedingte Dichte eine Wahrscheinlichkeitsdichte auf $\mathbb{R}^m$. Da $f_{Y|X}$ produktmeßbar ist, ist die bedingte Verteilung $p(x,dy)$ nach dem Satz von Fubini ein *stochastischer Kern* von $\mathbb{R}^n$ nach $\mathbb{R}^m$.

(2). Auf der Nullmenge $\{x \in \mathbb{R}^n : f_x(x) = 0\}$ sind die bedingte Dichte $f_{Y|X}(y|x)$ und die bedingte Verteilung von Y gegeben $X = x$ nicht eindeutig festgelegt - die oben getroffene Definition über die unbedingte Dichte ist relativ willkürlich.

Aus der Definition der bedingten Dichte ergibt sich unmittelbar eine Variante der Bayesschen Formel für absolutstetige Zufallsvariablen:

Satz 3.18 (Bayessche Formel). *Für $(x,y) \in \mathbb{R}^n \times \mathbb{R}^m$ mit $f_Y(y) > 0$ gilt*

$$f_{X|Y}(x|y) = \frac{f_X(x) f_{Y|X}(y|x)}{\int\limits_{\mathbb{R}^n} f_X(x) f_{Y|X}(y|x)\,dx}.$$

Beweis. Aus der Definition folgt

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{\int\limits_{\mathbb{R}^n} f_{X,Y}(x,y)\,dx},$$

und damit die Behauptung. □

In Modellen der Bayesschen Statistik interpretiert man $f_X(x)$ als Dichte der *a priori* angenommenen Verteilung eines unbekannten Parameters X, und $f_{Y|X}(y|x)$ als Maß für die Plausibilität („Likelihood“) des Parameterwertes x, wenn der Wert y der Zufallsgröße Y beobachtet wird. Die Bayessche Formel besagt dann, dass die Verteilung von X, von der man *a posteriori* (d.h. nach der Beobachtung von y) ausgeht, die Dichte

$$\begin{aligned} f_{X|Y}(x|y) &= \text{const.}(y) \cdot f_X(x) \cdot f_{Y|X}(y|x) \\ \text{A posteriori Dichte} &\propto \text{A priori Dichte} \times \text{Likelihood} \end{aligned}$$

hat. Trotz der einfachen Form der Bayesschen Formel ist es im Allgemeinen nicht trivial, Stichproben von der A-posteriori-Verteilung zu simulieren, und Erwartungswerte numerisch zu berechnen. Problematisch ist u.A., dass die Berechnung der Normierungskonstanten die Auswertung eines (häufig hochdimensionalen) Integrals erfordert. Ein wichtiges Verfahren zur Simulation von Stichproben in diesem Zusammenhang ist der Gibbs-Sampler.

Sind X und Y gemeinsam normalverteilt, dann kann man die wichtigsten Erwartungswerte bzgl. der A-posteriori-Verteilung im Prinzip exakt berechnen. Wir demonstrieren dies nun in einem grundlegenden Beispiel eines zweistufigen Modells. Ähnliche Modelle treten in zahlreichen Anwendungen auf.

Beispiel (**Signalverarbeitung**). Sei $S = T = \mathbb{R}^1$, also

$$S \times T = \mathbb{R}^2 = \{(x, y) \ : \ x, y \in \mathbb{R}\}.$$

Wir interpretieren die erste Komponente x als Größe eines nicht direkt beobachtbaren Signals, und die zweite Komponente y als verrauschte Beobachtung von x. In einem einfachen Bayesschen Modell nimmt man z.B. a priori an, dass Signal und Beobachtung normalverteilt sind:

$$\begin{aligned} \text{Signal} \quad x &\sim N(0, v)\,, \quad v > 0, \\ \text{Beobachtung} \quad y &\sim N(x, \varepsilon)\,, \quad \varepsilon > 0. \end{aligned}$$

Die Verteilung der ersten Komponente und der Übergangskern zur zweiten Komponente sind

$$\begin{aligned} \mu(dx) &= f_X(x)\ \lambda(dx), \\ p(x, dy) &= f_{Y|X}(y|x)\ \lambda(dy) \end{aligned}$$

mit den Dichten

$$\begin{aligned} f_X(x) &:= \frac{1}{\sqrt{2\pi v}}\, e^{-\frac{x^2}{2v}} && \text{(Dichte der Verteilung der ersten Komponente } X\text{)}, \\ f_{Y|X}(y|x) &:= \frac{1}{\sqrt{2\pi\varepsilon}}\, e^{-\frac{(y-x)^2}{2\varepsilon}} && \text{(bedingte Dichte der zweiten Komponente } Y \text{ gegeben } X = x\text{)}. \end{aligned}$$

Die gemeinsame Verteilung $\mu \otimes p$ von Signal und Beobachtungswert ist dann eine bivariate Normalverteilung (siehe Abschnitt 3.4) mit Dichte

$$\begin{aligned} f_{X,Y}(x,y) &= f_X(x)\, f_{Y|X}(y|x) \\ &= \frac{1}{2\pi\sqrt{v\varepsilon}}\, \exp\left(-\frac{(\varepsilon+v)x^2 - 2vxy + vy^2}{2v\varepsilon}\right) \\ &= \frac{1}{2\pi\sqrt{\det C}}\, \exp\left(-\frac{1}{2}\begin{pmatrix}x\\y\end{pmatrix}\cdot C^{-1}\begin{pmatrix}x\\y\end{pmatrix}\right). \end{aligned}$$

bzgl. des zweidimensionalen Lebesgue-Maßes. Hierbei ist

$$C = \begin{pmatrix} v & v \\ v & v+\varepsilon \end{pmatrix}$$

die Kovarianzmatrix der Verteilung, siehe Abschnitt 3.5. Als Dichte der Verteilung μp von Y ergibt sich

$$f_Y(y) = \int f_{X,Y}(x,y)\, dx.$$

Nach der Bayesschen Formel erhalten wir für die A-posteriori-Dichte des Signals gegeben die Beobachtung y:

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x) f_{Y|X}(y|x)}{\int f_X(x) f_{Y|X}(y|x)\, \lambda(dx)} \\ &= const(y) \cdot \exp\left(-\frac{\varepsilon+v}{2v\varepsilon}\, \left(x - \frac{v}{v+\varepsilon}y\right)^2\right). \end{aligned} \tag{3.3.5}$$

Die bedingte Verteilung des Signals gegeben die Beobachtung ist also $N(\hat{x}, u)$, wobei

$$\begin{aligned} \hat{x} &= \frac{v}{v+\varepsilon}\, y && \text{der Prognosewert ist, und} \\ u &= \frac{v\varepsilon}{v+\varepsilon} = \left(\frac{1}{v} + \frac{1}{\varepsilon}\right)^{-1} && \text{die Varianz der Prognose.} \end{aligned}$$

In einem Bayesschen Modell würden wir also nach der Beobachtung mit einer Standardabweichung $\sigma = \sqrt{u}$ prognostizieren, dass der Signalwert gleich $\hat{x}$ ist.

Ähnliche Modellierungsansätze werden auch in viel allgemeinerem Kontext verwendet. Beispielsweise wird in stochastischen Filterproblemen das Signal durch eine Markovkette (oder einen zeitstetigen Markovprozess) beschrieben, und die Folge der Beobachtungen durch einen von der Markovkette angetriebenen stochastischen Prozess. Sind alle gemeinsamen Verteilungen Gaußsch, dann kann man auch hier die a posteriori Verteilung im Prinzip exakt berechnen – andernfalls muss man auf numerische Näherungsmethoden (z.B. Partikelfilter) zurückgreifen.

3.4 Transformationen von mehreren Zufallsvariablen

3.4.1 Transformation von mehrdimensionalen Dichten

Seien $S, T \subseteq \mathbb{R}^n$ offen, und sei $X : \Omega \to S$ eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit absolutstetiger Verteilung μ_X mit Dichte f_X.

Satz 3.19 (Mehrdimensionaler Dichtetransformationssatz). *Ist $\phi : S \to T$ ein Diffeomorphismus (C^1) mit $\det D\phi(x) \neq 0$ für alle $x \in S$, dann ist die Verteilung von $\phi(X)$ absolutstetig mit Dichte*

$$f_{\phi(X)}(y) = f_X(\phi^{-1}(y)) \cdot |\det D\phi^{-1}(y)|,$$

wobei $\det D\phi^{-1}(y) = \det(\frac{\partial x_i}{\partial y_j})$ die Jacobideterminante der Koordinatentransformation ist.

Beweis. Die Behauptung folgt aus dem Transformationssatz der multivariaten Analysis:

$$\begin{aligned} P[\phi(X) \in B] &= P[X \in \phi^{-1}(B)] \\ &= \int_{\phi^{-1}(B)} f_X(x)\, dx \overset{Subst.}{=} \int_B f_X(\phi^{-1}(y)) \cdot |\det D\phi^{-1}(y)|\, dy\,. \end{aligned}$$

□

Beispiel (Sukzessive Wartezeiten). Seien T und $\widetilde{T}$ unabhängige, zum Parameter $\lambda > 0$ exponentialverteilte Zufallsvariablen (z.B. sukzessive Wartezeiten), und sei $S = T + \widetilde{T}$. Nach dem Dichtetransformationssatz gilt dann

$$\begin{aligned} f_{T,S}(t,s) &= f_{T,\widetilde{T}}(t, s-t) \cdot |\det \frac{\partial(t, s-t)}{\partial(t,s)}| \\ &\propto e^{-\lambda t} \cdot I_{(0,\infty)}(t) \cdot e^{-\lambda(s-t)} \cdot I_{(0,\infty)}(s-t) \\ &= e^{-\lambda s} \cdot I_{(0,s)}(t). \end{aligned}$$

Somit ist die bedingte Dichte $f_{S|T}(s|t)$ für festes $t > 0$ proportional zu $e^{-\lambda s} \cdot I_{(t,\infty)}(s)$. Dies ist auch anschaulich sofort plausibel, da s eine um die unabhängige Zufallsvariable T verschobene exponentialverteilte Zufallsvariable ist.
Interessanter ist die Berechnung der bedingten Dichte von T gegeben S: Für festes $s > 0$ ist $f_{T|S}(t|s)$ proportional zu $I_{(0,s)}(t)$, d.h.

$$f_{T|S}(t|s) = \frac{1}{s} \cdot I_{(0,s)}(t).$$

Gegeben die Summe S der beiden Wartezeiten ist die erste Wartezeit T also gleichverteilt auf $[0, S]$!

3.4.2 Multivariate Normalverteilungen

Sei $Z = (Z_1, Z_2, ..., Z_d)$ mit unabhängigen, standardnormalverteilten Zufallsvariablen Z_i. Die Verteilung des Zufallsvektors Z ist dann absolutstetig bzgl. des Lebesguemaßes im $\mathbb{R}^d$ mit Dichte

$$f_Z(x) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x_i^2}{2}} = (2\pi)^{-\frac{d}{2}} e^{-\frac{|x|^2}{2}} \quad \textit{(d-dimensionale Standardnormalverteilung).}$$

Sei nun $m \in \mathbb{R}^d$ und $\sigma \in \mathbb{R}^{d\times d}$ eine $d \times d$-Matrix. Wir betrachten den Zufallsvektor

$$Y = \sigma Z + m\ .$$

Ist σ *regulär*, dann können wir die Dichte der Verteilung von Y bzgl. des Lebesgue-Maßes im $\mathbb{R}^d$ mithilfe des Transformationssatzes explizit berechnen:

$$\begin{aligned} f_Y(y) &= f_X(\sigma^{-1}(y-m)) \cdot |\det \sigma^{-1}| \\ &= \frac{1}{\sqrt{(2\pi)^d |\det C|}} \exp\left(-\frac{1}{2}(y-m)\cdot C^{-1}(y-m)\right)\ . \end{aligned}$$

Ist σ nicht regulär, dann nimmt X nur Werte in einem echten Unterraum des $\mathbb{R}^d$ an. Die Verteilung von X ist deshalb *nicht absolutstetig* bzgl. des Lebesgue-Maßes im $\mathbb{R}^d$.

Definition (Normalverteilung im $\mathbb{R}^d$). *Sei $m \in \mathbb{R}^d$, und sei $C \in \mathbb{R}^{d\times d}$ eine symmetrische, positiv definite Matrix. Die Verteilung $N(m, C)$ im $\mathbb{R}^d$ mit Dichte f_Y heißt d-**dimensionale Normalverteilung** mit Mittel m und Kovarianzmatrix C.*

Beispiel (Zufällige Punkte in der Ebene). Seien X und Y unabhängige, $N(0, \sigma^2)$-verteilte Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ mit $\sigma > 0$. Dann ist die gemeinsame Verteilung $\mu_{X,Y}$ absolutstetig mit Dichte

$$f_{X,Y}(x,y) \quad = \quad \frac{1}{2\pi\sigma^2} \cdot \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right), \qquad (x,y) \in \mathbb{R}^2.$$

Insbesondere gilt $(X, Y) \neq (0, 0)$ P-fast sicher. Wir definieren den Radial- und Polaranteil

$$R\ :\ \Omega \to (0,\infty), \qquad \Phi\ :\ \Omega \to [0, 2\pi)$$

durch

$$X \quad = \quad R \cdot \cos\Phi \qquad \text{und} \qquad Y \quad = \quad R \cdot \sin\Phi,$$

d.h. $R = \sqrt{X^2 + Y^2}$ und $\Phi = \arg(X + iY)$ falls $(X, Y) \neq (0, 0)$. Auf der Nullmenge $\{(X, Y) = (0, 0)\}$ definieren wir (R, Φ) in beliebiger Weise, sodass sich messbare Funktionen ergeben. Wir berechnen nun die gemeinsame Verteilung von R und Φ:

$$\begin{aligned} P[R \le r_0, \Phi \le \phi_0] &= P[(X, Y) \in \text{„Kuchenstück“ mit Winkel } \phi_0 \text{ und Radius } r_0] \\ &= \int\int_{\text{Kuchenstück}} f_{X,Y}(x, y)\, dx\, dy \\ &= \int_0^{r_0}\int_0^{\phi_0} f_{X,Y}(r\cos\phi, r\sin\phi) \underbrace{r}_{\substack{\text{Jacobideterminante} \\ \text{der Koordinatentrans. } f}} d\phi\, dr \\ &= \int_0^{r_0}\int_0^{\phi_0} \frac{r}{2\pi\sigma^2} e^{-r^2/(2\sigma^2)}\, d\phi\, dr. \end{aligned}$$

Hierbei haben wir im 3. Schritt den Transformationssatz (Substitutionsregel) für mehrdimensionale Integrale verwendet - der Faktor r ist die Jacobideterminante der Koordinatentransformation. Es folgt, dass die gemeinsame Verteilung $\mu_{R,\Phi}$ absolutstetig ist mit Dichte

$$f_{R,\Phi}(r, \phi) = \frac{1}{2\pi} \cdot \frac{r}{\sigma^2} \cdot e^{-r^2/(2\sigma^2)}.$$

Da die Dichte Produktform hat, sind R und Φ unabhängig. Die Randverteilung μ_Φ ist absolutstetig mit Dichte

$$f_\Phi(\phi) = \text{const.} = \frac{1}{2\pi} \qquad (0 \le \phi < 2\pi),$$

d.h. Φ ist gleichverteilt auf $[0, 2\pi)$. Somit ist μ_R absolutstetig mit Dichte

$$\phi_R(r) = \frac{r}{\sigma^2} \cdot e^{-r^2/(2\sigma^2)} \qquad (r > 0).$$

Die Berechnung können wir verwenden, um Stichproben von der Standardnormalverteilung zu simulieren:

Beispiel (Simulation von normalverteilten Zufallsvariablen). Die Verteilungsfunktion einer $N(0, 1)$-verteilten Zufallsvariable X ist

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt\ .$$

Das Integral ist nicht explizit lösbar und die Inverse F_X^{-1} ist dementsprechend nur approximativ berechenbar. Daher ist die Simulation einer Standardnormalverteilung durch Inversion der Verteilungsfunktion relativ aufwendig. Ein einfacheres Simulationsverfahren ergibt sich, wenn wir

eine zweidimensionale Standardnormalverteilung betrachten und auf Polarkoordinaten transformieren. Dann gilt für den Radialanteil:

$$F_R(x) \;=\; \int_0^x e^{-r^2/2} r \, dr \;=\; 1 - e^{-x^2/2}.$$

Das Integral ist also explizit berechenbar, und

$$F_R^{-1}(u) \;=\; \sqrt{-2\log(1-u)}\,, \quad u \in (0,1).$$

Der Winkelanteil Φ ist unabhängig von R und gleichverteilt auf $[0, 2\pi)$. Wir können Zufallsvariablen mit der entsprechenden gemeinsamen Verteilung erzeugen, indem wir

$$\begin{aligned} \Phi &:= 2\pi U_1\,, \\ R &:= \sqrt{-2\log(1-U_2)} \qquad \Big(\text{bzw. } = \sqrt{-2\log U_2}\Big)\,, \end{aligned}$$

setzen, wobei U_1 und U_2 unabhängige, auf $(0,1)$ gleichverteilte Zufallsvariablen sind. Stichproben von U_1 und U_2 können durch Pseudozufallszahlen simuliert werden. die Zufallsvariablen

$$X \quad := \quad R\cos\Phi \qquad \text{und} \qquad Y \quad := \quad R \cdot \sin\Phi$$

sind dann unabhängig und $N(0,1)$-verteilt. Für $m \in \mathbb{R}$ und $\sigma > 0$ sind $\sigma X + m$ und $\sigma Y + m$ unabhängige $N(m, \sigma^2)$-verteilte Zufallsvariable.

Wir erhalten also den folgenden Algorithmus zur Simulation von Stichproben einer Normalverteilung:

Algorithmus 3.20 (Box-Muller-Verfahren). Input: $m \in \mathbb{R}, \sigma > 0$

Output: unabhängige Stichproben $\widetilde{x}, \widetilde{y}$ von $N(m, \sigma^2)$.

1. Erzeuge unabhängige Zufallszahlen $u_1, u_2 \sim \mathcal{U}_{(0,1)}$

2. $x := \sqrt{-2\log u_1}\cos(2\pi u_2)$, $y := \sqrt{-2\log u_1}\sin(2\pi u_2)$

3. $\widetilde{x} := \sigma x + m$, $\widetilde{y} = \sigma y + m$

3.4.3 Summen unabhängiger Zufallsvariablen, Faltung

Seien X und Y unabhängige reellwertige Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ mit Verteilungen μ bzw. ν. Wir wollen die Verteilung von $X + Y$ bestimmen. Für diskrete Zufallsvariablen ergibt sich:

$$P[X+Y=z] \;=\; \sum_{x \in X(\Omega)} \underbrace{P[X=x, Y=z-x]}_{=P[X=x]\cdot P[Y=z-x]} = \sum_{x \in X(\Omega)} \mu(x)\nu(z-x) \qquad (3.4.1)$$

Die Wahrscheinlichkeitsverteilung mit Massenfunktion

$$(\mu \star \nu)(z) \quad = \quad \sum_{x \in X(\Omega)} \mu(x)\nu(z-x)$$

heißt Faltung von μ und ν. Eine entsprechende Aussage erhält man auch im allgemeinen Fall:

Satz 3.21 (Verteilungen von Summen unabhängiger Zufallsvariablen). *Seien X und Y unabhängige reellwertige Zufallsvariablen mit Verteilungen μ bzw. ν. Dann ist die Verteilung von $X+Y$ die durch*

$$(\mu \star \nu)[B] \quad := \quad \int \mu(dx)\, \nu[B-x]\,, \qquad B \in \mathcal{B}(\mathbb{R}),$$

*definierte **Faltung** der Wahrscheinlichkeitsverteilungen μ und ν.*

Beweis. Sei $\widetilde{B} := \{(x,y) \mid x+y \in B\}$. Da X und Y unabhängig sind, erhalten wir mit dem Satz von Fubini

$$\begin{aligned} P[X+Y \in B] \quad &= \quad P[(X,Y) \in \widetilde{B}] \quad = \quad (\mu \otimes \nu)[\widetilde{B}] \\ &\overset{Fubini}{=} \int \mu(dx) \int \nu(dy) \underbrace{I_B(x+y)}_{=I_{B-x}(y)} \quad = \quad \int \mu(dx)\, \nu[B-x]. \end{aligned}$$

□

Bemerkung. Die Faltung $\mu \star \nu$ zweier Wahrscheinlichkeitsverteilungen μ und ν auf $\mathbb{R}^1$ ist wieder eine Wahrscheinlichkeitsverteilung auf $\mathbb{R}^1$. Da die Addition von Zufallsvariablen kommutativ und assoziativ ist, hat die Faltung von Wahrscheinlichkeitsverteilungen nach Satz 3.21 dieselben Eigenschaften:

$$\mu \star \nu \quad = \quad \nu \star \mu \qquad \text{(da } X+Y = Y+X\text{)} \tag{3.4.2}$$

$$(\mu \star \nu) \star \eta \quad = \quad \mu \star (\nu \star \eta) \qquad \text{(da } (X+Y)+Z = X+(Y+Z)\text{).} \tag{3.4.3}$$

Im diskreten Fall ist $\mu \star \nu$ nach (3.4.2) die Wahrscheinlichkeitsverteilung mit Gewichten

$$(\mu \star \nu)(z) \quad = \quad \sum_x \mu(x)\nu(z-x).$$

Eine entsprechende Berechnungsformel ergibt sich auch für absolutstetige Wahrscheinlichkeitsverteilungen:

Lemma 3.22. *Ist ν absolutstetig mit Dichte g, dann ist auch $\mu \star \nu$ absolutstetig mit Dichte*

$$\varrho(z) \;=\; \int \mu(dx)\; g(z-x).$$

Ist zusätzlich auch μ absolutstetig mit Dichte f, dann gilt

$$\varrho(z) \;=\; \int_{\mathbb{R}} f(x)\; g(z-x)\; dx \;=:\; (f \star g)(z)$$

Beweis. Wegen der Translationsinvarianz des Lebesguemaßes gilt

$$(\mu \star \nu)[B] = \int \mu(dx)\nu[B-x] = \int \mu(dx) \underbrace{\int_{B-x} g(y)dy}_{=\int_B g(z-x)dz} \stackrel{Fub.}{=} \int_B \left(\int \mu(dx) g(z-x) \right) dz \; .$$

Also ist $\mu \star \nu$ absolutstetig mit Dichte ϱ. Die zweite Behauptung folgt unmittelbar. □

Beispiel. (1). Sind X und Y unabhängig, und $\mathrm{Bin}(n,p)$ bzw. $\mathrm{Bin}(m,p)$-verteilt, dann ist $X+Y$ eine $\mathrm{Bin}(n+m,p)$-verteilte Zufallsvariable. Zum Beweis bemerkt man, dass die gemeinsame Verteilung von X und Y mit der gemeinsamen Verteilung von $Z_1 + ... + Z_n$ und $Z_{n+1} + ... + Z_{n+m}$ übereinstimmt, wobei die Zufallsvariablen Z_i $(1 \leq i \leq n+m)$ unabhängig und $\mathrm{Bernoulli}(p)$-verteilt sind. Also folgt:

$$\mu_{X+Y} = \mu_{Z_1+...+Z_n+Z_{n+1}+...+Z_{n+m}} = \mathrm{Bin}(n+m,p) \; .$$

Als Konsequenz erhalten wir (ohne zu rechnen):

$$\mathrm{Bin}(n,p) \star \mathrm{Bin}(m,p) \; = \; \mathrm{Bin}(n+m,p) \; ,$$

d.h. die Binomialverteilungen bilden eine *Faltungshalbgruppe*. Explizit ergibt sich:

$$\sum_{k=0}^{l} \binom{n}{k} p^k (1-p)^{n-k} \binom{m}{l-k} p^{l-k} (1-p)^{m-(l-k)} = \binom{n+m}{l} p^l (1-p)^{n+m-l} \; ,$$

d.h.

$$\sum_{k=0}^{l} \binom{n}{k} \binom{m}{l-k} = \binom{n+m}{l}. \tag{3.4.4}$$

Die kombinatorische Formel (3.4.4) ist auch als *Vandermonde-Identität* bekannt.

(2). Sind X und Y unabhängig und Poisson-verteilt mit Parametern λ bzw. $\widetilde{\lambda}$, dann ist $X+Y$ Poisson-verteilt mit Parameter $\lambda + \widetilde{\lambda}$, denn nach der Binomischen Formel gilt für $n \geq 0$:

$$\begin{aligned}
(\mu_X \star \mu_Y)(n) &= \sum_{k=0}^{n} \mu_X(k) \cdot \mu_Y(n-k) \\
&= \sum_{k=0}^{n} \frac{\lambda^k}{k!} e^{-\lambda} \cdot \frac{\widetilde{\lambda}^{n-k}}{(n-k)!} e^{-\widetilde{\lambda}} \\
&= e^{-\lambda+\widetilde{\lambda}} \cdot \sum_{k=0}^{n} \frac{\lambda^k}{k!} \frac{\widetilde{\lambda}^{n-k}}{(n-k)!} \\
&= e^{-\lambda+\widetilde{\lambda}} \cdot \frac{(\lambda+\widetilde{\lambda})^n}{n!} \; .
\end{aligned}$$

Also bilden auch die Poissonverteilungen eine Faltungshalbgruppe:

$$\mathrm{Poisson}(\lambda) \; \star \; \mathrm{Poisson}(\widetilde{\lambda}) \; = \; \mathrm{Poisson}(\lambda + \widetilde{\lambda})$$

(3). Sind X und Y unabhängig und normalverteilt mit Parametern (m, σ^2) bzw. $(\widetilde{m}, \widetilde{\sigma}^2)$, dann ist $X + Y$ normalverteilt mit Parametern $(m + \widetilde{m}, \sigma^2 + \widetilde{\sigma}^2)$, siehe **??**. Dies verifiziert man leicht mithilfe der charakteristischen Funktionen. Die Normalverteilungen bilden also eine zweiparametrige Faltungshalbgruppe.

3.4.4 Wartezeiten, Gamma-Verteilung

Seien $T_1, T_2, ...$ sukzessive Wartezeiten auf das Eintreten eines unvorhersehbaren Ereignisses. In einem einfachen Modell nehmen wir an, dass die T_i $(i \in \mathbb{N})$ unabhängige exponentialverteilte Zufallsvariablen sind, d.h. die Verteilungen der T_i sind absolutstetig mit Dichte

$$f(t) \;=\; \lambda \cdot e^{-\lambda t} \cdot I_{(0,\infty)}(t) \; .$$

Die Verteilung der Gesamtwartezeit

$$S_n = T_1 + ... + T_n$$

bis zum n-ten Ereignis ist dann

$$\mu_{S_n} = \mu_{T_1} \star \mu_{T_2} \star ... \star \mu_{T_n} \; .$$

Insbesondere ist die Verteilung von S_2 absolutstetig mit Dichte

$$(f \star f)(s) = \int_{\mathbb{R}} \underbrace{f(x)}_{\substack{=0 \\ für\ x<0}} \underbrace{f(s-x)}_{\substack{=0 \\ für\ x>s}} = \int_0^s \lambda^2 e^{-\lambda x} e^{-\lambda(s-x)} dx = \lambda^2 e^{-\lambda s} \int_0^s dx = \lambda^2 s e^{-\lambda s}$$

für $s \geq 0$, bzw. $(f \star f)(s) = 0$ für $s < 0$. Durch Induktion ergibt sich allgemein:

Lemma 3.23. *Die Verteilung von S_n ist absolutstetig mit Dichte*

$$f_{\lambda,n}(s) = \frac{\lambda^n}{\Gamma(n)} \cdot s^{n-1} \cdot e^{-\lambda s} \cdot I_{(0,\infty)}(s) \; ,$$

wobei

$$\Gamma(n) \;:=\; \int_0^\infty t^{n-1}\, e^{-t}\, dx \;\overset{n\in\mathbb{N}}{=}\; (n-1)! \quad .$$

Definition. *Die Wahrscheinlichkeitsverteilung auf $\mathbb{R}_+$ mit Dichte $f_{\lambda,n}$ heißt **Gammaverteilung** mit Parametern $\lambda, n \in (0, \infty)$.*

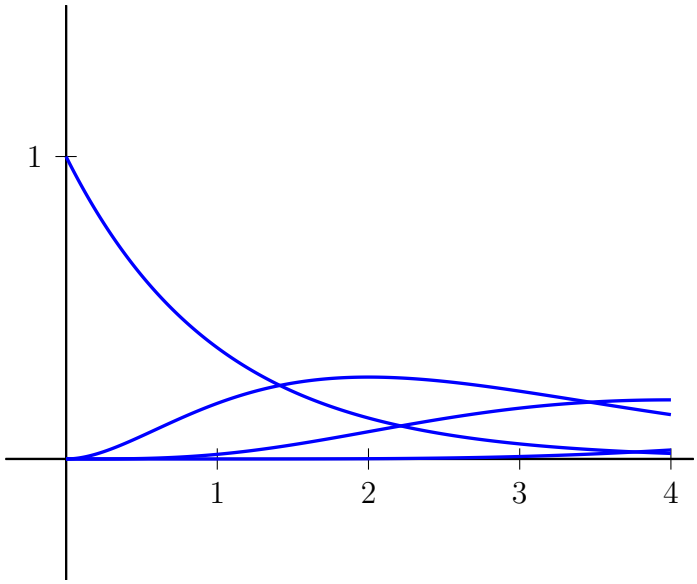


Abbildung 3.3: Dichtefunktionen der Gammaverteilung $\Gamma_{1,n}$ für verschiedene n.

Die Gammaverteilung ist auch für nicht-ganzzahlige n definiert, Γ ist dann die Eulersche Gammafunktion. Für $n = 1$ ergibt sich die Exponentialverteilung als Spezialfall der Gammaverteilung. Allgemein gilt:

$$\Gamma(\lambda, r) \star \Gamma(\lambda, s) \quad = \quad \Gamma(\lambda, r+s) \,,$$

d.h. die Gammaverteilungen mit festem Parameter λ bilden eine Faltungshalbgruppe.

Bemerkung (Poissonprozess). Die Anzahl der bis zur Zeit $t \geq 0$ eingetretenen Ereignisse im obigen Modell ist

$$N_t \;=\; \max\{n \geq 0 \mid S_n \leq t\} \,.$$

Die Zufallsvariablen N_t sind Poissonverteilt mit Parameter $\lambda \cdot t$ (Übung). Die Kollektion N_t $(t \geq 0)$ der Zufallsvariablen heißt ***Poissonprozess mit Intensität*** λ. Der Poissonprozess ist ein monoton wachsender stochastischer Prozess mit ganzzahligen Werten. Er ist selbst eine zeitstetige Markovkette und ist von grundlegender Bedeutung für die Konstruktion allgemeiner Markovketten in kontinuierlicher Zeit. Wir werden den Poissonprozess in der Vorlesung „Stochastische Prozesse“ genauer betrachten.

3.5 Kovarianz und lineare Prognosen

Für einen gegebenen Wahrscheinlichkeitsraum und $p \in [1, \infty)$ bezeichnen wir mit $\mathcal{L}^p(\Omega, \mathcal{A}, P)$ den Raum aller Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ mit endlichem p-tem Moment:

$$\mathcal{L}^p(\Omega, \mathcal{A}, P) \quad = \quad \{X : \Omega \to \overline{\mathbb{R}} \text{ messbar } : E[|X|^p] < \infty\}.$$

Der Raum ist ein Unterraum des Vektorraums aller $\mathcal{A}/\mathcal{B}(\overline{\mathbb{R}})$ messbaren Abbildungen, denn für $X, Y \in \mathcal{L}^p(\Omega, \mathcal{A}, P)$ und $a \in \mathbb{R}$ gilt mit einer endlichen Konstante c_p:

$$E[|aX + Y|^p] \quad \leq \quad E[c_p(|aX|^p + |Y|^p)] \quad = \quad c_p|a|^p E[|X|^p] + c_p E[|Y|^p] \quad < \quad \infty.$$

Auf dem Vektorraum

$$L^p(\Omega, \mathcal{A}, P) \quad = \quad \mathcal{L}^p(\Omega, \mathcal{A}, P)/\sim$$

der Äquivalenzklassen von P-fast sicher gleichen Zufallsvariablen in $\mathcal{L}^p(\Omega, \mathcal{A}, P)$ wird durch

$$\|X\|_{L^p} \quad := \quad E\left[|X|^p\right]^{1/p}$$

eine Norm definiert. In der Analysis wird gezeigt, dass $L^p(\Omega, \mathcal{A}, P)$ ein Banachraum, also vollständig bzgl. der L^p-Norm ist.

Wir beschränken uns hier zunächst auf die Fälle $p = 1$ und $p = 2$. Der Vektorraum $L^2(\Omega, \mathcal{A}, P)$ hat den großen Vorteil, dass die L^2-Norm von dem Skalarprodukt

$$(X, Y)_{L^2} \quad := \quad E[XY]$$

erzeugt wird. Hierbei ist der Erwartungswert $E[XY]$ wegen $|XY| \leq (X^2 + Y^2)/2$ definiert. Insbesondere gilt die **Cauchy-Schwarz-Ungleichung**

$$|E[XY]| \quad \leq \quad E[X^2]^{1/2} \cdot E[Y^2]^{1/2} \qquad \text{für alle } X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P).$$

Es folgt

$$\mathcal{L}^2(\Omega, \mathcal{A}, P) \quad \subseteq \quad \mathcal{L}^1(\Omega, \mathcal{A}, P),$$

denn für alle $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ gilt

$$E[|X|] \quad \leq \quad E[1^2]^{1/2} \cdot E[X^2]^{1/2} \quad < \quad \infty.$$

Dabei haben wir wesentlich benutzt, dass P ein endliches Maß ist - für unendliche Maße ist der Raum $\mathcal{L}^2$ *nicht* in $\mathcal{L}^1$ enthalten ! Nach (3.2.2) ist umgekehrt eine Zufallsvariable aus $\mathcal{L}^1$ genau dann in $\mathcal{L}^2$ enthalten, wenn sie endliche Varianz hat.

Seien nun X und Y quadratintegrierbare reellwertige Zufallsvariablen, die auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definiert sind. Wie schon für diskrete Zufallsvariablen definieren wir wieder die **Kovarianz**

$$\mathrm{Cov}[X,Y] \quad := \quad E[(X - E[X])(Y - E[Y])] \quad = \quad E[XY] - E[X] \cdot E[Y],$$

und den **Korrelationskoeffizienten**

$$\varrho[X,Y] \quad := \quad \frac{\mathrm{Cov}[X,Y]}{\sigma[X]\sigma[Y]}, \qquad \text{falls } \sigma[X] \cdot \sigma[Y] \neq 0.$$

Die Zufallsvariablen X und Y heißen **unkorreliert**, falls $\mathrm{Cov}[X,Y] = 0$ gilt, d.h. falls

$$E[XY] \quad = \quad E[X] \cdot E[Y].$$

Aus dem Transformationssatz für den Erwartungswert folgt, dass wir die Kovarianz $\mathrm{Cov}[X,Y]$ aus der gemeinsamen Verteilung $\mu_{X,Y} = P \circ (X,Y)^{-1}$ der Zufallsvariablen X und Y berechnen können:

$$\mathrm{Cov}[X,Y] \quad = \quad \int \left(x - \int z \, \mu_X(dz) \right) \left(y - \int z \, \mu_Y(dz) \right) \mu_{X,Y}(dx\, dy).$$

Aus der Linearität des Erwartungswertes ergibt sich, dass die Abbildung $\mathrm{Cov} : \mathcal{L}^2 \times \mathcal{L}^2 \to \mathbb{R}$ symmetrisch und bilinear ist. Die Varianz $\mathrm{Var}[X] = \mathrm{Cov}[X,X]$ ist die zugehörige quadratische Form. Insbesondere gilt wie im diskreten Fall:

$$\mathrm{Var}\left[\sum_{i=1}^n X_i\right] \quad = \quad \sum_{i=1}^n \mathrm{Var}[X_i] + 2 \cdot \sum_{\substack{i,j=1 \\ i<j}}^n \mathrm{Cov}[X_i, X_j].$$

Sind die Zufallsvariablen $X_1, \ldots, X_n$ unkorreliert, dann folgt:

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

3.5.1 Lineare Prognosen

Angenommen wir wollen den Ausgang eines Zufallsexperiments vorhersagen, dass durch eine reellwertige Zufallsvariable $Y : \Omega \to \mathbb{R}$ beschrieben wird. Welches ist der *beste Prognosewert* b für $Y(\omega)$, wenn uns keine weiteren Informationen zur Verfügung stehen?

Die Antwort hängt offensichtlich davon ab, wie wir den Prognosefehler messen. Häufig verwendet man den mittleren quadratischen Fehler (**M**ean **S**quare **E**rror)

$$\mathrm{MSE} \quad = \quad E[(X-a)^2]$$

bzw. die Wurzel (**R**oot **M**ean **S**quare **E**rror)

$$\text{RMSE} \quad = \quad \text{MSE}^{1/2} \quad = \quad \|X - a\|_{L^2(\Omega,\mathcal{A},P)}.$$

Eine andere Möglichkeit ist die Verwendung der L^1- statt der L^2-Norm.

Satz 3.24 (Erwartungswert und Median als beste Prognosewerte).

(1). Ist $Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$, dann gilt für alle $b \in \mathbb{R}$:

$$E[(Y-b)^2] \quad = \quad \text{Var}[Y] + (b - E[Y])^2 \quad \geq \quad E[(Y - E[Y])^2].$$

(2). Ist $Y \in L^1(\Omega, \mathcal{A}, P)$ und m ein Median der Verteilung von Y, dann gilt für alle $b \in \mathbb{R}$:

$$E[|Y-b|] \quad \geq \quad E[|Y-m|].$$

Der mittlere quadratische Fehler des Prognosewertes b ist also die Summe der Varianz von X und des Quadrats des systematischen bzw. mittleren Prognosefehlers (engl. *Bias*) $b - E[X]$:

$$\text{MSE} \quad = \quad \text{Varianz} \; + \; \text{Bias}^2.$$

Insbesondere ist der mittlere quadratische Fehler genau für $b = E[X]$ minimal. Der L^1-Fehler ist dagegen bei einem Median minimal.

Beweis. (1). Für $b \in \mathbb{R}$ gilt wegen der Linearität des Erwartungswertes:

$$\begin{aligned} E[(Y-b)^2] &= E[(Y - E[Y] + E[Y] - b)^2] \\ &= E[(Y-E[Y])^2] + 2E[(Y-E[Y]) \cdot (E[Y]-b)] + E[(E[Y]-b)^2] \\ &= \text{Var}[Y] + (E[Y]-b)^2. \end{aligned}$$

(2). Für $m \geq b$ folgt die Behauptung aus der Identität

$$|Y-m| - |Y-b| \quad \leq \quad (m-b) \cdot \big(I_{(-\infty,m)}(Y) - I_{[m,\infty)}(Y)\big)$$

durch Bilden des Erwartungswertes, denn für einen Median m gilt $P[Y < m] \leq 1/2$ und $P[Y \geq m] \geq 1/2$. Der Beweis für $m \leq b$ verläuft analog.

□

Seien nun $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ quadratintegrierbare Zufallsvariablen mit $\sigma[X] \neq 0$. Angenommen, wir kennen bereits den Wert $X(\omega)$ in einem Zufallsexperiment und suchen die beste *lineare* Vorhersage

$$\hat{Y}(\omega) \quad = \quad aX(\omega) + b, \qquad (a, b \in \mathbb{R}) \tag{3.5.1}$$

für $Y(\omega)$ im quadratischen Mittel. Zu minimieren ist jetzt der mittlere quadratischen Fehler

$$\text{MSE} \quad := \quad E[(\hat{Y} - Y)^2],$$

unter allen Zufallsvariablen $\hat{Y}$, die affine Funktionen von X sind. In diesem Fall erhalten wir

$$\begin{aligned} \text{MSE} &= \operatorname{Var}[Y - \hat{Y}] + E[Y - \hat{Y}]^2 \\ &= \operatorname{Var}[Y - aX] + (E[Y] - aE[X] - b)^2. \end{aligned}$$

Den zweiten Term können wir für gegebenes a minimieren, indem wir

$$b \quad = \quad E[Y] - aE[X]$$

wählen. Für den ersten Term ergibt sich

$$\begin{aligned} \operatorname{Var}[Y - aX] &= \operatorname{Cov}[Y - aX, Y - aX] = \operatorname{Var}[Y] - 2a \operatorname{Cov}[X, Y] + a^2 \operatorname{Var}[X] \\ &= \left(a \cdot \sigma[X] - \frac{\operatorname{Cov}[X, Y]}{\sigma[X]} \right)^2 + \operatorname{Var}[Y] - \frac{\operatorname{Cov}[X, Y]^2}{\operatorname{Var}[X]}. \end{aligned} \tag{3.5.2}$$

Dieser Ausdruck wird minimiert, wenn wir $a = \operatorname{Cov}[X, Y]/\sigma[X]^2$ wählen. Die bzgl. des mittleren quadratischen Fehlers optimale Prognose für Y gestützt auf X ist dann

$$\hat{Y}_{\text{opt}} \quad = \quad aX + b \quad = \quad E[Y] + a(X - E[X]).$$

Damit haben wir gezeigt:

Satz 3.25 (Lineare Prognose/Regression von Y gestützt auf X). *Der mittlere quadratische Fehler $E[(\hat{Y} - Y)^2]$ ist minimal unter allen Zufallsvariablen der Form $\hat{Y} = aX + b$ mit $a, b \in \mathbb{R}$ für*

$$\hat{Y}(\omega) \quad = \quad E[Y] \; + \; \frac{\operatorname{Cov}[X, Y]}{\operatorname{Var}[X]} \cdot (X(\omega) - E[X]).$$

Das Problem der linearen Prognose steht in engem Zusammenhang mit der Cauchy-Schwarz-Ungleichung für die Kovarianz. In der Tat ergibt sich diese Ungleichung unmittelbar aus Gleichung (3.5.2):

Satz 3.26 (Cauchy-Schwarz-Ungleichung). *Für $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ gilt*

$$|\operatorname{Cov}[X,Y]| \quad \leq \quad \operatorname{Var}[X]^{1/2} \cdot \operatorname{Var}[Y]^{1/2} \quad = \quad \sigma[X] \cdot \sigma[Y]. \tag{3.5.3}$$

Insbesondere gilt für den Korrelationskoeffizienten im Fall $\sigma[X] \cdot \sigma[Y] \neq 0$:

$$|\varrho[X,Y]| \quad \leq \quad 1. \tag{3.5.4}$$

Gleichheit in (3.5.3) bzw. (3.5.4) gilt genau dann, wenn $a, b \in \mathbb{R}$ existieren, sodass $Y = aX + b$ P-fast sicher gilt. Hierbei ist $\varrho[X,Y] = 1$ im Falle $a > 0$ und $\varrho[X,Y] = -1$ für $a < 0$.

Beweis. Im Fall $\sigma[X] = 0$ gilt $X = E[X]$ P-fast sicher, und die Ungleichung (3.5.3) ist trivialerweise erfüllt. Wir nehmen nun an, dass $\sigma[X] \neq 0$ gilt. Wählt man dann wie oben $a = \operatorname{Cov}[X,Y]/\sigma[X]^2$, so folgt aus (3.5.2) die Cauchy-Schwarz-Ungleichung

$$\operatorname{Var}[Y] - \frac{\operatorname{Cov}[X,Y]^2}{\operatorname{Var}[X]} \quad \geq \quad 0.$$

Die Ungleichung (3.5.4) folgt unmittelbar. Zudem erhalten wir genau dann Gleichheit in (3.5.2) bzw. (3.5.3) bzw. (3.5.4), wenn $\operatorname{Var}[Y - aX] = 0$ gilt, also $Y - aX$ P-fast sicher konstant ist. In diesem Fall folgt $\operatorname{Cov}[X,Y] = \operatorname{Cov}[X,aX] = a \operatorname{Var}[X]$, also hat $\varrho[X,Y]$ dasselbe Vorzeichen wie a. □

Bemerkung (Nichtlinearen Prognosen). Die Zufallsvariable $\hat{Y}$ minimiert den mittleren quadratischen Fehler nur unter allen Zufallsvariablen, die affine Funktionen von X sind. Als *beste nichtlineare Prognose für Y gestützt auf X* bezüglich der L^2-Norm ergibt sich dagegen die bedingte Erwartung $E[Y|X]$, die wir in Kapitel **??** definieren werden.

3.5.2 Regressionsgerade, Methode der kleinsten Quadrate

Ist die zugrundeliegende Wahrscheinlichkeitsverteilung auf Ω eine *empirische Verteilung*

$$P \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \delta_{\omega_i}$$

von n Elementen $\omega_1, \ldots, \omega_n$ aus einer Grundgesamtheit (z.B. alle Bonner Mathematikstudenten des ersten Jahrgangs, oder eine Stichprobe daraus), dann ist die gemeinsame Verteilung zweier reellwertiger Abbildungen $X, Y : \Omega \to \mathbb{R}$ (*statistische Merkmale*, z.B. Punktzahlen im Abitur und in der Analysis-Klausur) gerade die empirische Verteilung der auftretenden Werte:

$$\mu_{X,Y} \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \delta_{(x_i,y_i)}, \qquad x_i = X(\omega_i), \;\; y_i = Y(\omega_i).$$

Die Gewichte der empirischen Verteilung sind die relativen Häufigkeiten

$$\mu_{X,Y}[\{a,b\}] \quad = \quad h(a,b)/n, \qquad h(a,b) \quad = \quad |\{1 \le i \le n : x_i = a, y_i = b\}|.$$

Der Erwartungswert $E[X]$ ist in diesem Fall das **empirische Mittel**

$$E[X] \quad = \quad \sum_{a \in \{x_1,\ldots,x_n\}} a \cdot h(a)/n \quad = \quad \frac{1}{n}\sum_{i=1}^{n} x_i \quad =: \quad \overline{x}_n,$$

und die Varianz ist die **empirische Varianz**

$$\begin{aligned} \mathrm{Var}[X] \;&=\; E[(X - E[X])^2] \;=\; \sum_{a \in \{x_1,\ldots,x_n\}} (a - \overline{x}_n)^2 \cdot h(a)/n \\ &=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 \;= \overline{(x^2)}_n - (\overline{x}_n)^2 \;=:\; \sigma_n^2. \end{aligned}$$

Nach Satz 3.24 minimieren daher das empirische Mittel und jeder Median einer Stichprobe $x_1, \ldots, x_n \in \mathbb{R}$ die Summe der quadratischen bzw. absoluten Abweichungen $\sum (x_i - a)^2$ bzw. $\sum |x_i - a|$.

Als Kovarianz und als Korrelationskoeffizient ergibt sich

$$\mathrm{Cov}[X,Y] \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n) \quad = \quad \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x}_n \overline{y}_n,$$

$$\varrho[X,Y] \quad = \quad \frac{\mathrm{Cov}[X,Y]}{\sigma[X]\sigma[Y]} \quad = \quad \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\left(\sum_{i=1}^{n}(x_i - \overline{x}_n)^2\right)^{1/2}\left(\sum_{i=1}^{n}(y_i - \overline{y}_n)^2\right)^{1/2}} \quad =: \quad r_n.$$

Den **empirischen Korrelationskoeffizienten** r_n der Daten (x_i, y_i), $1 \le i \le n$, verwendet man als Schätzer für die Korrelation von Zufallsgrößen mit unbekannten Verteilungen. Die Grafiken in Abbildung 3.4 zeigen Stichproben von bivariaten Normalverteilungen mit verschiedenen Korrelationskoeffizienten ϱ.

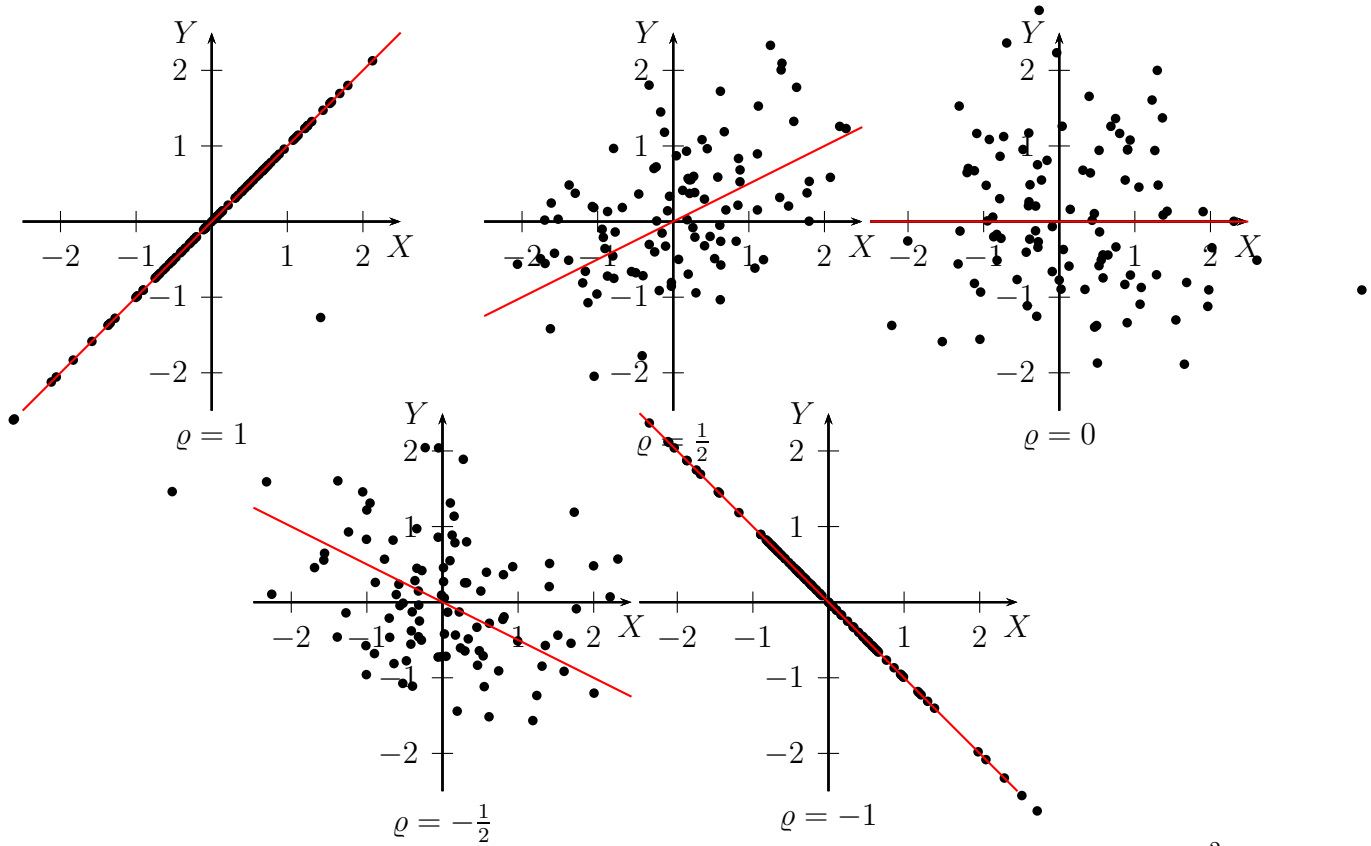


Abbildung 3.4: Stichproben von 100 Punkten von verschiedenen Normalverteilungen im $\mathbb{R}^2$

Als beste lineare Prognose von Y gestützt auf X im quadratischen Mittel erhalten wir die **Regressionsgerade** $y = ax + b$, die die Quadratsumme

$$\sum_{i=1}^{n} (ax_i + b - y_i)^2 \quad = \quad n \cdot \text{MSE}$$

der Abweichungen minimiert. Hierbei gilt nach Satz 3.25:

$$a = \frac{\text{Cov}[X,Y]}{\sigma[X]^2} = \frac{\sum (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum (x_i - \overline{x}_n)^2} \qquad \text{und} \qquad b = E[Y] - a \cdot E[X] = \overline{y}_n - a \cdot \overline{x}_n.$$

Die Regressionsgeraden sind in Abbildung 3.4 eingezeichnet.

3.5.3 Unabhängigkeit und Unkorreliertheit

Auch für allgemeine Zufallsvariablen X und Y folgt aus Unabhängigkeit die Unkorreliertheit von beliebigen Funktionen $f(X)$ und $g(Y)$. Dies werden wir unter anderem beim Beweis des Kolmogorovschen Gesetzes der großen Zahlen in Satz 4.10 ausnutzen, um die Aussage auf nicht-quadratintegrierbare Zufallsvariablen zu erweitern.

Satz 3.27 (Unabhängigkeit und Unkorreliertheit von allgemeinen Zufallsvariablen). *Für Zufallsvariablen $X : \Omega \to S$ und $Y : \Omega \to T$ mit Werten in meßbaren Räumen $(S, \mathcal{S})$ und $(T, \mathcal{T})$ sind äquivalent:*

(1). Die Zufallsvariablen X und Y sind unabhängig, d.h.

$$P[X \in A, Y \in B] \quad = \quad P[X \in A] \cdot P[Y \in B] \qquad \text{für alle } A \in \mathcal{S} \text{ und } B \in \mathcal{T}.$$

(2). Die Zufallsvariablen $f(X)$ und $g(Y)$ sind unkorreliert für alle meßbaren Funktionen f, g mit $f, g \geq 0$ bzw. $f(X), g(Y) \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$, d.h.

$$E[f(X) \cdot g(Y)] \quad = \quad E[f(X)] \cdot E[g(Y)]. \tag{3.5.5}$$

Beweis. Offensichtlich folgt (1) aus (2) durch Wahl von $f = I_A$ und $g = I_B$. Die umgekehrte Implikation folgt durch maßtheoretische Induktion: Gilt (1), dann ist (3.5.5) für Indikatorfunktionen f und g erfüllt. Wegen der Linearität beider Seiten dieser Gleichung in f und g gilt (3.5.5) auch für beliebige Elementarfunktionen. Für messbare $f, g \geq 0$ betrachten wir Folgen von Elementarfunktionen f_n, g_n mit $f_n \nearrow f, g_n \nearrow g$. Die Aussage (3.5.5) folgt durch monotone Konvergenz. Allgemeine Funktionen zerlegen wir in ihren Positiv- und Negativanteil, und wenden die Aussage auf diese an. Also gilt $\mathrm{Cov}[f(X), g(Y)] = 0$ für alle messbaren f, g mit $f, g \geq 0$ bzw. $f(X), g(Y) \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$. □

Das Produkt zweier integrierbarer Zufallsvariablen ist im Allgemeinen nicht wieder integrierbar. Für unabhängige Zufallsvariablen erhalten wir jedoch:

Korollar 3.28. *Sind $X, Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ unabhängig, so gilt:*

$$X \cdot Y \quad \in \quad \mathcal{L}^1(\Omega, \mathcal{A}, P) \qquad \text{und} \qquad E[XY] \quad = \quad E[X] \cdot E[Y].$$

Beweis. Nach Satz 3.27 gilt:

$$E[|XY|] \quad = \quad E[|X|] \cdot E[|Y|] \quad < \quad \infty.$$

Die Formel für $E[XY]$ folgt durch die Zerlegungen $X = X^+ - X^-$ und $Y = Y^+ - Y^-$. □

Beispiel (Kovarianzmatrix von multivariaten Normalverteilungen). Sei $m \in \mathbb{R}^d$, und sei $C \in \mathbb{R}^{d \times d}$ eine symmetrische, positiv definite $d \times d$-Matrix. Die Normalverteilung $N(m, C)$ ist die Verteilung des Zufallsvektors $Y = \sigma Z + m$, wobei $Z = (Z_1, Z_2, ..., Z_d)$ ein Zufallsvektor mit unabhängigen, standardnormalverteilten Komponenten Z_i, und σ eine beliebige $d \times d$-Matrix

mit $C = \sigma\sigma^T$ ist. Die Zufallsvariablen Z_i haben Erwartungswert 0, und nach Satz 3.27 gilt $\mathrm{Cov}[Z_i, Z_j] = \delta_{ij}$. Hieraus folgt, dass der Vektor m der Erwartungswert, und die Matrix C die Kovarianzmatrix der Verteilung $N(m, C)$ ist, denn

$$\begin{aligned} E[Y_i] &= \sum_{k=1}^{d} \sigma_{ik} E[Z_k] + m_i \;=\; m_i, \qquad \text{und} \\ \mathrm{Cov}[Y_i, Y_j] &= \mathrm{Cov}\left[\sum_k \sigma_{ik} Z_k + m_i \,,\, \sum_l \sigma_{jl} Z_l + m_j\right] \\ &= \sum_{k,l} \sigma_{ik}\sigma_{jl} \cdot \mathrm{Cov}[Z_k, Z_l] \;=\; \sum_k \sigma_{ik}\sigma_{jk} \;=\; C_{ij}. \end{aligned}$$

Bemerkung (Linearkombinationen von gemeinsam normalverteilten Zufallsvariablen). Ist Y ein Zufallsvektor mit Verteilung $N(m, C)$, dann ist jede Linearkombination $\alpha \cdot Y = \sum_{i=1}^{d} \alpha_i Y_i$ der Komponenten mit $\alpha \in \mathbb{R}^d$ eine normalverteilte Zufallsvariable mit Mittelwert $\alpha \cdot m$ und Varianz $\alpha \cdot C\alpha$. Dies kann man zum Beispiel durch eine explizite Berechnung der Verteilungsfunktion aus der Dichte von Y zeigen. Wir werden multivariate Normalverteilungen systematischer in Abschnitt 5.3 untersuchen, und dort auch einen eleganteren Beweis der letzten Aussage mithilfe von charakteristischen Funktionen geben.

Teil II

Grenzwertsätze

Sind $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$, unabhängige identisch verteilte (i.i.d.) Zufallsvariablen mit Erwartungswert m, dann gibt es drei unterschiedliche Arten von fundamentalen Aussagen für die Asymptotik der Mittelwerte S_n/n der Summen $S_n = \sum_{i=1}^n X_i$ für große n:

- *Gesetze der großen Zahlen* besagen, dass die Mittelwerte bzgl. eines geeigneten Konvergenzbegriffs für Zufallsvariablen gegen den Erwartungswert m konvergieren, siehe Kapitel 4.

- *Zentrale Grenzwertsätze* beschreiben „typische“ Fluktuationen um den Grenzwert aus einem Gesetz der großen Zahlen, d.h. die asymptotische Form der Verteilung von S_n/n in Bereichen der Größenordnung $O(1/\sqrt{n})$ um den Erwartungswert m, siehe Kapitel 5.

- Aussagen über *große Abweichungen* beschreiben asymptotisch die Wahrscheinlichkeiten der seltenen Abweichungen der Größenordnung $O(1)$ von S_n/n vom Erwartungswert m. Diese Wahrscheinlichkeiten fallen unter geeigneten Voraussetzungen exponentiell ab, siehe Kapitel 6.

Mit dem starken Gesetz der großen Zahlen für Bernoulli-Folgen (Satz 1.5), dem Grenzwertsatz von de Moivre/Laplace, bzw. der Bernsteinungleichung haben wir bereits entsprechende Aussagen kennengelernt, falls die X_i Bernoulli-verteilte Zufallsvariablen sind. In den folgenden Kapiteln werden wir sehen, dass keine spezifische Form der Verteilung vorausgesetzt werden muss, sondern die Aussagen ganz allgemein unter geeigneten Integrierbarkeitsbedingungen gelten.

Kapitel 4

Gesetze der großen Zahlen

In diesem Kapitel beweisen wir verschiedene Gesetze der großen Zahlen, d.h. wir leiten Bedingungen her, unter denen die Mittelwerte $\frac{1}{n}\sum_{i=1}^{n} X_i$ einer Folge $(X_i)_{i\in\mathbb{N}}$ von reellwertigen Zufallsvariablen gegen ihren Erwartungswert konvergieren. Dabei unterscheiden wir verschiedene Arten der Konvergenz, die wir zunächst genauer untersuchen wollen.

4.1 Grundlegende Ungleichungen und Konvergenz von Zufallsvariablen

4.1.1 Konvergenzbegriffe für Zufallsvariablen

Seien $Y_n, n \in \mathbb{N},$ und Y reellwertige Zufallsvariablen, die auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definiert sind. Wir betrachten die folgenden Konvergenzbegriffe für die Folge $(Y_n)_{n\in\mathbb{N}}$:

Definition. *(1). **Fast sichere Konvergenz:***

Die Folge $(Y_n)_{n\in\mathbb{N}}$ konvergiert P-fast sicher gegen Y, falls gilt:

$$P\left[\lim_{n\to\infty} Y_n = Y\right] \quad = \quad P[\{\omega \in \Omega | Y_n(\omega) \to Y(\omega)\}] \quad = \quad 1.$$

*(2). **Stochastische Konvergenz** (Convergence in probability):*

Die Folge $(Y_n)_{n\in\mathbb{N}}$ konvergiert P-stochastisch gegen Y (Notation $Y_n \xrightarrow{P} Y$), falls

$$\lim_{n\to\infty} P[|Y_n - Y| > \varepsilon] \quad = \quad 0 \qquad \textit{für alle } \varepsilon > 0 \textit{ gilt.}$$

(3). ***$\mathcal{L}^p$-Konvergenz*** *($1 \leq p < \infty$):*
Die Folge $(Y_n)_{n\in\mathbb{N}}$ konvergiert in $\mathcal{L}^p(\Omega, \mathcal{A}, P)$ gegen Y, falls

$$\lim_{n\to\infty} E[|Y_n - Y|^p] \quad = \quad 0.$$

Ein Gesetz der großen Zahlen bezüglich fast sicherer Konvergenz heißt **starkes Gesetz der großen Zahlen**, ein G.d.g.Z. bezüglich stochastischer Konvergenz heißt **schwaches Gesetz der großen Zahlen**. Wir wollen nun die Zusammenhänge zwischen den verschiedenen Konvergenzbegriffen untersuchen.

Satz 4.1. *(1). Fast sichere Konvergenz impliziert stochastische Konvergenz.*

(2). Die umgekehrte Implikation gilt im Allgemeinen nicht.

Beweis. (1). Konvergiert Y_n P-fast sicher gegen Y, dann gilt für $\varepsilon > 0$:

$$\begin{aligned}
1 &= P[|Y_n - Y| < \varepsilon \quad \text{schließlich}] \\
&= P\left[\bigcup_m \bigcap_{n\geq m} \{|Y_n - Y| < \varepsilon\}\right] \\
&= \lim_{m\to\infty} P\left[\bigcap_{n\geq m} \{|Y_n - Y| < \varepsilon\}\right] \\
&\leq \lim_{m\to\infty} \inf_{n\geq m} P[|Y_n - Y| < \varepsilon] \\
&= \liminf_{n\to\infty} P[|Y_n - Y| < \varepsilon].
\end{aligned}$$

Es folgt $\lim_{n\to\infty} P[|Y_n - Y| < \varepsilon] = 1$ für alle $\varepsilon > 0$, d.h. Y_n konvergiert auch P-stochastisch gegen Y.

(2). Sei andererseits P das Lebesguemaß auf $\Omega = (0, 1]$ mit Borelscher σ-Algebra. Wir betrachten die Zufallsvariablen

$$Y_1 = I_{(0,1]}, Y_2 = I_{(0,\frac{1}{2}]}, Y_3 = I_{(\frac{1}{2},1]}, Y_4 = I_{(0,\frac{1}{4}]}, Y_5 = I_{(\frac{1}{4},\frac{1}{2}]}, Y_6 = I_{(\frac{1}{2},\frac{3}{4}]}, Y_6 = I_{(\frac{3}{4},1]}, \ldots$$

Dann gilt

$$P[|Y_n| > \varepsilon] \quad = \quad P[Y_n = 1] \quad \to \quad 0 \qquad \text{für alle } \varepsilon > 0,$$

also konvergiert Y_n stochastisch gegen 0, obwohl

$$\limsup Y_n(\omega) \quad = \quad 1 \qquad \text{für alle } \omega \in \Omega \text{ gilt.}$$

□

Hier ist ein weiteres Beispiel, das den Unterschied zwischen stochastischer und fast sicherer Konvergenz zeigt:

Beispiel. Sind $T_1, T_2, \ldots$ unter P unabhängige $\mathrm{Exp}(1)$-verteilte Zufallsvariablen, dann konvergiert $T_n/\log n$ P-stochastisch gegen 0, denn

$$P\left[\left|\frac{T_n}{\log n}\right| \geq \varepsilon\right] \quad = \quad P[T_n \geq \varepsilon \cdot \log n] \quad = \quad n^{-\varepsilon} \quad \stackrel{n\to\infty}{\to} \quad 0$$

für alle $\varepsilon > 0$. Andererseits gilt nach (2.1.5) aber

$$\limsup_{n\to\infty} \frac{T_n}{\log n} \quad = \quad 1 \qquad P\text{-fast sicher,}$$

also konvergiert $T_n/\log n$ nicht P-fast sicher.

Obwohl die stochastische Konvergenz selbst nicht fast sichere Konvergenz impliziert, kann man aus einer Verschärfung von stochastischer Konvergenz die fast sichere Konvergenz schließen. Wir sagen, dass eine Folge $Y_n, n \in \mathbb{N}$, von Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ **schnell stochastisch** gegen Y **konvergiert**, falls

$$\sum_{n=1}^{\infty} P[|Y_n - Y| \geq \varepsilon] \quad < \quad \infty \qquad \text{für alle } \varepsilon > 0.$$

Lemma 4.2. *Aus schneller stochastischer Konvergenz folgt fast sichere Konvergenz.*

Beweis. Wir können o.B.d.A. $Y = 0$ annehmen. Konvergiert Y_n schnell stochastisch gegen 0, dann gilt:

$$P[\limsup |Y_n| \leq \varepsilon] \quad \geq \quad P[|Y_n| \geq \varepsilon \text{ nur endlich oft}] \quad = \quad 1.$$

Es folgt

$$P[\limsup |Y_n| \neq 0] \quad = \quad P\left[\bigcup_{\varepsilon\in\mathbb{Q}_+} \{\limsup |Y_n| > \varepsilon\}\right] \quad = \quad 0.$$

□

Ähnlich zeigt man:

Lemma 4.3. *Konvergiert Y_n P-stochastisch gegen Y, dann existiert eine Teilfolge Y_{n_k}, die P-fast sicher gegen Y konvergiert.*

Beweis. Wieder können wir o.B.d.A. $Y = 0$ annehmen. Konvergiert Y_n stochastisch gegen 0, dann existiert eine Teilfolge Y_{n_k} mit

$$P\left[|Y_{n_k}| \geq \frac{1}{k}\right] \quad \leq \quad \frac{1}{k^2}.$$

Nach dem Lemma von Borel-Cantelli folgt

$$P\left[|Y_{n_k}| \geq \frac{1}{k} \text{ nur endlich oft}\right] \quad = \quad 1,$$

also $Y_{n_k} \to 0$ P-fast sicher. □

Als nächstes beweisen wir eine Erweiterung der Čebyšev-Ungleichung, die wir an vielen Stellen verwenden werden. Insbesondere impliziert sie, dass stochastische Konvergenz schwächer ist als $\mathcal{L}^p$-Konvergenz.

4.1.2 Die Markov-Čebyšev-Ungleichung

Sei $X : \Omega \to \overline{\mathbb{R}}$ eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Wir verwenden die folgende Notation:

Notation: $E[X\,;\,A] := E[X \cdot I_A] = \int_A X dP$.

Satz 4.4 (Allgemeine Markov-Ungleichung). *Sei $h : [0,\infty] \to [0,\infty]$ monoton wachsend und Borel-messbar. Dann gilt*

$$P[|X| \geq c] \;\leq\; \frac{E[h(|X|)\,;\,|X| \geq c]}{h(c)} \;\leq\; \frac{E[h(|X|)]}{h(c)} \qquad \textit{für alle } c > 0 \textit{ mit } h(c) \neq 0.$$

Beweis. Da h nichtnegativ und monoton wachsend ist, gilt

$$h(|X|) \;\geq\; h(|X|) \cdot I_{\{|X| \geq c\}} \;\geq\; h(c) \cdot I_{\{|X| \geq c\}},$$

also auch

$$E[h(|X|)] \;\geq\; E[h(|X|)\,;\,|X| \geq c] \;\geq\; h(c) \cdot P[|X| \geq c].$$

□

Wichtige Spezialfälle:

(1). **Markov - Ungleichung:** Für $h(x) = x$ erhalten wir:

$$P[|X| \geq c] \;\leq\; \frac{E[|X|]}{c} \qquad \text{für alle } \; c > 0.$$

Insbesondere gilt für eine Zufallsvariable X mit $E[|X|] = 0$:

$$P[|X| \geq c] = 0 \qquad \text{für alle } \; c > 0,$$

also auch $P[|X| > 0] = 0$, d.h. $X = 0$ P-fast sicher.

(2). **Čebyšev - Ungleichung:** Für $h(x) = x^2$ und $X = Y - E[Y]$ mit $Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ erhalten wir:

$$P[|Y - E[Y]| \geq c] \leq \frac{E[(Y - E[Y])^2]}{c^2} = \frac{\mathrm{Var}[Y]}{c^2} \qquad \text{für alle } c > 0.$$

Diese Ungleichung haben wir bereits in Abschnitt **??** im Beweis des schwachen Gesetzes der großen Zahlen verwendet.

(3). **Exponentielle Abschätzung:** Für $h(x) = \exp(tx)$ mit $t > 0$ erhalten wir wegen $I_{\{X \geq c\}} \leq e^{-tc} e^{tX}$:

$$P[X \geq c] = E[I_{\{X \geq c\}}] \leq e^{-tc} \cdot E[e^{tX}].$$

Die Abbildung $t \mapsto E[e^{tX}]$ heißt **momentenerzeugende Funktion** der Zufallsvariablen X. Exponentielle Ungleichungen werden wir in Abschnitt **??** zur Kontrolle der Wahrscheinlichkeiten *großer Abweichungen* vom Gesetz der großen Zahlen verwenden.

Als erste Anwendung der allgemeinen Markovungleichung zeigen wir für reellwertige Zufallsvariablen X, X_n $(n \in \mathbb{N})$:

Korollar 4.5 ($\mathcal{L}^p$-Konvergenz impliziert stochastische Konvergenz). *Für $1 \leq p < \infty$ gilt:*

$$E[|X_n - X|^p] \to 0 \quad \Rightarrow \quad P[|X_n - X| > \varepsilon] \to 0 \quad \textit{für alle } \varepsilon > 0.$$

Beweis. Nach der Markovungleichung mit $h(x) = x^p$ gilt:

$$P[|X_n - X| \geq \varepsilon] \quad \leq \quad \frac{1}{\varepsilon^p} E[|X_n - X|^p].$$

□

Bemerkung. Aus stochastischer Konvergenz folgt im Allgemeinen nicht $\mathcal{L}^p$-Konvergenz (Übung). Es gilt aber: Konvergiert $X_n \to X$ stochastisch, und ist die Folge der Zufallsvariablen $|X_n|^p$ $(n \in \mathbb{N})$ **gleichmäßig integrierbar**, d.h.

$$\sup_{n \in \mathbb{N}} E[|X_n|^p \,;\, |X_n| \geq c] \to 0 \quad \text{für } c \to \infty,$$

dann konvergiert X_n gegen X in $\mathcal{L}^p$ *(Verallgemeinerter Satz von Lebesgue)*. Wir benötigen diese Aussage im Moment nicht, und werden sie daher erst in der Vorlesung »Stochastische Prozesse« beweisen.

Als nächstes wollen wir den Zusammenhang zwischen $\mathcal{L}^p$-Konvergenz für verschiedene Werte von $p \geq 1$ untersuchen. Dazu verwenden wir eine weitere fundamentale Ungleichung:

4.1.3 Die Jensensche Ungleichung

Ist $\ell(x) = ax + b$ eine affine Funktion auf $\mathbb{R}$, und $X \in \mathcal{L}^1$ eine integrierbare Zufallsvariable, dann folgt aus der Linearität des Lebesgueintegrals:

$$E[\ell(X)] \quad = \quad E[aX + b] \quad = \quad aE[X] + b \quad = \quad \ell(E[X]) \tag{4.1.1}$$

Da konvexe Funktionen Suprema einer Familie von affinen Funktionen (nämlich der Tangenten an den Funktionsgraphen der konvexen Funktion) sind, ergibt sich für konvexe Funktionen eine entsprechende *Ungleichung*:

Satz 4.6 (Jensensche Ungleichung). *Ist P eine Wahrscheinlichkeitsverteilung, $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ eine reellwertige Zufallsvariable, und $h : \mathbb{R} \to \mathbb{R}$ eine konvexe Abbildung, dann ist $E[h(X)^-] < \infty$, und es gilt*

$$h(E[X]) \;\leq\; E[h(X)].$$

Warnung: Diese Aussage gilt (wie auch (4.1.1)) nur für die Integration bzgl. eines Wahrscheinlichkeitsmaßes!
Bevor wir die Jensensche Ungleichung beweisen, erinnern wir kurz an die Definition und elementare Eigenschaften von konvexen Funktionen:

Bemerkung. Eine Funktion $h : \mathbb{R} \to \mathbb{R}$ ist genau dann konvex, wenn

$$h(\lambda x + (1-\lambda)y) \quad \leq \quad \lambda h(x) + (1-\lambda)h(y) \quad \text{für alle } \lambda \in [0,1] \text{ und } x, y \in \mathbb{R}$$

gilt, d.h. wenn alle Sekanten oberhalb des Funktionsgraphen liegen.

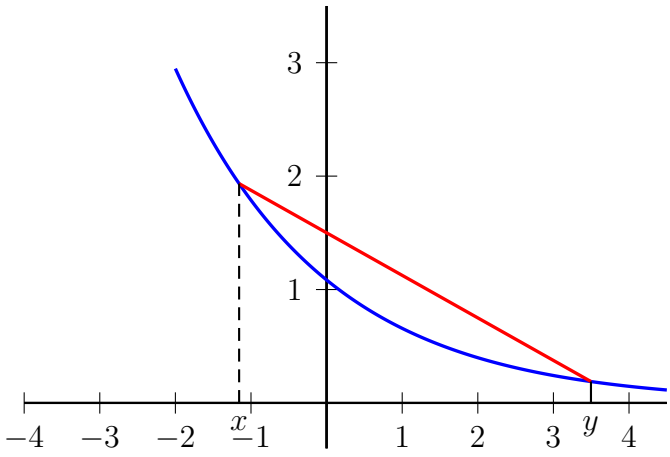


Abbildung 4.1: Sekante an konvexer Funktion

Hieraus folgt, dass jede konvexe Funktion stetig ist: Für $a < b < x < y < c < d$ gilt nämlich

$$\frac{h(b) - h(a)}{b - a} \quad \leq \quad \frac{h(y) - h(x)}{y - x} \quad \leq \quad \frac{h(d) - h(c)}{d - c}.$$

Also sind die Differenzenquotienten $\frac{h(y)-h(x)}{y-x}$ gleichmäßig beschränkt auf (b, c), und somit ist h gleichmäßig stetig auf (b, c). Da konvexe Funktionen stetig sind, sind sie auch messbar. Die Existenz des Erwartungswertes $E[h(X)]$ in $(-\infty, \infty]$ folgt dann aus $E[h(X)^-] < \infty$.

Wir beweisen nun die Jensensche Ungleichung:

Beweis. Ist h konvex, dann existiert zu jedem $x_0 \in \mathbb{R}$ eine affine Funktion ℓ (*Stützgerade*) mit $\ell(x_0) = h(x_0)$ und $\ell \leq h$, siehe die Analysis Vorlesung oder [A. KLENKE: „WAHRSCHEINLICHKEITSTHEORIE“, Abschnitt 7.2].

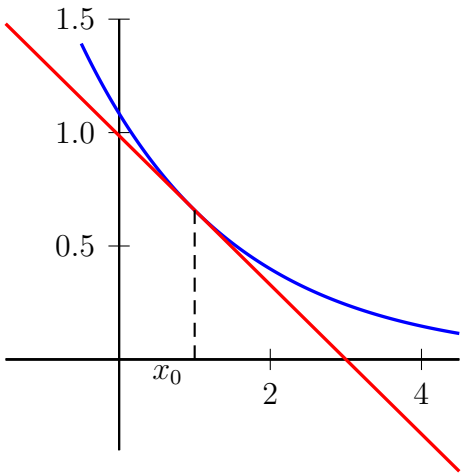


Abbildung 4.2: Darstellung von $\ell(x)$ und $h(x)$

Wählen wir $x_0 := E[X]$, dann folgt

$$h(E[X]) = \ell(E[X]) = E(\ell[X]) \leq E[h(X)].$$

Der Erwartungswert auf der rechten Seite ist definiert, da $h(X)$ durch die integrierbare Zufallsvariable $\ell(X)$ nach unten beschränkt ist. Insbesondere gilt $E[h(X)^-] \leq E[\ell(X)^-] < \infty$. □

Korollar 4.7 ($\mathcal{L}^q$-Konvergenz impliziert $\mathcal{L}^p$-Konvergenz). *Für $1 < p \leq q$ gilt:*

$$\|X\|_p \quad := \quad E[|X|^p]^{\frac{1}{p}} \quad \leq \quad \|X\|_q.$$

Insbesondere folgt $\mathcal{L}^p$-Konvergenz aus $\mathcal{L}^q$-Konvergenz.

Beweis. Nach der Jensenschen Ungleichung gilt

$$E[|X|^p]^{\frac{q}{p}} \quad \leq \quad E[|X|^q],$$

da die Funktion $h(x) = |x|^{q/p}$ für $q \geq p$ konvex ist. □

Nach dem Korollar gilt für $p \leq q$:

$$\mathcal{L}^p(\Omega, \mathcal{A}, P) \quad \supseteq \quad \mathcal{L}^q(\Omega, \mathcal{A}, P),$$

und

$$X_n \to X \text{ in } \mathcal{L}^q \quad \Rightarrow \quad X_n \to X \text{ in } \mathcal{L}^p.$$

Man beachte, dass diese Aussage nur für **endliche Maße** wahr ist, da im Beweis die Jensensche Ungleichung verwendet wird.
Mithilfe der Jensenschen Ungleichung beweist man auch die **Hölderungleichung**:

$$E[|XY|] \quad \leq \quad \|X\|_p \cdot \|Y\|_q \qquad \text{für } p, q \in [1, \infty] \text{ mit} \qquad \frac{1}{p} + \frac{1}{q} \quad = \quad 1.$$

4.2 Starke Gesetze der großen Zahlen

Wir werden nun Gesetze der großen Zahlen unter verschiedenen Voraussetzungen an die zugrundeliegenden Zufallsvariablen beweisen. Zunächst nehmen wir an, dass $X_1, X_2, \ldots \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ quadratintegrierbare Zufallsvariablen sind, deren Varianzen gleichmäßig beschränkt sind, und deren Korrelationen hinreichend schnell abklingen:

Annahme: *„Schnelles Abklingen der positiven Korrelation“*

(A) Es existiert eine Folge $c_n \in \mathbb{R}_+$ $(n \in \mathbb{N})$ mit

$$\sum_{n=0}^{\infty} c_n < \infty$$

und

$$\operatorname{Cov}[X_i, X_j] \quad \leq \quad c_{|i-j|} \qquad \text{für alle } i, j \in \mathbb{N}. \tag{4.2.1}$$

Die Bedingung (A) ist insbesondere erfüllt, wenn die *Korrelationen exponentiell abfallen*, d.h. wenn

$$|\operatorname{Cov}[X_i, X_j]| \quad \leq \quad c \cdot \alpha^{|i-j|}$$

für ein $\alpha \in (0, 1)$ und $c \in \mathbb{R}^+$ gilt. Sind etwa die Zufallsvariablen X_i unkorreliert, und ist die Folge der *Varianzen beschränkt*, d.h. gilt

(A1) $\mathrm{Cov}[X_i, X_j] = 0$ für alle $i, j \in \mathbb{N}$, und

(A2) $v := \sup_i \mathrm{Var}[X_i] < \infty$,

dann ist die Annahme (A) mit $c_0 = v$ und $c_n = 0$ für $n > 0$ erfüllt. In diesem Fall haben wir bereits in Abschnitt **??** ein schwaches Gesetz der großen Zahlen bewiesen.

Wichtig: Es wird **keine Unabhängigkeit vorausgesetzt!**

Sei nun

$$S_n \quad = \quad X_1 + \ldots + X_n$$

die Summe der ersten n Zufallsvariablen.

4.2.1 Das schwache Gesetz der großen Zahlen

Den Beweis des schwachen Gesetzes der großen Zahlen aus Abschnitt **??** können wir auf den hier betrachteten allgemeinen Fall erweitern:

Satz 4.8 (Schwaches Gesetz der großen Zahlen, $\mathcal{L}^2$-Version)**.** *Unter der Voraussetzung (A) gilt für alle $n \in \mathbb{N}$ und $\varepsilon > 0$:*

$$E\left[\left(\frac{S_n}{n} - \frac{E[S_n]}{n}\right)^2\right] \leq \frac{v}{n}, \qquad \textit{und} \tag{4.2.2}$$

$$P\left[\left|\frac{S_n}{n} - \frac{E[S_n]}{n}\right| \geq \varepsilon\right] \leq \frac{v}{\varepsilon^2 n} \tag{4.2.3}$$

mit $v := c_0 + 2 \cdot \sum_{n=1}^{\infty} c_n < \infty$. Gilt insbesondere $E[X_i] = m$ für alle $i \in \mathbb{N}$, dann folgt

$$\frac{S_n}{n} \quad \to \quad m \qquad \textit{in } \mathcal{L}^2(\Omega, \mathcal{A}, P) \textit{ und } P\textit{-stochastisch.}$$

Beweis. Unter Verwendung der Voraussetzung an die Kovarianzen erhalten wir

$$\begin{aligned} E\left[\left(\frac{S_n}{n} - \frac{E[S_n]}{n}\right)^2\right] &= \mathrm{Var}\left[\frac{S_n}{n}\right] = \frac{1}{n^2}\mathrm{Var}[S_n] \\ &= \frac{1}{n^2}\sum_{i,j=1}^{n}\mathrm{Cov}[X_i, X_j] \leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{|i-j|} \\ &\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=-\infty}^{\infty} c_{|k|} = \frac{v}{n} \end{aligned}$$

Die zweite Behauptung folgt daraus durch Anwenden der Čebyšev-Ungleichung. □

Bemerkung. (1). Im Fall unkorrelierter Zufallsvariablen X_i (Annahmen (A1) und (A2)) ist die Aussage ein Spezialfall einer allgemeinen funktionalanalytischen Sachverhalts:

Das Mittel von beschränkten orthogonalen Vektoren im Hilbertraum
$L^2(\Omega, \mathcal{A}, P) = \mathcal{L}^2(\Omega, \mathcal{A}, P)/\sim$ *konvergiert gegen* 0.

Unkorreliertheit der X_i bedeutet gerade, dass die Zufallsvariablen

$$Y_i \quad := \quad X_i - E[X_i]$$

orthogonal in L^2 sind - beschränkte Varianzen der X_i ist gleichbedeutend mit der Beschränktheit der L^2 Normen der Y_i. Es gilt

$$S_n - E[S_n] \quad = \quad \sum_{i=1}^{n} Y_i,$$

also

$$\begin{aligned} E\left[\left(\frac{S_n}{n} - \frac{E[S_n]}{n}\right)^2\right] &= \left\|\frac{1}{n}\sum_{i=1}^{n} Y_i\right\|_{L^2}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle Y_i, Y_j\rangle_{L^2} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\|Y_i\|_{L^2}^2 \leq \frac{1}{n}\sup_i \|Y_i\|_{L^2}^2. \end{aligned}$$

(2). Die $\mathcal{L}^2$-Konvergenz und stochastische Konvergenz von $(S_n - E[S_n])/n$ gegen 0 gilt auch, falls die Korrelationen „langsam“ abklingen, d.h. falls (4.2.1) für eine nicht summierbare Nullfolge c_n erfüllt ist. In diesem Fall erhält man allerdings im Allgemeinen keine Abschätzung der Ordnung $O(\frac{1}{n})$ für den Fehler in (4.2.2) bzw. (4.2.3).

(3). Eine für große n deutlich bessere Abschätzung des Fehlers in (4.2.3) (mit exponentiellem Abfall in n) erhält man bei Unabhängigkeit und exponentieller Integrierbarkeit der X_i mithilfe der *exponentiellen Ungleichung*, siehe Satz 4.13 unten.

4.2.2 Das starke Gesetz für quadratintegrierbare Zufallsvariablen

Unter derselben Voraussetzung wie in Satz 4.8 gilt sogar P-fast sichere Konvergenz:

Satz 4.9 (Starkes Gesetz großer Zahlen, $\mathcal{L}^2$-Version). *Unter der Voraussetzung (A) konvergiert*

$$\frac{S_n(\omega)}{n} - \frac{E[S_n]}{n} \quad \longrightarrow \quad 0$$

für P-fast alle $\omega \in \Omega$. Insbesondere gilt

$$\frac{S_n}{n} \quad \longrightarrow \quad m \qquad P\text{-fast sicher,}$$

falls $E[X_i] = m$ für alle i.

Der Übersichtlichkeit halber führen wir den Beweis zunächst unter den stärkeren Voraussetzungen (A1) und (A2). Der allgemeine Fall ist eine Übungsaufgabe, die sich gut zum Wiederholen der Beweisschritte eignet:

Beweis unter den Annahmen (A1) und (A2). Wir können o.B.d.A. $E[X_i] = 0$ für alle i voraussetzen – andernfalls betrachten wir die zentrierten Zufallsvariablen $\widetilde{X_i} := X_i - E[X_i]$; diese sind wieder unkorreliert mit beschränkten Varianzen. Zu zeigen ist dann:

$$\frac{S_n}{n} \to 0 \qquad P\text{-fast sicher.}$$

Wir unterteilen den Beweis in mehrere Schritte:

(1). *Schnelle stochastische Konvergenz gegen 0 entlang der Teilfolge $n_k = k^2$:* Aus der Čebyšev-Ungleichung folgt:

$$P\left[\left|\frac{S_{k^2}}{k^2}\right| \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2} \operatorname{Var}\left[\frac{S_{k^2}}{k^2}\right] \leq \frac{1}{\varepsilon^2 k^2} \sup_i \operatorname{Var}[X_i].$$

Da die Varianzen beschränkt sind, ist der gesamte Ausdruck durch die Summanden einer summierbaren Reihe beschränkt. Somit ergibt sich nach Borel-Cantelli:

$$\frac{S_{k^2}(\omega)}{k^2} \quad \to \quad 0$$

für alle ω außerhalb einer Nullmenge N_1.

(2). Wir untersuchen nun die Fluktuationen der Folge S_n zwischen den Werten der Teilfolge $n_k = k^2$. Sei

$$D_k \quad := \quad \max_{k^2 \leq l < (k+1)^2} |S_l - S_{k^2}|.$$

Wir zeigen *schnelle stochastische Konvergenz gegen 0 für D_k/k^2*. Für $\varepsilon > 0$ haben wir

$$\begin{aligned} P\left[\frac{D_k}{k^2} \geq \varepsilon\right] &= P\left[\bigcup_{k^2 \leq l < (k+1)^2} \{\, |S_l - S_{k^2}| > \varepsilon k^2 \,\}\right] \\ &\leq \sum_{l=k^2}^{k^2+2k} P[|S_l - S_{k^2}| > \varepsilon k^2] \leq \frac{\text{const.}}{k^2}, \end{aligned}$$

denn nach der Čebyšev-Ungleichung gilt für $k^2 \le l \le k^2 + 2k$:

$$\begin{aligned} P[|S_l - S_{k^2}| > \varepsilon k^2] &\le \frac{1}{\varepsilon^2 k^4} \operatorname{Var}[S_l - S_{k^2}] \le \frac{1}{\varepsilon^2 k^4} \operatorname{Var}\left[\sum_{i=k^2+1}^{l} X_i\right] \\ &\le \frac{l-k^2}{\varepsilon^2 k^4} \sup_i \operatorname{Var}[X_i] \le \text{const} \cdot \frac{k}{k^4}. \end{aligned}$$

Nach Lemma 4.2 folgt daher

$$\frac{D_k(\omega)}{k^2} \to 0$$

für alle ω außerhalb einer Nullmenge N_2.

(3). Zu gegebenem n wählen wir nun $k = k(n)$ mit $k^2 \le n < (k+1)^2$. Durch Kombination der ersten beiden Schritte erhalten wir:

$$\left|\frac{S_n(\omega)}{n}\right| \le \frac{|S_{k^2}(\omega)| + D_k(\omega)}{n} \le \left|\frac{S_{k^2}(\omega)}{k^2}\right| + \frac{D_k(\omega)}{k^2} \longrightarrow 0 \qquad \text{für } n \to \infty$$

für alle ω außerhalb der Nullmenge $N_1 \cup N_2$. Also konvergiert S_n/n P-fast sicher gegen 0.

□

Beispiel (Random Walk im $\mathbb{R}^d$). Sei $S_n = X_1 + ... + X_n$ ein Random Walk im $\mathbb{R}^d$ mit unabhängigen identisch verteilten Inkrementen X_i mit Verteilung μ. Gilt

$$E[\|X_i\|^2] = \int_{\mathbb{R}^d} \|x\|^2 \, \mu(dx) < \infty,$$

dann folgt nach dem schwachen Gesetz der großen Zahlen (angewandt auf die Komponenten $S_n^{(k)} = \sum_{i=1}^{n} X_i^{(k)}$ des Vektors S_n):

$$\frac{S_n(\omega)}{n} \longrightarrow m \qquad \text{für } P\text{-fast alle } \omega,$$

wobei $m = \int_{\mathbb{R}^d} x\mu(dx)$ der Schwerpunkt der Inkrementverteilung ist. Insbesondere gilt für $m \neq 0$:

$$S_n \sim m \cdot n \qquad \text{für } n \to \infty \quad P\text{-fast sicher,}$$

d.h. S_n wächst linear mit Geschwindigkeit m. Im Fall $m = 0$ gilt dagegen

$$\frac{S_n(\omega)}{n} \to 0 \quad P\text{-fast sicher,}$$

d.h. der Random Walk wächst sublinear. Eine viel präzisere Beschreibung der pfadweisen Asymptotik des Random Walk im Fall $m = 0$ liefert der *Satz vom iterierten Logarithmus*:

$$\begin{aligned} \limsup_{n\to\infty} \frac{S_n(\omega)}{\sqrt{n \log\log n}} &= +1 \qquad P\text{-fast sicher,} \\ \liminf_{n\to\infty} \frac{S_n(\omega)}{\sqrt{n \log\log n}} &= -1 \qquad P\text{-fast sicher,} \end{aligned}$$

siehe z.B. [BAUER: „WAHRSCHEINLICHKEITSTHEORIE“].

Beispiel (Wachstum in zufälligen Medien). Um ein zufälliges Populationswachstum zu beschreiben, definieren wir Zufallsvariablen X_n $(n \in \mathbb{N})$ durch

$$X_0 = 1, \quad X_n = Y_n \cdot X_{n-1},$$

d.h. $X_n = \prod_{i=1}^n Y_i$. Hierbei nehmen wir an, dass die Wachstumsraten Y_i unabhängige identisch verteilte Zufallsvariablen mit $Y_i > 0$ P-f.s. sind. Sei $m = E[Y_i]$.

(1). ASYMPTOTIK DER ERWARTUNGSWERTE: Da die Y_i unabhängig sind, gilt:

$$E[X_n] \;=\; \prod_{i=1}^n E[Y_i] \;=\; m^n.$$

Die mittlere Populationsgröße wächst also im *superkritischen Fall* $m > 1$ exponentiell und fällt im *subkritischen Fall* $m < 1$ exponentiell ab.

Konkretes Beispiel: In einem Glücksspiel setzt der Spieler in jeder Runde die Hälfte seines Kapitals. Mit Wahrscheinlichkeit $\frac{1}{2}$ erhält er das c-fache des Einsatzes zurück, und mit Wahrscheinlichkeit $\frac{1}{2}$ erhält er nichts zurück. Hier gilt:

$$Y_i \;=\; \begin{cases} \frac{1}{2}(1+c) & \text{mit } p = \frac{1}{2} \\ \frac{1}{2} & \text{mit } p = \frac{1}{2} \end{cases},$$

also

$$m \;=\; E[Y_i] = \frac{1}{4}(1+c) + \frac{1}{4} = \frac{2+c}{4}.$$

Das Spiel ist also „fair“ für $c = 2$ und „superfair“ für $c > 2$.

(2). ASYMPTOTIK VON $X_n(\omega)$: Wir nehmen nun an, dass $\log Y_1 \in \mathcal{L}^2$ gilt. Nach dem starken Gesetz der großen Zahlen folgt dann:

$$\frac{1}{n} \log X_n = \frac{1}{n} \sum_{i=1}^n \log Y_i \;\to\; E[\log Y_1] =: \alpha \qquad P\text{-f.s.}$$

Also existiert für $\varepsilon > 0$ ein $N(\omega)$ mit $N(\omega) < \infty$ P-fast sicher,

$$X_n(\omega) \;\leq\; e^{(\alpha+\varepsilon)n} \quad \text{und} \quad X_n(\omega) \;\geq\; e^{(\alpha-\varepsilon)n} \qquad \text{für alle } n \geq N(\omega).$$

Für $\alpha < 0$ fällt X_n also P-fast sicher exponentiell ab, während X_n für $\alpha > 0$ P-fast sicher exponentiell wächst.

(3). ZUSAMMENHANG VON α UND m: Nach der Jensenschen Ungleichung gilt:

$$\alpha = E[\log Y_1] \leq \log E[Y_1] = \log m.$$

Hierbei haben wir benutzt, dass der Logarithmus eine konkave, bzw. $-\log$ eine konvexe Funktion ist. Im subkritischen Fall $m < 1$ ist also auch α strikt negativ, d.h. X_n fällt auch P-f.s. exponentiell ab. Im superkritischen Fall $m > 1$ kann es aber passieren, dass *trotzdem* $\alpha < 0$ gilt, d.h. obwohl die Erwartungswerte exponentiell wachsen, fällt X_n P-fast sicher exponentiell! Im Beispiel

$$Y_i = \begin{cases} \frac{1}{2}(1+c) & \text{mit } p = \frac{1}{2} \\ \frac{1}{2} & \text{mit } p = \frac{1}{2} \end{cases}$$

von oben wachsen die Erwartungswerte exponentiell für $c > 2$, aber es gilt

$$\alpha = E[\log Y_i] = \frac{1}{2}\left(\log \frac{1+c}{2} + \log \frac{1}{2}\right) = \frac{1}{2} \log \frac{1+c}{4} \geq 0 \;\Leftrightarrow\; c \geq 3.$$

Für $c \in (2,3)$ ist das Spiel also superfair mit fast sicherem exponentiellem Bankrott!

Die Voraussetzungen des Satzes von Lebesgue sind in dieser Situation nicht erfüllt, denn es gilt:

$$E[X_n] \nearrow \infty, \quad \text{obwohl } X_n \to 0 \quad P\text{-fast sicher.}$$

4.2.3 Von $\mathcal{L}^2$ nach $\mathcal{L}^1$ mit Unabhängigkeit

Sind Zufallsvariablen $X, Y : \Omega \to S$ unabhängig, so sind $f(X)$ und $g(Y)$ für beliebige beschränkte oder nichtnegative Funktionen $f, g : S \to \mathbb{R}$ unkorreliert. Bisher konnten wir zeigen, dass das starke Gesetz der großen Zahlen für unkorrelierte (bzw. schwach korrelierte) Zufallsvariablen $X_n \in \mathcal{L}^2$ mit gleichmäßig beschränkten Varianzen gilt. Die Unabhängigkeit der X_n ermöglicht es, diese Aussage auf integrierbare Zufallsvariablen (d.h. $\mathcal{L}^1$ statt $\mathcal{L}^2$) zu erweitern:

Satz 4.10 (Kolmogorovs Gesetz der großen Zahlen). *Seien $X_1, X_2, \ldots \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ paarweise unabhängig und identisch verteilt mit $E[X_i] = m$. Dann gilt:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i \;=\; m \qquad \textit{P-fast sicher.}$$

Kolmogorov hatte eine entsprechende Aussage unter der Annahme von Unabhängigkeit (statt paarweiser Unabhängigkeit) bewiesen. Der Beweis unter der schwächeren Voraussetzung stammt von Etemadi (1981).

Bemerkung (Dynamische Systeme, Ergodensatz). In einer dynamischen Interpretation bedeutet die Aussage

$$\frac{1}{n} \sum_{i=1}^{n} X_i(\omega) \longrightarrow m = \int x \, \mu_{X_i}(dx) \qquad P\text{-fast sicher,}$$

des starken Gesetzes der großen Zahlen, dass die „zeitlichen Mittelwerte“ der Zufallsvariablen X_i gegen den „räumlichen Mittelwert“ m konvergieren. Dies ist ein Spezialfall eines viel allgemeineren *Ergodensatzes*, der eine entsprechende Aussage für ergodische dynamische Systeme liefert, siehe z.B. BREIMAN: PROBABILITY oder DURRETT: PROBABILITY: THEORY AND EXAMPLES.

Beweis von Satz 4.10. Wir führen den Beweis in mehreren Schritten.

(1). *Reduktion auf nichtnegative Zufallsvariablen.*

Wir können o.B.d.A. $X_i \geq 0$ für alle $i \in \mathbb{N}$ voraussetzen. Andernfalls zerlegen wir $X_i = X_i^+ - X_i^-$. Die Zufallsvariablen $X_i^+, i \in \mathbb{N}$, bzw. $X_i^-, i \in \mathbb{N}$, sind jeweils Funktionen der X_i, und daher wieder paarweise unabhängig. Aus dem Gesetz der großen Zahlen für X_i^+ und X_i^- folgt das Gesetz der großen Zahlen für die Zufallsvariablen X_i.

(2). *Reduktion auf Gesetz der großen Zahlen für $Y_i := X_i \cdot I_{\{X_i \leq i\}}$.*

Nach dem Lemma von Borel-Cantelli gilt

$$P[Y_i \neq X_i \text{ unendlich oft}] = 0,$$

denn

$$\begin{aligned}
\sum_{i=1}^{\infty} P[Y_i \neq X_i] &= \sum_{i=1}^{\infty} P[X_i > i] \\
&= \sum_{i=1}^{\infty} P[X_1 > i] \qquad (X_i \text{ identisch verteilt}) \\
&\leq \int_0^{\infty} P[X_1 > x] \, dx \qquad (P[X_1 > x] \text{ monoton fallend}) \\
&= E[X_1] \quad < \quad \infty.
\end{aligned}$$

Also konvergiert $\frac{1}{n} \sum_{i=1}^{n} X_i$ P-fast sicher gegen m, falls dasselbe für $\frac{1}{n} \sum_{i=1}^{n} Y_i$ gilt.

Sei nun

$$S_n \;=\; \sum_{i=1}^{n} Y_i.$$

Die Zufallsvariablen Y_i sind wieder paarweise unabhängig, und es gilt $0 \leq Y_i \leq i$.

(3). *Konvergenz der Erwartungswerte.*

Da die Zufallsvariablen Y_i nicht mehr identisch verteilt sind, bestimmen wir zunächst den Grenzwert der Erwartungswerte der Mittelwerte S_n/n. Nach dem Satz von der monotonen Konvergenz gilt

$$E[Y_i] = E[X_i \,;\, X_i \leq i] = E[X_1 \cdot I_{\{X_1 \leq i\}}] \longrightarrow E[X_1] = m, \qquad \text{für } i \to \infty,$$

also auch

$$E\left[\frac{S_n}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] \longrightarrow m \qquad \text{für } n \to \infty.$$

(4). *P-fast sichere Konvergenz von $\frac{S_n}{n}$ entlang der Teilfolgen $k_n = \lfloor \alpha^n \rfloor, \alpha > 1$.*

Vorbemerkung: Es gilt

$$\sum_{n \geq m} \frac{1}{k_n^2} \;=\; \frac{1}{\lfloor \alpha^m \rfloor^2} + \frac{1}{\lfloor \alpha^{m+1} \rfloor^2} + \ldots \;\leq\; \frac{\text{const.}}{\lfloor \alpha^m \rfloor^2} \;=\; \frac{\text{const.}}{k_m^2}$$

mit einer von m unabhängigen Konstanten.

Behauptung:

$$\frac{S_{k_n}}{k_n} \;\longrightarrow\; \lim_{n\to\infty} E\left[\frac{S_{k_n}}{k_n}\right] = m \qquad P\text{-fast sicher.}$$

Beweis der Behauptung: Nach dem Lemma von Borel-Cantelli genügt es,

$$\sum_{n=1}^{\infty} P\left[\left|\frac{S_{k_n} - E[S_{k_n}]}{k_n}\right| \geq \varepsilon\right] \;<\; \infty$$

zu zeigen. Dies ist der Fall, wenn

$$\sum_{n=1}^{\infty} \mathrm{Var}\left[\frac{S_{k_n}}{k_n}\right] \;< \infty$$

gilt. Wegen

$$\mathrm{Var}[Y_i] \leq E[Y_i^2] = E[X_i^2 \,;\, X_i \leq i] = E[X_1^2 \,;\, X_1 \leq i]$$

erhalten wir mithilfe der Vorbemerkung

$$\begin{aligned}
\sum_{n=1}^{\infty} \operatorname{Var}\left[\frac{S_{k_n}}{k_n}\right] &= \sum_{n=1}^{\infty} \frac{1}{k_n^2} \cdot \sum_{i=1}^{k_n} \operatorname{Var}[Y_i] \\
&\leq \sum_{i=1}^{\infty} E[X_1^2\,;\, X_1 \leq i] \cdot \sum_{n:k_n \geq i} \frac{1}{k_n^2} \\
&\leq \text{const.} \cdot \sum_{i=1}^{\infty} E\left[X_1^2\,;\, X_1 \leq i\right] \cdot \frac{1}{i^2} \\
&\leq \text{const.} \cdot \sum_{i=1}^{\infty} \sum_{j=1}^{i} j^2 \cdot P[X_1 \in (j-1,j]] \cdot \frac{1}{i^2} \\
&= \text{const.} \cdot \sum_{j=1}^{\infty} j^2 \cdot P[X_1 \in (j-1,j]] \cdot \sum_{i=j}^{\infty} \frac{1}{i^2} \\
&\leq \text{const.} \cdot \sum_{j=1}^{\infty} j \cdot P[X_1 \in (j-1,j]] \\
&= \text{const.} \cdot E\left[\sum_{j=1}^{\infty} j \cdot I_{\{X_1 \in (j-1,j]\}}\right] \\
&\leq \text{const.} \cdot E[X_1 + 1] \;<\; \infty.
\end{aligned}$$

(5). *P-fast sichere Konvergenz von $\frac{S_n}{n}$.*

Für $l \in \mathbb{N}$ mit $k_n \leq l \leq k_{n+1}$ gilt wegen $Y_i \geq 0$:

$$S_{k_n} \leq S_l \leq S_{k_{n+1}}.$$

Es folgt

$$\frac{k_n}{k_{n+1}} \cdot \frac{S_{k_n}}{k_n} = \frac{S_{k_n}}{k_{n+1}} \leq \frac{S_l}{l} \leq \frac{S_{k_{n+1}}}{k_n} = \frac{k_{n+1}}{k_n} \cdot \frac{S_{k_{n+1}}}{k_{n+1}}.$$

Für $n \to \infty$ erhalten wir wegen $\frac{k_{n+1}}{k_n} \to \alpha$ und $\frac{S_{k_n}(\omega)}{k_n} \to m$:

$$\frac{m}{\alpha} \quad \leq \quad \liminf \frac{S_l(\omega)}{l} \quad \leq \quad \limsup \frac{S_l(\omega)}{l} \quad \leq \quad \alpha m$$

für alle ω außerhalb einer von α abhängenden Nullmenge N_α. Für ω außerhalb der Nullmenge $\bigcup_{\substack{\alpha > 1 \\ \alpha \in \mathbb{Q}}} N_\alpha$ folgt somit:

$$\lim_{l \to \infty} \frac{S_l(\omega)}{l} \;=\; m.$$

□

Korollar 4.11 (Gesetz der großen Zahlen ohne Integrierbarkeit). *Seien $X_1, X_2 ...$ paarweise unabhängige, identisch verteilte, nicht-negative Zufallsvariablen. Dann gilt:*

$$\lim_{n\to\infty} \frac{1}{n} \cdot \sum_{i=1}^{n} X_i(\omega) \;=\; E[X_1] \;\in [0,\infty] \qquad P\text{-fast sicher.}$$

Beweis. Nach Satz 4.10 gilt die Aussage im Fall $E[X_1] < \infty$. Für $E[X_1] = \infty$ erhalten wir für $k \in \mathbb{N}$:

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i \quad \geq \quad \liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (X_i \wedge k) \quad = \quad E[X_1 \wedge k] \qquad P\text{-fast sicher.}$$

Für $k \to \infty$ folgt dann mit monotoner Konvergenz

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i \quad \geq \quad E[X_1] \quad = \quad \infty,$$

und damit die Behauptung. □

4.3 Momentenerzeugende Funktionen und exponentielle Abschätzungen

In diesem Abschnitt führen wir momentenerzeugende und charakteristische Funktionen von reellen Zufallsvariablen ein und beweisen einige grundlegende Aussagen über diese Funktionen. Anschließend zeigen wir, wie nicht-asymptotische obere Schranken für die Wahrscheinlichkeiten großer Abweichungen vom Gesetz der großen Zahlen mithilfe momentenerzeugender Funktionen hergeleitet werden können. Charakteristische Funktionen werden wir in Kapitel 5 zum Beweis von zentralen Grenzwertsätzen verwenden.

4.3.1 Momentenerzeugende und charakteristische Funktionen

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}^d$ eine Zufallsvariable mit Verteilung μ. Wir definieren den Erwartungswert bzw. das Lebesgue-Integral einer komplexwertigen Zufallsvariable $Z = U + iV$ mit Real- und Imaginärteil $U, V : \Omega \to \mathbb{R}$ durch $E[Z] = E[U] + iE[V]$.

Definition (Momentenerzeugende und charakteristische Funktion).
Die Funktionen $M : \mathbb{R}^d \to (0,\infty]$ bzw. $\phi : \mathbb{R}^d \to \mathbb{C}$,

$$\begin{aligned} M(t) \;&:=\; E[e^{t\cdot X}] \;=\; \int_{\mathbb{R}^d} e^{t\cdot x}\, \mu(dx), \\ \phi(t) \;&:=\; E[e^{it\cdot X}] \;=\; \int_{\mathbb{R}^d} e^{it\cdot x}\, \mu(dx), \end{aligned}$$

heißen ***momentenerzeugende*** *bzw.* ***charakteristische Funktion*** *der Zufallsvariable X oder der Verteilung μ.*

Da die Funktionen $t \mapsto e^{t \cdot x}$ und $t \mapsto e^{it \cdot x}$ für $t \in \mathbb{R}^d$ nichtnegativ bzw. beschränkt sind, sind die Erwartungswerte definiert. Dabei nimmt $M(t)$ den Wert $+\infty$ an, falls $\exp(t \cdot X)$ nicht integrierbar ist. Für den Betrag der komplexen Zahl $\phi(t)$ gilt dagegen

$$|\phi(t)| \;\leq\; E[|\exp(it \cdot x)|] \;=\; 1 \qquad \text{für alle } t \in \mathbb{R}^d.$$

Bemerkung (Fourier- und Laplace-Transformation). Die Funktion $\phi(-t) = \int e^{-itx}\, \mu(dx)$ ist die *Fourier-Transformation* des Maßes μ. Ist μ absolutstetig bzgl. des Lebesguemaßes mit Dichte f, dann ist $\phi(-t)$ die Fourier-Transformation der Funktion f, d.h.

$$\phi(-t) \quad = \quad \int_{\mathbb{R}^d} e^{-it \cdot x} f(x)\, dx \quad =: \quad \widehat{f}(t).$$

Entsprechend ist

$$M(-t) \quad = \quad \int_{\mathbb{R}^d} e^{-t \cdot x}\, \mu(dx) \qquad (t > 0)$$

die *Laplace-Transformation* des Maßes μ bzw. der Dichte f.

Rechenregeln. Die folgenden Rechenregeln ergeben sich unmittelbar aus den Definitionen der momentenerzeugenden bzw. charakteristischen Funktionen:

(1). Sind $X, Y : \Omega \to \mathbb{R}^d$ unabhängige Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$, dann gilt

$$M_{X+Y}(t) \quad = \quad M_X(t) \cdot M_Y(t) \qquad \text{und} \qquad \phi_{X+Y}(t) \quad = \quad \phi_X(t) \cdot \phi_Y(t)$$

für alle $t \in \mathbb{R}^d$.

(2). Ist $X = (X_1, \ldots, X_d) : \Omega \to \mathbb{R}^d$ ein Zufallsvektor mit unabhängigen Komponenten $X_1, \ldots, X_d$, dann gilt für $t = (t_1, \ldots, t_d) \in \mathbb{R}^d$:

$$M_X(t) \quad = \quad \prod_{i=1}^{d} M_{X_i}(t_i) \qquad \text{und} \qquad \phi_X(t) \quad = \quad \prod_{i=1}^{d} \phi_{X_i}(t_i).$$

(3). Für $A \in \mathbb{R}^{d \times d}$ und $b \in \mathbb{R}^d$ gilt

$$M_{AX+b}(t) \quad = \quad e^{t \cdot b}\, M_X(A^T t) \qquad \text{und} \qquad \phi_{AX+b}(t) \quad = \quad e^{it \cdot b} \phi_X(A^T t).$$

(4). Es gilt stets $M(0) = \phi(0) = 1$ und $\phi(-t) = \overline{\phi(t)}$ für alle $t \in \mathbb{R}$.

Beispiel (Binomialverteilung). Die Binomialverteilung $Bin(n,p)$ ist die Verteilung der Summe $\sum_{i=1}^n Y_i$ von unabhängigen $\mathrm{Bernoulli}(p)$-verteilten Zufallsvariablen $Y_1, ..., Y_n$. Also sind

$$\phi(t) \;=\; \prod_{i=1}^n \phi_{Y_i}(t) = \left(1-p+pe^{it}\right)^n \;, \quad \text{und} \quad M(t) \;=\; \left(1-p+pe^{t}\right)^n$$

die charakteristische und momentenerzeugende Funktion von $\mathrm{Bin}(n,p)$.

Der Übersichtlichkeit halber beschränken wir uns nun auf den Fall $d = 1$. Wir zeigen, dass sich die Momente $E[X^n]$ einer Zufallsvariable $X : \Omega \to \mathbb{R}$ unter geeigneten Voraussetzungen aus der momentenerzeugenden bzw. charakteristischen Funktion berechnen lassen. Die nötigen Voraussetzungen sind allerdings im Fall der momentenerzeugenden Funktion viel stärker.

Satz 4.12 (Momentenerzeugung). *(1). Ist M endlich auf $(-\delta,\delta)$ für ein $\delta > 0$, dann existiert der Erwartungswert $M(z) := E[e^{zX}]$ für alle $z \in \mathbb{C}$ mit $|Re(z)| < \delta$, und es gilt*

$$E[e^{zX}] \quad = \quad \sum_{n=0}^\infty \frac{z^n}{n!} E[X^n] \qquad \textit{für alle } z \in \mathbb{C} \textit{ mit } |z| < \delta.$$

Insbesondere folgt

$$M^{(n)}(0) = E[X^n] \qquad \textit{für alle } n \in \mathbb{Z}_+ \,.$$

(2). Ist $E[|X|^n] < \infty$ für ein $n \in \mathbb{N}$, dann gilt $\phi \in C^n(\mathbb{R})$ und

$$\phi^{(n)}(t) \quad = \quad i^n \cdot E[X^n e^{itX}] \qquad \textit{für alle } t \in \mathbb{R} \,. \tag{4.3.1}$$

Man beachte, dass die Voraussetzung im ersten Teil des Satzes erfüllt ist, falls $M(s) < \infty$ und $M(-s) < \infty$ für ein festes $s > 0$ gilt. Nach der Jensenschen Ungleichung folgt nämlich aus $M(s) < \infty$ auch

$$M(t) \;=\; E[e^{tX}] \;\le\; E[e^{sX}]^{t/s} \;<\; \infty \qquad \text{für alle } t \in [0,s].$$

Entsprechend folgt $M < \infty$ auf $[-s,0]$ aus $M(-s) < \infty$.

Beweis. (1). Aus der Voraussetzung und dem Satz von der monotonen Konvergenz ergibt sich für $s \in (0,\delta)$:

$$\sum_{n=0}^\infty \frac{s^n}{n!} E[|X|^n] \quad = \quad E\left[e^{s|X|}\right] \quad \le \quad E\left[e^{sX}\right] + E\left[e^{-sX}\right] \quad < \quad \infty \,.$$

Insbesondere existieren alle Momente $E[X^n], n \in \mathbb{N}$, sowie die exponentiellen Momente $E[e^{zX}]$ für $z \in \mathbb{C}$ mit $|\mathrm{Re}(z)| < \delta$. Nach dem Satz von Lebesgue erhalten wir für $z \in \mathbb{C}$ mit $|z| < \delta$ zudem

$$\sum_{n=0}^{\infty} \frac{z^n}{n!} E[X^n] = \lim_{m\to\infty} E\left[\sum_{n=0}^{m} \frac{(zX)^n}{n!}\right] = E\left[\lim_{m\to\infty} \sum_{n=0}^{m} \frac{(zX)^n}{n!}\right] = E[e^{zX}]\,,$$

da $e^{s|X|}$ für $s \geq |z|$ eine Majorante der Partialsummen ist.

(2). Wir zeigen die Behauptung durch Induktion nach n. Für $n = 0$ gilt (4.3.1) nach Definition von $\phi(t)$. Ist $E[|X|^{n+1}] < \infty$, dann folgt nach Induktionsvoraussetzung und mit dem Satz von Lebesgue:

$$\begin{aligned} \frac{\phi^{(n)}(t+h) - \phi^{(n)}(t)}{h} &= \frac{1}{h} E\left[(iX)^n \left(e^{i(t+h)X} - e^{itX}\right)\right] \\ &= E\left[(iX)^n \frac{1}{h} \int_t^{t+h} iXe^{isX}\, ds\right] \quad \to \quad E\left[(iX)^{n+1}\, e^{itX}\right] \end{aligned}$$

für $h \to 0$, also

$$\phi^{(n+1)}(t) \quad = \quad E[(iX)^{n+1} e^{itX}].$$

Die Stetigkeit der rechten Seite in t folgt ebenfalls aus dem Satz von Lebesgue und der Voraussetzung $E[|X|^{n+1}] < \infty$.

□

Beispiele. (1). Für eine Zufallsvariable X mit Verteilungsdichte $f(x) = const. \cdot e^{-|x|^{1/2}}$ gilt $E[|X|^n] < \infty$ für alle $n \in \mathbb{N}$. Also ist die charakteristische Funktion beliebig oft differenzierbar. Die momentenerzeugende Funktion $M_X(t)$ ist hingegen nur für $t = 0$ endlich.

(2). Ein Standardbeispiel einer Verteilung, deren Momente nicht existieren, ist die *Cauchy-Verteilung* mit Dichte

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

Für eine Cauchy-verteilte Zufallsvariable X gilt $M_X(t) = \infty$ für alle $t \neq 0$. Trotzdem existiert

$$\phi_X(t) = e^{-|t|} \qquad \text{für alle } t \in \mathbb{R}\,.$$

Die charakteristische Funktion ist allerdings bei 0 nicht differenzierbar.

Bemerkung (Zusammenhang von M und ϕ). Gilt $M < \infty$ auf $(-\delta, \delta)$ für ein $\delta > 0$, dann hat die Funktion M eine eindeutige analytische Fortsetzung auf den Streifen $\{z \in \mathbb{C} \,:\, |Re(z)| < \delta\}$ in der komplexen Zahlenebene, die durch $M(z) = E[\exp(zX)]$ gegeben ist. In diesem Fall gilt

$$\phi(t) \quad = \quad M(it) \qquad \text{für alle } t \in \mathbb{R},$$

insbesondere ist die charakteristische Funktion dann durch die momentenerzeugende Funktion eindeutig bestimmt.

Die letzte Bemerkung ermöglicht manchmal eine vereinfachte Berechnung von charakteristischen Funktionen.

Beispiel (**Normalverteilungen**). (1). Für eine standardnormalverteilte Zufallsvariable Z gilt:

$$M_Z(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx-x^2/2} dx = e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx = e^{t^2/2} < \infty \quad \text{für } t \in \mathbb{R}.$$

Die eindeutige analytische Fortsetzung auf $\mathbb{C}$ ist die als Potenzreihe darstellbare Funktion $M_Z(z) = \exp(z^2/2)$. Also ist die charakteristische Funktion gegeben durch

$$\phi_Z(t) \quad = \quad M_Z(it) \quad = \quad e^{-t^2/2} \qquad \text{für alle } \ t \in \mathbb{R}.$$

(2). Eine normalverteilte Zufallsvariable X mit Mittel m und Varianz σ^2 können wir darstellen als $X = \sigma Z + m$ mit $Z \sim N(0,1)$. Also gilt:

$$\begin{aligned} M_X(t) &= e^{mt}\, M_Z(\sigma t) = \exp\left(mt + \sigma^2 t^2/2\right), \\ \phi_X(t) &= \exp\left(imt - \sigma^2 t^2/2\right). \end{aligned}$$

Bemerkung (**Satz von Bochner**). Eine Funktion $\phi : \mathbb{R} \to \mathbb{C}$ ist genau dann eine charakteristische Funktion einer Wahrscheinlichkeitsverteilung auf $\mathbb{R}$, wenn gilt:

(1). $\phi(0) = 1$ und $|\phi(t)| \leq 1$ für alle $t \in \mathbb{R}$,

(2). ϕ ist gleichmäßig stetig,

(3). ϕ ist *nicht-negativ definit*, d.h.

$$\sum_{i,j=1}^{n} \phi(t_i - t_j) z_i \overline{z_j} \geq 0 \quad \forall\, n \in \mathbb{N},\ t_1, ..., t_n \in \mathbb{R},\ z_1, ..., z_n \in \mathbb{C}.$$

Dass jede charakteristische Funktion einer Wahrscheinlichkeitsverteilung die Eigenschaften (1)-(3) hat, prüft man leicht nach. Der Beweis der umgekehrten Aussage findet sich z.B. in Vol. II des Lehrbuchs von Feller.

4.3.2 Große Abweichungen vom Gesetz der großen Zahlen

Seien $X_1, X_2, ... \in \mathscr{L}^1(\Omega, \mathcal{A}, P)$ unabhängige, identisch verteilte Zufallsvariablen mit Erwartungswert m und momentenerzeugender Funktion

$$M(t) = E[e^{tX_1}] \,,$$

und sei $S_n = X_1 + ... + X_n$.
Der folgende Satz verschärft die *nicht-asymptotische* obere Schranke für die Wahrscheinlichkeit großer Abweichungen vom Gesetz der großen Zahlen aus der Bernstein-Ungleichung (Satz **??**), und verallgemeinert diese auf nicht Bernoulli-verteilte Zufallsvariablen.

Satz 4.13 (Chernoff). *Für alle $n \in \mathbb{N}$ und $a \in \mathbb{R}$ gilt:*

$$\begin{aligned} P\left[\frac{S_n}{n} \geq a\right] &\leq e^{-nI(a)} \quad \textit{falls } a \geq m, \textit{ bzw.} \\ P\left[\frac{S_n}{n} \leq a\right] &\leq e^{-nI(a)} \quad \textit{falls } a \leq m, \end{aligned}$$

wobei die exponentielle Abfallrate $I(a)$ gegeben ist durch

$$I(a) = \sup_{t \in \mathbb{R}} (at - \log M(t)).$$

Beweis. Wir zeigen diese Aussage im Fall $a \geq m$ – der Beweis für $a \leq m$ verläuft analog. Der Beweis erfolgt in drei Schritten:

(1). *Zentrieren:* Wir können o.B.d.A. $m = 0$ annehmen. Andernfalls betrachten wir die zentrierten Zufallsvariablen $\widetilde{X}_i = X_i - E[X_i]$, die wieder unabhängig und identisch verteilt sind. Man überzeugt sich leicht, dass aus der Behauptung für $\widetilde{X_i}$ die Behauptung für X_i folgt (Übung).

(2). *Exponentielle Markovungleichung:* Für alle $t \geq 0$ und $n \in \mathbb{N}$ gilt:

$$\begin{aligned} P\left[\frac{S_n}{n} \geq a\right] &= P[S_n \geq na] \leq e^{-tna} E[e^{tS_n}] \\ &\overset{X_i \text{ iid}}{=} e^{-tna}\, E[e^{tX_1}]^n = e^{-(at - \log M(t)) \cdot n}. \end{aligned}$$

(3). *Optimieren der Abschätzung:* Bilden wir das Infimum der für verschiedene $t \geq 0$ erhaltenen Abschätzungen, dann ergibt sich:

$$P\left[\frac{S_n}{n} \geq a\right] \leq \inf_{t \geq 0} e^{-(at - \log M(t)) \cdot n} = e^{-\sup_{t \geq 0}(at - \log M(t)) \cdot n}.$$

Es bleibt zu zeigen, dass

$$\sup_{t\geq 0}(at-\log M(t)) \quad = \quad \sup_{t\in\mathbb{R}}(at-\log M(t)) \quad = \quad I(a).$$

Dies ist in der Tat der Fall, denn für $t<0$ und $a\geq m$ gilt nach der Jensenschen Ungleichung und der Voraussetzung $m=0$:

$$\begin{aligned} at-\log M(t) &\leq -\log E[e^{tX_1}] \quad \leq \quad E[-\log e^{tX_1}] \\ &= -tm \quad = \quad 0 \quad = \quad a\cdot 0-\log M(0). \end{aligned}$$

□

Die Analyse der Asymptotik der Wahrscheinlichkeiten großer Abweichungen auf der exponentiellen Skala werden wir in Kapitel 6 durch den Beweis einer asymptotischen unteren Schranke mit derselben Ratenfunktion I vervollständigen. Die Chernoff-Schranke aus dem Satz oben hat aber den Vorteil, dass sie nicht nur asymptotisch (d.h. für $n\to\infty$), sondern für jedes feste n gilt !

Um die Aussage aus dem Satz von Chernoff zu interpretieren, untersuchen wir die Ratenfunktion I genauer. Insbesondere interessiert uns, wann $I(a)$ strikt positiv ist, denn in diesem Fall fallen die Wahrscheinlichkeiten großer Abweichungen exponentiell in n ab. Wir beginnen mit einer Bemerkung zur Funktion $\Lambda:=\log M$:

Bemerkung (Kumulantenerzeugende Funktion). Die Funktion $\Lambda(t)=\log M(t)$, $t\in\mathbb{R}$, heißt *logarithmische momentenerzeugende* oder *kumulantenerzeugende Funktion* von X_1. Sie hat unter anderem die folgenden Eigenschaften:

(1). Λ ist konvex.

(2). $\Lambda(0)=0$.

(3). Gilt $M(t)<\infty$ auf $(-\delta,\delta)$ für ein $\delta>0$, dann ist

$$\begin{aligned} \Lambda'(0) &= \frac{M'(0)}{M(0)} = m, \quad \text{und} \\ \Lambda''(0) &= \frac{M''(0)}{M(0)}-\frac{M'(0)^2}{M(0)^2} \quad = \quad E[X_1^2]-E[X_1]^2 \quad = \quad \mathrm{Var}[X_1]. \end{aligned}$$

Die höheren Ableitungen von Λ heißen *Kumulanten* von X_1.

Die Ratenfunktion I ist die *Legendre-Transformation* der Funktion Λ, d.h.

$$I(a) \;=\; \sup_{t\in\mathbb{R}} f_a(t) \qquad \text{mit} \qquad f_a(t) \;=\; at - \Lambda(t).$$

Die Legendre-Transformation einer konvexen Funktion hat eine einfache geometrische Bedeutung: Wie man aus Abbildung 4.3.2 sieht, ist der Wert $I(a)$ der negative Achsenabschnitt der (eindeutigen) Tangente an den Graphen von Λ mit Steigung a (wobei wir $I(a) = \infty$ setzen, falls keine solche Tangente existiert).

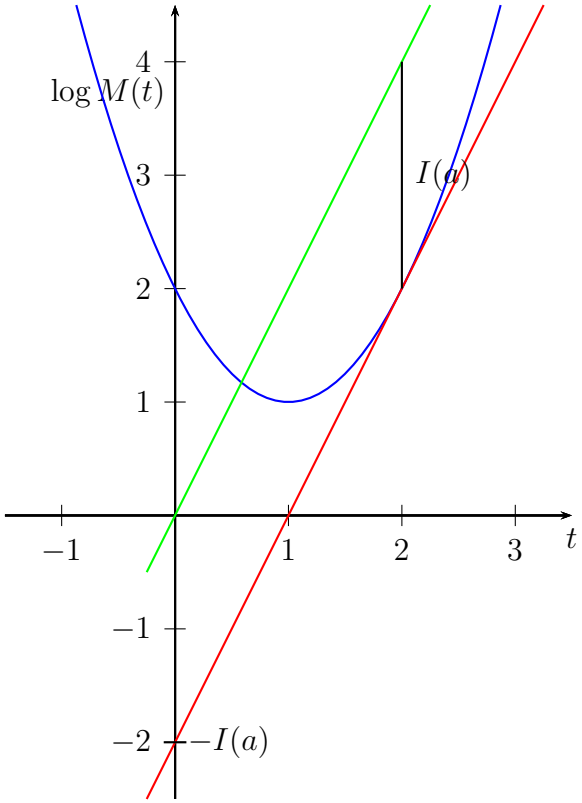


Abbildung 4.3: Geometrische Darstellung der Rate $I(a)$ als negativer Achsenabschnitt der eindeutigen Tangente mit Steigung a (rot) an die Kumulantenerzeugende Funktion (blau)

Wichtige Eigenschaften der Ratenfunktion sind:

(1). I ist wieder konvex.

(2). Es gilt $I(a) \geq f_a(0) = 0$ für alle $a \in \mathbb{R}$.

(3). Ist $M(t) < \infty$ auf $(-\delta, \delta)$ für ein $\delta > 0$, dann folgt $f_a \in C^\infty(-\delta, \delta)$ mit $f_a(0) = 0$ und $f_a'(0) = a - m$. In diesem Fall ist $I(a)$ für $a \neq m$ strikt positiv:

$$I(a) \quad = \quad \sup f_a \quad > \quad 0 \qquad \text{für alle } a \neq m.$$

Unter der Voraussetzung in (3) ergibt sich ein *exponentieller Abfall der Wahrscheinlichkeiten großer Abweichungen !* Sind die Zufallsvariablen X_i dagegen nicht exponentiell integrierbar, dann kann es auch passieren, dass die Abfallrate $I(a)$ für $a \neq m$ gleich 0 ist. Die Wahrscheinlichkeiten großer Abweichungen fallen in diesem Fall langsamer als exponentiell ab, denn es gilt auch eine asymptotische untere Schranke mit derselben Ratenfunktion I, siehe Satz 6.2 unten.

Für konkrete Verteilungen der Zufallsvariablen X_i kann man die Kumulantenerzeugende Funktion Λ und die Ratenfunktion I manchmal explizit berechnen:

Beispiel (Normalverteilung). Für normalverteilte Zufallsvariablen $X_i \sim N(m, \sigma^2)$ ergibt sich $I(a) = \frac{(a-m)^2}{2\sigma^2}$, also

$$P\left[\frac{S_n}{n} \geq a\right] \leq e^{-\frac{(a-m)^2 n}{2\sigma^2}} \qquad \text{für alle } a \geq m.$$

Die Ratenfunktion hat eine Nullstelle beim Erwartungswert m, da die Mittelwert S_n/n gegen diesen konvergieren. Jenseits von m erhalten wir eine Abschätzung der Wahrscheinlichkeiten mit einer exponentiellen Abfallrate, die quadratisch in a wächst. Da in diesem Fall S_n/n wieder normalverteilt ist, kann man die Wahrscheinlichkeiten auch präziser mithilfe von Lemma 1.17 abschätzen. Es zeigt sich, dass die Chernoff-Abschätzung hier zwar die optimale exponentielle Rate liefert; mit der genaueren Gaußschen Abschätzung (1.4.3) gewinnt man aber einen zusätzlichen Faktor der Größenordnung $n^{-1/2}$ (Übung).

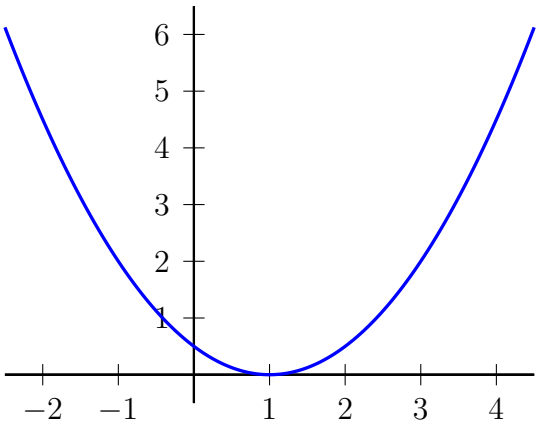


Abbildung 4.4: Legendre-Transformation der logarithmischen momentenerzeugenden Funktion einer $\mathcal{N}(1,1)$-verteilten Zufallsvariable.

Beispiel (Exponentialverteilung). Für $X_i \sim \mathrm{Exp}(\lambda)$ ergibt sich die Ratenfunktion

$$I(a) = \begin{cases} \lambda a - 1 - \log(\lambda a) & \text{für } a > 0, \\ \infty & \text{für } a \leq 0. \end{cases}.$$

Diese hat eine Nullstelle beim Erwartungswert $1/\lambda$. Da nicht positive Werte mit Wahrscheinlichkeit 1 nicht auftreten, hat die Ratenfunktion auf dem Intervall $(-\infty, 0]$ den Wert $+\infty$.

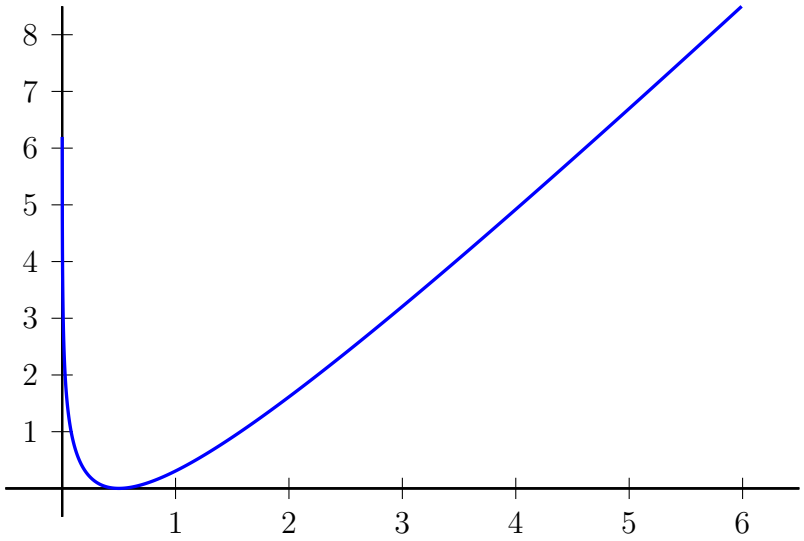


Abbildung 4.5: Legendre-Transformierte der logarithmischen momentenerzeugenden Funktion einer $\mathrm{Exp}(2)$-verteilten Zufallsvariable

Beispiel (Bernoulli-Verteilung; Bernstein-Ungleichung). Für $X_i \sim \mathrm{Bernoulli}(p)$ erhält man

$$I(a) = a \log\left(\frac{a}{p}\right) + (1-a)\log\left(\frac{1-a}{1-p}\right) \quad \text{für } a \in [0,1], \qquad I(a) = +\infty \quad \text{sonst,}$$

wobei wir $0 \log 0 := 0$ setzen. Wegen $I(a) \geq 2(a-p)^2$ verschärft die Abschätzung aus dem Satz von Chernoff in diesem Fall die in Satz **??** hergeleitete obere Schranke

$$P\left[\frac{S_n}{n} \geq a\right] \leq e^{-2(a-p)^2 n} \qquad \text{für } a \geq p.$$

Wir werden später sehen, dass $I(a)$ sich als relative Entropie der Bernoulli(a)-Verteilung bzgl. der Bernoulli (p)-Verteilung interpretieren lässt.

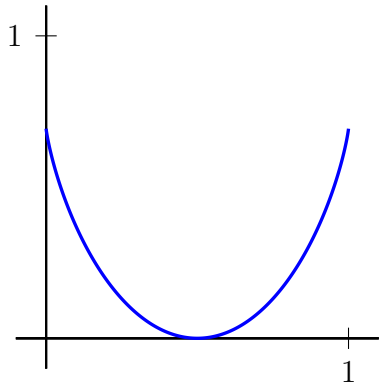


Abbildung 4.6: Legendre-Transformation der logarithmischen momentenerzeugenden Funktion einer Bernoulli$(1/2)$-verteilten Zufallsvariable

Beispiel (Ehrenfestmodell im Gleichgewicht). Es befinden sich $n = 10^{23}$ Moleküle in einem Gefäß. Jedes Molekül sei jeweils mit Wahrscheinlichkeit $1/2$ in der linken bzw. rechten Hälfte. Seien X_i $(1 \leq i \leq n)$ Bernoulli$(1/2)$-verteilte unabhängige Zufallsvariablen, wobei $X_i = 1$ dafür steht, dass sich das i-te Molekül in der linken Hälfte befindet. Der Anteil S_n/n der Moleküle in dieser Hälfte konvergiert nach dem Gesetz der großen Zahlen fast sicher gegen $1/2$.

Wie groß ist die Wahrscheinlichkeit $p := P\left[\frac{S_n}{n} \geq \frac{1}{2} + 10^{-10}\right]$ *?*

Eine Abschätzung mit der Čebyšev-Ungleichung liefert:

$$p \quad \leq \quad 10^{20} \cdot \mathrm{Var}\left[S_n/n\right] \quad = \quad \frac{1}{4} \cdot 10^{-3} \quad = \quad \frac{1}{4000}.$$

Durch Anwenden der exponentiellen Abschätzung erhält man dagegen die viel präzisere Aussage

$$p \quad \leq \quad e^{-2n(10^{-10})^2} \quad = \quad e^{-2000}\,.$$

Eine Abweichung von der Größenordnung 10^{-10} vom Mittelwert ist also *praktisch unmöglich* ! Die makroskopische Größe S_n/n ist daher de facto deterministisch.

4.3.3 Inversion der Fouriertransformation

Die folgende zentrale Aussage zeigt, dass eine Wahrscheinlichkeitsverteilung auf $\mathbb{R}$ *eindeutig* durch ihre charakteristische Funktion ϕ festgelegt ist. Der Satz liefert sogar eine *explizite Formel* zur Rekonstruktion der Verteilung aus ϕ. Gilt zudem $M < \infty$ auf einem Intervall $(-\delta, \delta)$ mit $\delta > 0$, dann erhält man die charakteristische Funktion wie oben bemerkt durch analytische Fortsetzung der momentenerzeugenden Funktion M auf die imaginäre Achse. In diesem Fall ist die Verteilung somit auch durch die momentenerzeugende Funktion eindeutig bestimmt !

Satz 4.14 (Lévys Inversionsformel). *Sei ϕ die charakteristische Funktion einer Zufallsvariable X mit Verteilung μ. Dann gilt:*

(1). Für $a, b \in \mathbb{R}$ mit $a < b$ gilt

$$\frac{1}{2}\,\mu[\{a\}] \;+\; \mu[(a,b)] \;+\frac{1}{2}\,\mu[\{b\}] \;=\; \frac{1}{2\pi}\,\lim_{T\to\infty}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\,\phi(t)\,dt\;.$$

(2). Ist $\int_{-\infty}^{\infty}|\phi(t)|\,dt<\infty$, dann ist μ absolutstetig mit stetiger Dichte

$$f(x) \;=\; \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\,\phi(t)\,dt.$$

Bemerkung. (1). Die Verteilung μ ist durch (1) eindeutig festgelegt, denn für $c, d \in \mathbb{R}$ mit $c < d$ gilt:

$$\frac{1}{2}\,\mu[\{a\}]+\mu[(a,b)]+\frac{1}{2}\,\mu[\{b\}] \quad=\quad \frac{1}{2}\,\Big(\mu\Big[[a,b]\Big]+\mu\Big[(a,b)\Big]\Big) \quad\to\quad \mu[(c,d)]\;,$$

für $a \searrow c$ und $b \nearrow d$.

(2). Ist die Verteilung μ absolutstetig mit quadratintegrierbarer Dichte f, dann ist auch die entsprechende charakteristische Funktion

$$\phi(t) \quad=\quad \int_{-\infty}^{\infty} e^{itx} f(x)\,dx$$

quadratintegrierbar. Die Aussage (2) aus Satz 4.14 ist in diesem Fall die klassische *Fourierinversionsformel der Analysis*, siehe z.B. Forster „Analysis 3“.

(3). Die Aussagen lassen sich auf Wahrscheinlichkeitsmaße auf $\mathbb{R}^d$ erweitern - auch diese sind durch ihre charakteristische Funktion eindeutig bestimmt.

Beweis von Satz 4.14. (1). Sei $T > 0$ und $a < b$. Nach dem Satz von Fubini können wir die Integrationsreihenfolge in dem folgendem Doppelintegral vertauschen, und erhalten:

$$\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita}\;-\;e^{-itb}}{it}\;\underbrace{\phi(t)}_{=\int e^{itx}\,\mu(dx)}\,dt \quad=\quad \frac{1}{\pi}\int\;\underbrace{\int_{-T}^{T}\frac{e^{it(x-a)}\;-\;e^{it(x-b)}}{2it}\,dt}_{=:\,g(T,x)}\,\mu(dx) \tag{4.3.2}$$

Dabei haben wir benutzt, dass der Integrand produktintegrierbar ist, da aus der Lipschitz-Stetigkeit der Abbildung $y \mapsto e^{iy}$ mit Konstante $L = 1$ folgt, dass

$$\left|\;\frac{e^{it(x-a)}\;-\;e^{it(x-b)}}{it}\;\right| \le \frac{|t\cdot(x-a)-t\cdot(x-b)|}{|t|} \;=\; |a-b| \qquad \text{gilt.}$$

Weiterhin erhalten wir, wegen $e^{it(x-a)} = \cos(t\cdot(x-a))+i\sin(t\cdot(x-a))$, $\cos(x) = \cos(-x)$ und $\sin(x) = -\sin(-x)$:

$$\begin{aligned} g(T,x) &= \int_0^T \frac{\sin(t\cdot(x-a))}{t}\,dt \;-\; \int_0^T \frac{\sin(t\cdot(x-b))}{t}\,dt \\ &= \int_0^{T\cdot(x-a)} \frac{\sin u}{u}\,du \;-\; \int_0^{T\cdot(x-b)} \frac{\sin u}{u}\,du \\ &= S(T\cdot(x-a)) \;-\; S(T\cdot(x-b)) \end{aligned}$$

wobei

$$S(t) \;:=\; \int_0^t \frac{\sin u}{u}\,du$$

der Integralsinus ist. Mithilfe des Residuensatzes (siehe Funktionentheorie) zeigt man:

$$\lim_{t\to\infty} S(t) \;=\; \frac{\pi}{2} \quad , \qquad \lim_{t\to-\infty} S(t) \;=\; -\frac{\pi}{2}\,.$$

Damit erhalten wir:

$$\lim_{T\to\infty} g(T,x) \;=\; \frac{\pi}{2}\,\mathrm{sgn}(x-a) \;-\; \frac{\pi}{2}\,\mathrm{sgn}(x-b) \;=\; \pi\cdot I_{(a,b)}(x) \;+\; \frac{\pi}{2}\cdot I_{\{a,b\}}(x)\,,$$

wobei wir $\mathrm{sgn}(0) := 0$ setzen. Da S beschränkt ist, ist auch $g(T,x)$ beschränkt in T und x. Nach dem Satz von Lebesgue folgt daher aus (4.3.2) für $T\to\infty$

$$\begin{aligned} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ita}-e^{-itb}}{it}\,\phi(t)\,dt \;&=\; \frac{1}{\pi}\int g(T,x)\,\mu(dx) \\ &\overset{T\to\infty}{\longrightarrow} \quad \mu[(a,b)] \;+\; \frac{1}{2}\,\mu[\{a,b\}]\,. \end{aligned}$$

(2). Ist ϕ integrierbar, dann ist die Funktion $(t,x)\mapsto e^{-itx}\,\phi(t)$ produktintegrierbar auf $[a,b]\times\mathbb{R}$ für alle $-\infty < a < b < \infty$. Also ist die Funktion

$$f(x) \;:=\; \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\,\phi(t)\,dt$$

integrierbar auf $[a,b]$, und es gilt nach dem Satz von Fubini und (1):

$$\int_a^b f(x)\,dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi(t)\,\underbrace{\int_a^b e^{-itx}\,dx}_{=\,\frac{e^{-ita}-e^{-itb}}{it}}\,dt \;\overset{(1)}{=}\; \frac{1}{2}\,\mu[\{a\}] + \mu[(a,b)] + \frac{1}{2}\,\mu[\{b\}]\,.$$

Insbesondere folgt

$$\int_{a+\varepsilon}^{b-\varepsilon} f(x)\,dx \;\le\; \mu\,[(a,b)] \;\le\; \int_a^b f(x)\,dx \quad \forall\,\varepsilon > 0,$$

also für $\varepsilon \searrow 0$:

$$\mu[(a,b)] \;=\; \int_a^b f(x)\,dx\;.$$

□

Beispiel (Summen von unabhängigen normalverteilten Zufallsvariablen). Sind X und Y unter P unabhängige Zufallsvariablen mit Verteilung $N(a,u)$ bzw. $N(b,v)$, dann hat $X+Y$ die charakteristische Funktion

$$\phi_{X+Y}(t) \;=\; \phi_X(t)\cdot\phi_Y(t) \;=\; \exp(i(a+b)t-(u+v)t^2/2).$$

Da die rechte Seite die charakteristische Funktion der Normalverteilung mit Mittel $a+b$ und Varianz $u+v$ ist, folgt

$$X+Y \;\sim\; N(a+b,u+v).$$

Insbesondere ist die Verteilung $N(a+b,u+v)$ also die Faltung der Normalverteilungen $N(a,u)$ und $N(b,v)$.

Das Argument aus dem Beispiel ist auch allgemein anwendbar: Da die Faltung $\mu*\nu$ von Wahrscheinlichkeitsverteilungen μ und ν auf $\mathbb{R}$ die Verteilung der Summe unabhängiger Zufallsvariablen $X\sim\mu$ und $Y\sim\nu$ ist, gilt für die charakteristischen Funktionen:

$$\phi_{\mu*\nu}(t) \;=\; \phi_\mu(t)\cdot\phi_\nu(t) \qquad \text{für alle } t\in\mathbb{R}.$$

4.4 Empirische Verteilungen

4.4.1 Schätzen von Kenngrößen einer unbekannten Verteilung

Angenommen, wir haben eine Stichprobe aus reellen Beobachtungswerten $X_1, X_2, \ldots, X_n$ gegeben, und möchten die zugrundeliegende Wahrscheinlichkeitsverteilung μ auf $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ möglichst weitgehend rekonstruieren. Im einfachsten Modell interpretieren wir die Beobachtungswerte als Realisierungen unabhängiger Zufallsvariablen $X_1, X_2, \ldots$ mit Verteilung μ.

(1). SCHÄTZEN DES ERWARTUNGSWERTES: Sei $\int |x|\,\mu(dx) < \infty$. Um den Erwartungswert

$$m \quad = \quad \int x\,\mu(dx)$$

zu schätzen, verwenden wir das **empirische Mittel**

$$\overline{X}_n \quad := \quad \frac{1}{n}\sum_{i=1}^n X_i.$$

Das empirische Mittel ist ein *erwartungstreuer Schätzer* für m, d.h. $\overline{X}_n$ ist eine Funktion von den Beobachtungswerten $X_1, \ldots, X_n$ mit $E[\overline{X}_n] = m$. Obere Schranken für den Schätzfehler $P[|\overline{X}_n - m| > \varepsilon], \varepsilon > 0$, erhält man z.B. mithilfe der Čebyšev- oder der exponentiellen Markov-Ungleichung. Für $n \to \infty$ gilt nach dem Gesetz der großen Zahlen

$$\overline{X}_n \quad \longrightarrow \quad m \qquad P\text{-fast sicher},$$

d.h. $\overline{X}_n$ ist eine *konsistente* Folge von Schätzern für m.

(2). SCHÄTZEN DER VARIANZ: Um die Varianz

$$v \quad = \quad \int (x-m)^2 \, \mu(dx)$$

der zugrundeliegenden Verteilung zu schätzen, verwendet man meistens die **renormierte Stichprobenvarianz**

$$\widetilde{V}_n \quad = \quad \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

Der Vorfaktor $\frac{1}{n-1}$ (statt $\frac{1}{n}$) gewährleistet unter anderem, dass $\widetilde{V}_n$ ein *erwartungstreuer* Schätzer für v ist, denn aus

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \quad &= \quad \frac{1}{n} \sum_{i=1}^{n} (X_i - m)^2 - (\overline{X}_n - m)^2 \qquad (4.4.1) \\ \text{Stichprobenvarianz} \quad &= \quad \text{MSE} \quad - \quad \text{Stichprobenbias}^2 \end{aligned}$$

folgt

$$E\left[\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2\right] \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}[X_i] - \mathrm{Var}[\overline{X}_n] \quad = \quad \frac{n-1}{n} v,$$

also $E[\widetilde{V}_n] = v$.

Um zu zeigen, dass $\widetilde{V}_n$ eine konsistente Folge von Schätzern für v ist, können wir erneut das Gesetz der großen Zahlen anwenden. Da die Zufallsvariablen $X_i - \overline{X}_n, 1 \leq i \leq n$, selbst nicht unabhängig sind, verwenden wir dazu die Zerlegung (4.4.1). Nach dem starken Gesetz der großen Zahlen für nichtnegative Zufallsvariablen erhalten wir

$$\frac{n-1}{n} \widetilde{V}_n \quad = \quad \frac{1}{n} \sum_{i=1}^{n} (X_i - m)^2 - (\overline{X}_n - m)^2 \quad \longrightarrow \quad v \qquad P\text{-fast sicher},$$

also auch $\widetilde{V}_n \to v$ P-fast sicher.

(3). SCHÄTZEN VON INTEGRALEN: Allgemeiner können wir für jede Funktion $f \in \mathcal{L}^1(S, \mathcal{S}, \mu)$ das Integral

$$\theta \quad = \quad \int f \, d\mu$$

erwartungstreu durch die **empirischen Mittelwerte**

$$\widehat{\theta}_n \quad = \quad \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

schätzen. Dies haben wir schon in Kapitel **??** für Monte Carlo Verfahren verwendet. Da die Zufallsvariablen $f(X_i)$ wieder unabhängig und identisch verteilt sind mit Erwartungswert θ, gilt nach dem starken Gesetz der großen Zahlen:

$$\widehat{\theta}_n \quad \longrightarrow \quad \theta \qquad P\text{-fast sicher.} \tag{4.4.2}$$

(4). SCHÄTZEN DER VERTEILUNG: Die gesamte Verteilung μ können wir durch die **empirische Verteilung**

$$\widehat{\mu}_n(\omega) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}$$

der Zufallsstichprobe schätzen. $\widehat{\mu}_n$ ist eine *„zufällige Wahrscheinlichkeitsverteilung,“* d.h. eine Zufallsvariable mit Werten im Raum $WV(\mathbb{R})$ der Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Aus (4.4.2) ergibt sich die folgende Approximationseigenschaft der empirischen Verteilungen:

$$\int f \, d\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \quad \stackrel{n\to\infty}{\longrightarrow} \quad \int f \, d\mu \tag{4.4.3}$$

P-fast sicher für alle $f \in \mathcal{L}^1(S, \mathcal{S}, \mu)$.

4.4.2 Konvergenz der empirischen Verteilungsfunktionen

Für die *empirischen Verteilungsfunktionen*

$$F_n(c) \quad = \quad \widehat{\mu}_n[(-\infty, c]] \quad = \quad \frac{1}{n} \, |\{1 \le i \le n \, : \, X_i \le c\}|$$

von unabhängigen, identisch verteilten, reellwertigen Zufallsvariablen $X_1, X_2, \ldots$ mit Verteilungsfunktion F ergibt sich wegen $F_n(c) = \int I_{(-\infty,c]} \, d\widehat{\mu}_n$:

$$\lim_{n\to\infty} F_n(c) \quad = \quad F(c) \qquad P\text{-fast sicher für alle } c \in \mathbb{R}. \tag{4.4.4}$$

Diese Aussage kann man noch etwas verschärfen:

Satz 4.15 (Glivenko-Cantelli). *Sind $X_1, X_2, \ldots$ unabhängig und identisch verteilt mit Verteilungsfunktion F, dann gilt für die empirischen Verteilungsfunktionen F_n:*

$$\sup_{c\in\mathbb{R}} |F_n(c) - F(c)| \quad \longrightarrow \quad 0 \qquad P\text{-fast sicher.} \tag{4.4.5}$$

Beweis. Wir führen den Beweis unter der zusätzlichen Annahme, dass F stetig ist – für den allgemeinen Fall siehe z.B. *Klenke: Wahrscheinlichkeitstheorie*. Sie $\varepsilon > 0$ gegeben. Ist F stetig, dann existieren $k \in \mathbb{N}$ und Konstanten

$$-\infty \;=\; c_0 \;<\; c_1 \;<\; c_2 \;<\; \ldots \;<\; c_k \;=\; \infty \qquad \text{mit } F(c_i) - F(c_{i-1}) \le \frac{\varepsilon}{2}$$

für alle $1 \le i \le k$. Da F_n nach 4.4.4 mit Wahrscheinlichkeit 1 punktweise gegen F konvergiert, existiert zudem ein $n_0 \in \mathbb{N}$ mit

$$\max_{0\le i\le n} |F_n(c_i) - F(c_i)| \quad < \quad \frac{\varepsilon}{2} \qquad \text{für alle } n \ge n_0.$$

Wegen der Monotonie der Verteilungsfunktionen folgt dann

$$F_n(c) - F(c) \quad \le \quad F_n(c_i) - F(c_{i-1}) \quad \le \quad \frac{\varepsilon}{2} + F_n(c_i) - F(c_i) \quad < \quad \varepsilon,$$

und entsprechend

$$F(c) - F_n(c) \quad \le \quad F(c_i) - F_n(c_{i-1}) \quad \le \quad \frac{\varepsilon}{2} + F(c_i) - F_n(c_i) \quad < \quad \varepsilon,$$

für alle $n \ge n_0, c \in \mathbb{R}$, und $1 \le i \le k$ mit $c_{i-1} \le c \le c_i$. Also gilt auch

$$\sup_{c\in\mathbb{R}} |F_n(c) - F(c)| \quad < \quad \varepsilon \qquad \text{für alle } n \ge n_0.$$

□

Bemerkung (QQ-Plot). In parametrischen statistischen Modellen nimmt man von vornherein an, dass die beobachteten Daten Realisierungen von Zufallsvariablen sind, deren Verteilung aus einer bestimmten Familie von Wahrscheinlichkeitsverteilungen stammt, z.B. der Familie aller Normalverteilungen. Um zu entscheiden, ob eine solche Annahme für gegebene reellwertige Daten $x_1, \ldots, x_n$ gerechtfertigt ist, kann man die empirische Verteilungsfunktion mit der tatsächlichen Verteilungsfunktion vergleichen. Ein praktikables graphisches Verfahren ist der Quantil-Quantil-Plot, bei dem die Quantile der empirischen und der theoretischen Verteilung gegeneinander aufgetragen werden. Um auf Normalverteilung zu testen, plottet man beispielsweise die Punkte

$$\left(\Phi^{-1}\left(\frac{k-\frac{1}{2}}{n}\right), x_{(k)}\right), \qquad k = 1, 2, \ldots, n,$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist, und

$$x_{(1)} \quad \leq \quad x_{(2)} \quad \leq \quad \dots \quad \leq \quad x_{(n)}$$

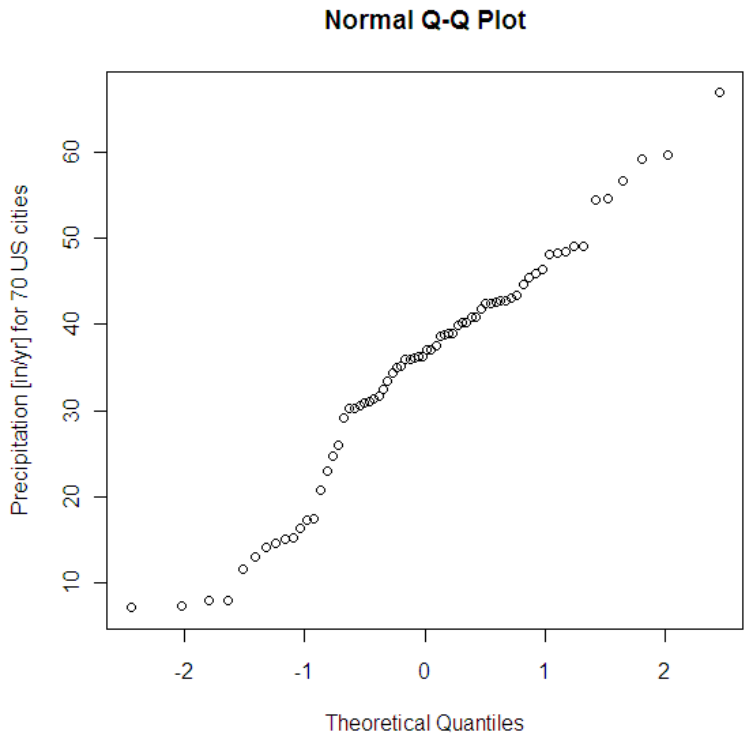
die Ordnungsstatistiken von $x_1, \dots, x_n$, also die $(k - \frac{1}{2})/n$-Quantile der empirischen Verteilung sind. Ist die zugrundeliegende Verteilung eine Normalverteilung mit Mittel m und Standardabweichung σ, dann liegen die Punkte für große n näherungsweise auf einer Geraden mit Steigung σ und Achsenabschnitt m, da für die Verteilungsfunktion und die Quantile der theoretischen Verteilung dann

$$F(c) \quad = \quad P[X \leq c] \quad = \quad P[\sigma Z + m \leq c] \quad = \quad P\left[Z \leq \frac{c-m}{\sigma}\right] \quad = \quad \Phi\left(\frac{c-m}{\sigma}\right),$$

bzw.

$$F^{-1}(u) \quad = \quad m + \sigma \Phi^{-1}(u)$$

gilt. Die folgende Grafik zeigt einen QQ-Plot bzgl. der Standardnormalverteilung.



4.4.3 Histogramme und Multinomialverteilung

Die empirische Verteilung $\widehat{\mu}_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}$ von Zufallsvariablen $X_1, \dots, X_n$ ist selbst eine Zufallsvariable mit Werten im Raum der Wahrscheinlichkeitsverteilungen. Wir wollen nun

die Verteilung dieser Zufallsvariablen explizit berechnen, falls die X_i unabhängig und identisch verteilt mit endlichem Wertebereich S sind. Haben die Zufallsvariablen keinen endlichen Wertebereich, dann kann man die Aussagen trotzdem anwenden, indem man den Wertebereich in endlich viele Teilmengen (Klassen) zerlegt.

Das *Histogramm* von n Beobachtungswerten $x_1, \ldots, x_n$, die in einer endlichen Menge S liegen, ist der Vektor

$$\vec{h} = (h_a)_{a \in S}, \qquad h_a = |\{1 \leq i \leq n | x_i = a\}|,$$

der Häufigkeiten der möglichen Werte $a \in S$ unter $x_1, \ldots, x_n$. Graphisch stellt man ein Histogramm durch ein Balkendiagramm dar:

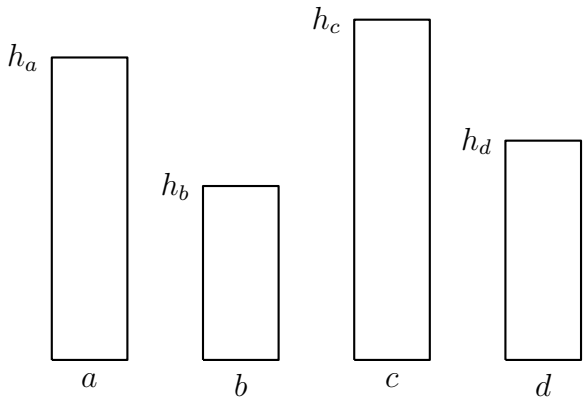


Abbildung 4.7: Histogramm der Klassen a, b, c und d mit den jeweiligen Häufigkeiten h_a, h_b, h_c und h_d

Der Raum Hist(n, S) aller möglichen Histogramme von n Beobachtungswerten ist eine Teilmenge von $\{0, 1, \ldots, n\}^S$:

$$\text{Hist}(n, S) \quad = \quad \{\vec{h} = (h_a)_{a \in S} | h_a \in \mathbb{Z}_+, \sum_{a \in S} h_a = n\} \quad \subseteq \quad \{0, 1, \ldots, n\}^S.$$

Sei nun μ eine Wahrscheinlichkeitsverteilung auf der endlichen Menge S. Wir wollen die Verteilung des Histogrammvektors bestimmen, wenn die Beobachtungswerte unabhängige Stichproben von der Verteilung μ sind. Wir betrachten also unabhängige Zufallsvariablen $X_1, \ldots, X_n$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathscr{A}, P)$ mit Verteilung μ und die Häufigkeiten

$$H_a(\omega) \quad := \quad |\{1 \leq i \leq n : X_i(\omega) = a\}|$$

der möglichen Werte $a \in S$. Die Zufallsvariable H_a ist $\mathrm{Bin}(n,p)$-verteilt mit $p = \mu[\{a\}]$. Wir berechnen nun die **gemeinsame Verteilung** aller dieser Häufigkeiten, d.h. die Verteilung μ_H des Zufallsvektors

$$H \quad = \quad (H_a)_{a\in S} \quad : \quad \Omega \longrightarrow \text{Hist}(n,S)$$

mit Werten im Raum der Histogramme. Dazu verwenden wir die Unabhängigkeit der X_i. Mit $I = \{1,\ldots,n\}$ erhalten wir:

$$\begin{aligned}
\mu_H(\vec{k}) &= P[H_a = k_a \quad \forall a \in S] \\
&= P[X_i = a \text{ genau } k_a\text{-mal für alle } a \in S] \\
&= \sum_{\substack{I = \dot{\bigcup}_{a\in S} I_a \\ |I_a| = k_a}} P[X_i = a \quad \forall\, i \in I_a \quad \forall\, a \in S] \\
&= \sum_{\substack{I = \dot{\bigcup}_{a\in S} I_a \\ |I_a| = k_a}} \prod_{a\in S} \mu[\{a\}]^{k_a} \\
&= \binom{n}{\vec{k}} \prod_{a\in S} \mu[\{a\}]^{k_a}.
\end{aligned}$$

Hierbei laufen die Summen über alle disjunkten Zerlegungen von $I = \{0,1,\ldots,n\}$ in Teilmengen $i_a, a \in S$, mit jeweils k_a Elementen, und der **Multinomialkoeffizient**

$$\binom{n}{\vec{k}} \quad := \quad \frac{n!}{\prod\limits_{a\in S} k_a!}, \qquad k_a \in \{0,1,\ldots,n\} \text{ mit } \sum_{a\in S} k_a = n,$$

gibt die Anzahl der Partitionen von n Elementen in Teilmengen von jeweils k_a Elementen an.

Definition. *Die Verteilung des Histogrammvektors H heißt **Multinomialverteilung für n Stichproben mit Ergebniswahrscheinlichkeiten** $\mu(a), a \in S$.*

Bemerkung. Im Fall $|S| = 2$ ist $H(\omega)$ eindeutig festgelegt durch $H_1(\omega)$, und die Zufallsvariable H_1 ist binomialverteilt mit Parametern n und $p = \mu[\{1\}]$. In diesem Sinn ergibt sich die Binomialverteilung als Spezialfall der Multinomialverteilung.

Kapitel 5

Zentrale Grenzwertsätze

Seien $X_1, X_2, ... \in \mathscr{L}^2(\Omega, \mathcal{A}, P)$ unabhängige und identisch verteilte Zufallsvariablen mit $E[X_i] = 0$ für alle i, und sei $S_n = X_1 + ... + X_n$. Nach dem Gesetz der großen Zahlen gilt

$$\frac{S_n}{n} \quad \to \quad 0 \qquad P\text{-fast sicher und in } L^2(\Omega, \mathcal{A}, P).$$

Wie sieht die Verteilung von S_n für große n aus?

Um eine asymptotische Darstellung zu erhalten, reskalieren wir zunächst so, dass die Varianz konstant ist. Es gilt

$$\mathrm{Var}[S_n] \;=\; n \cdot \mathrm{Var}[X_1],$$

also ist

$$\mathrm{Var}\left[\frac{S_n}{\sqrt{n}}\right] \;=\; \frac{1}{n} \cdot \mathrm{Var}[S_n] \;=\; \mathrm{Var}[X_1] \;=:\; \sigma^2$$

unabhängig von n.

Um die Asymptotik der Verteilungen der entsprechend standardisierten Summen $S_n/\sqrt{n}$ zu bestimmen, betrachten wir die charakteristischen Funktionen. Da die Summanden X_i unabhängig und identisch verteilt sind, erhalten wir

$$\phi_{\frac{S_n}{\sqrt{n}}}(t) \;=\; \phi_{S_n}\left(\frac{t}{\sqrt{n}}\right) \;\overset{X_i \text{ iid}}{=}\; \left[\phi_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n .$$

Wegen $X_1 \in \mathscr{L}^2(\Omega, \mathcal{A}, P)$ ist ϕ_{X_1} zweimal stetig differenzierbar, und die Taylorentwicklung bei $t = 0$ ist gegeben durch

$$\phi_{X_1}(t) \quad = \quad 1 + i \cdot E[X_1] \cdot t - \frac{1}{2}\, E[X_1^2] \cdot t^2 + o(t^2) \quad = \quad 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2).$$

Damit folgt für $t \in \mathbb{R}$:

$$\phi_{\frac{S_n}{\sqrt{n}}}(t) \;=\; \left(1 - \frac{\sigma^2 t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \;\overset{n\nearrow\infty}{\longrightarrow}\; \exp\left(-\frac{\sigma^2 t^2}{2}\right) \;=\; \phi_{N(0,\sigma^2)}(t).$$

Wir werden im nächsten Abschnitt zeigen, dass aus der Konvergenz der charakteristischen Funktionen unter geeigneten Voraussetzungen die schwache Konvergenz (Definition s.u.) der Verteilungen folgt. Somit ergibt sich:

Zentraler Grenzwertsatz. Die Verteilungen der standardisierten Summen $S_n/\sqrt{n}$ konvergieren schwach gegen die Normalverteilung $N(0, \sigma^2)$.

Den detaillierten Beweis werden wir in Abschnitt 5.2 führen. Der zentrale Grenzwertsatz erklärt, warum die Normalverteilungen in der Stochastik von so großer Bedeutung sind:

Bemerkung (Universalität der Normalverteilung). Die *Limesverteilung im zentralen Grenzwertsatz ist unabhängig von der Verteilung von* X_1, vorausgesetzt, es gilt $X_1 \in \mathscr{L}^2(\Omega, \mathcal{A}, P)$.

5.1 Verteilungskonvergenz

Sei S ein metrischer Raum mit Borelscher σ-Algebra $\mathcal{B}(S)$, zum Beispiel $S = \mathbb{R}$ oder $S = \mathbb{R}^d$. Wir wollen nun einen für den zentralen Grenzwertsatz angemessenen Konvergenzbegriff für die Verteilungen μ_n einer Folge Y_n von Zufallsvariablen mit Werten in S einführen. Naheliegend wäre es zu definieren, dass eine Folge μ_n von Wahrscheinlichkeitsverteilungen auf $(S, \mathcal{B}(S))$ gegen eine Wahrscheinlichkeitsverteilung μ konvergiert, wenn $\mu[A] = \lim \mu_n[A]$ für *jede* Menge $A \in \mathcal{B}(S)$ gilt. Ein solcher Konvergenzbegriff erweist sich jedoch sofort als zu restriktiv, z.B. würde eine Folge von diskreten Wahrscheinlichkeitsverteilungen in diesem Sinne niemals gegen eine Normalverteilung konvergieren. Einen angemesseneren Grenzwertbegriff erhält man durch Berücksichtigung der Topologie auf S:

Definition. *(1). **Schwache Konvergenz von Wahrscheinlichkeitsmaßen**: Eine Folge $(\mu_n)_{n\in\mathbb{N}}$ von Wahrscheinlichkeitsmaßen auf der Borelscher σ-Algebra $\mathcal{B}(S)$ **konvergiert schwach** gegen ein Wahrscheinlichkeitsmaß μ auf $\mathcal{B}(S)$ ($\mu_n \xrightarrow{w} \mu$), falls*

$$\int f \, d\mu_n \quad \longrightarrow \quad \int f \, d\mu \qquad \text{für alle stetigen, beschränkten } f : S \to \mathbb{R} \text{ gilt.}$$

*(2). **Konvergenz in Verteilung von Zufallsvariablen**: Eine Folge $(Y_n)_{n\in\mathbb{N}}$ von Zufallsvariablen mit Werten in S **konvergiert in Verteilung** gegen eine Zufallsvariable Y bzw. gegen die Verteilung von Y, falls*

$$\text{Verteilung}\,(Y_n) \quad \xrightarrow{w} \quad \text{Verteilung}\,(Y),$$

d.h. falls

$$E[f(Y_n)] \quad \longrightarrow \quad E[f(Y)] \qquad \text{für alle } f \in C_b(S) \text{ gilt.}$$

Konvergenz in Verteilung bezeichnet man auf Englisch als *„convergence in distribution“* oder *„convergence in law“*. Entsprechend verwendet man die Kurzschreibweisen $Y_n \xrightarrow{\mathcal{D}} Y$ oder $Y_n \xrightarrow{\mathcal{L}} Y$, falls Y_n in Verteilung gegen Y konvergiert.

Bemerkung. (1). Die Zufallsvariablen $Y_n, n \in \mathbb{N},$ und Y können bei der Verteilungskonvergenz *auf verschiedenen Wahrscheinlichkeitsräumen* definiert sein!

(2). Die hier definierte Form der schwachen Konvergenz entspricht **nicht** der im funktionalanalytischen Sinn definierten schwachen Konvergenz auf dem Vektorraum aller beschränkten signierten Maße auf $(S, \mathcal{B}(S))$, sondern einer schwach$*$-Konvergenz auf diesem Raum, siehe z.B. ALT: LINEARE FUNKTIONALANALYSIS.

(3). Um zu garantieren, dass die Grenzwerte bzgl. schwacher Konvergenz eindeutig sind, benötigt man eine zusätzliche Annahme an den Raum S. Zum Beispiel ist dies der Fall, wenn S ein vollständiger, separabler metrischer Raum ist.

(4). Wir werden in Satz 5.3 zeigen, dass im Fall $S = \mathbb{R}$ die Folge μ_n genau dann schwach gegen μ konvergiert, wenn für die Verteilungsfunktionen

$$F_{\mu_n}(x) \quad \longrightarrow \quad F_\mu(x) \qquad \text{für alle Stetigkeitsstellen } x \text{ von } F,$$

d.h. für alle $x \in \mathbb{R}$ mit $\mu[\{x\}] = 0$, gilt.

Das folgende Beispiel zeigt, dass die schwache Konvergenz von Wahrscheinlichkeitsmaßen auf S mit dem Konvergenzbegriff auf S konsistent ist:

Beispiel (Schwache Konvergenz von Dirac-Maßen). Ist $(x_n)_{n\in\mathbb{N}}$ eine Folge in S, dann konvergiert die Folge der Dirac-Maße δ_{x_n} genau dann schwach, wenn x_n in S konvergiert. Der Grenzwert von δ_{x_n} ist in diesem Fall das Dirac-Maß δ_x an der Stelle $x = \lim x_n$.

5.1.1 Schwache Konvergenz von Wahrscheinlichkeitsmaßen

Das Beispiel der Dirac-Maße zeigt auch, dass bei schwacher Konvergenz die Wahrscheinlichkeiten beliebiger Mengen nicht unbedingt konvergieren. Ist zum Beispiel $A \subset S$ eine abgeschlossene Menge, die den Grenzwert x einer Folge $x_n \in S$ enthält, aber nicht die Folgenglieder selbst, dann gilt

$$\lim \delta_{x_n}[A] \;=\; 0 \;<\; 1 \;=\; \delta_x[A].$$

Umgekehrt gilt für eine offene Menge $O \subset S$, die die Folgenglieder aber nicht den Grenzwert enthält:

$$\lim \delta_{x_n}[O] \;=\; 1 \;>\; 0 \;=\; \delta_x[O].$$

Der folgende Satz liefert eine Charakteriserung der schwachen Konvergenz von Wahrscheinlichkeitsmaßen über die Konvergenz der Wahrscheinlichkeiten von Mengen:

Satz 5.1 (**Portemanteau-Theorem**). *Seien μ_n ($n \in \mathbb{N}$) und μ Wahrscheinlichkeitsmaße auf der Borelschen σ-Algebra über einem metrischen Raum S. Dann sind folgende Aussagen äquivalent:*

(1). Die Folge $(\mu_n)_{n\in\mathbb{N}}$ konvergiert schwach gegen μ.

(2). Für jede beschränkte Lipschitz-stetige Funktion $f : S \to \mathbb{R}$ konvergiert $\int f\, d\mu_n$ gegen $\int f\, d\mu$.

(3). Für jede abgeschlossene Teilmenge $A \subseteq S$ gilt

$$\limsup \mu_n[A] \;\leq\; \mu[A].$$

(4). Für jede offene Teilmenge $O \subseteq S$ gilt

$$\liminf \mu_n[O] \;\geq\; \mu[O].$$

(5). Für jede Menge $B \in \mathcal{B}(S)$ mit $\mu[\partial B] = 0$ gilt

$$\lim \mu_n[A] \;=\; \mu[A].$$

Beweis. Die Implikation „(1) $\Rightarrow$ (2)“ ist offensichtlich wahr.

„(2) $\Rightarrow$ (3)“: Sei $A \subseteq S$ abgeschlossen. Wir approximieren die Indikatorfunktion der Menge A durch die beschränkten Lipschitz-stetigen Funktionen

$$f_\varepsilon(x) \;:=\; (1 - \operatorname{dist}(x, A)/\varepsilon)^+ .$$

Nach (2) gilt für jedes $\varepsilon > 0$:

$$\limsup \mu_n[A] \;\leq\; \limsup \int f_\varepsilon\, d\mu_n \;=\; \int f_\varepsilon\, d\mu.$$

Hieraus folgt (3), da die rechte Seite für $\varepsilon \downarrow 0$ gegen $\mu[A]$ konvergiert.

„(3) $\Leftrightarrow$ (4)“: Durch Anwenden der Aussagen (3) bzw. (4) auf das Komplement einer Menge B sieht man, dass (3) und (4) äquivalent sind, da $\mu_n[B^C] = 1 - \mu_n[B]$ und $\mu[B^C] = 1 - \mu[B]$ gilt. Man beachte, dass wir hier benutzt haben, dass alle Maße Wahrscheinlichkeitsmaße sind, also dieselbe Gesamtmasse haben !

„(3) und (4) $\Leftrightarrow$ (5)“: Aus der Kombination der (äquivalenten) Aussagen (3) und (4) folgt (5). Ist

$B \subseteq S$ nämlich eine Borel-Menge mit $\mu[\partial B] = 0$, dann stimmt das Maß von B mit den Maßen der abgeschlossenen Hülle $B \cup \partial B$ und des offenen Kerns $B \setminus \partial B$ von B überein. Damit folgt aus (3) und (4):

$$\limsup \mu_n[B] \;\leq\; \limsup \mu_n[B \cup \partial B] \;\leq\; \mu[B] \leq\; \liminf \mu_n[B \setminus \partial B] \;\leq\; \liminf \mu_n[B],$$

also $\liminf \mu_n[B] = \limsup \mu_n[B] = \mu[B]$.

„(5) $\Leftrightarrow$ (1)“: Sei $f \in C_b(S)$ und $\varepsilon > 0$. Um von der Konvergenz der Maße von μ-randlosen Mengen auf die Konvergenz von $\int f\, d\mu_n$ zu schließen, bemerken wir, dass $\mu[f = c] > 0$ nur für abzählbar viele $c \in \mathbb{R}$ gilt. Da f beschränkt ist, können wir endlich viele reelle Zahlen $c_0 < c_1 < \ldots < c_k$ mit $c_0 < \inf f$ und $c_k > \sup f$ finden, so dass $c_{i+1} - c_i < \varepsilon$ und $\mu[f = c_i] = 0$ für alle i gilt. Da f zudem stetig ist, sind die Ränder der Mengen $f^{-1}([c_{i-1}, c_i])$ in $\bigcup_i \{f = c_i\}$ enthalten, und haben daher Maß Null bzgl. μ. Somit folgt aus (5):

$$\begin{aligned}\limsup \int f\, d\mu_n \;&\leq\; \limsup \sum_{i=1}^{k} c_i\, \mu_n\left[f^{-1}([c_{i-1}, c_i])\right] \;=\; \sum_{i=1}^{k} c_i\, \mu\left[f^{-1}([c_{i-1}, c_i])\right] \\ &\leq\; \sum_{i=1}^{k} (\varepsilon + c_{i-1})\, \mu\left[f^{-1}((c_{i-1}, c_i))\right] \;\leq\; \varepsilon + \int f\, d\mu.\end{aligned}$$

Für $\varepsilon \downarrow 0$ folgt $\limsup \int f\, d\mu_n \leq \int f\, d\mu$, und, durch Anwenden dieser Aussage auf $-f$, auch $\liminf \int f\, d\mu_n \geq \int f\, d\mu$, also $\int f\, d\mu_n \to \int f\, d\mu$. □

Neben schwacher Konvergenz betrachtet man häufig auch unter anderem die folgenden Konvergenzarten auf positiven bzw. beschränkten signierten Maßen:

- **Vage Konvergenz:** Die Folge μ_n *konvergiert vage gegen* μ, falls

  $$\int f\, d\mu_n \quad\longrightarrow\quad \int f\, d\mu$$

  für alle stetigen Funktionen f mit kompaktem Träger gilt.

- **Konvergenz in Variationsdistanz:** μ_n *konvergiert in (totaler) Variation gegen* μ, falls die *Variationsnormen*

  $$\|\mu - \mu_n\|_{\mathrm{TV}} \quad := \quad \sup_{B \in \mathcal{B}(S)} |\mu[B] - \mu_n[B]|$$

  der signierten Maße $\mu - \mu_n$ gegen Null konvergieren. Dies ist eine sehr starke Konvergenzform: Die Wahrscheinlichkeiten aller Borel-Mengen konvergieren gleichmäßig. Die

Variationsdistanz zweier Wahrscheinlichkeitsmaße μ und ν lässt sich auch wie folgt darstellen:

$$\|\mu - \nu\|_{\text{TV}} \quad = \quad \frac{1}{2} \sup_{\substack{f:S\to\mathbb{R} \text{ messbar} \\ \text{mit } |f| \le 1}} \left| \int f \, d\mu - \int f \, d\nu \right|.$$

Im diskreten Fall gilt

$$\|\mu - \nu\|_{\text{TV}} \quad = \quad \frac{1}{2} \sum_{x \in S} |\mu[\{x\}] - \nu[\{x\}]| \,.$$

Diesen Abstandsbegriff haben wir bereits in Abschnitt **??** bei der Konvergenz ins Gleichgewicht von Markov-Ketten verwendet.

- **Konvergenz in Kantorovich-Distanz:** Die *Kantorovich-Distanz* zweier Wahrscheinlichkeitsmaße μ und ν auf S ist definiert durch

$$d_K(\mu, \nu) \; = \; \inf_{\eta \in \Pi(\mu,\nu)} \int d(x,y) \; \eta(dx \, dy).$$

Hierbei bezeichnet $\Pi(\mu, \nu)$ die Menge aller **Kopplungen** von μ und ν, d.h. aller Wahrscheinlichkeitsmaße $\eta(dx\,dy)$ auf dem Produktraum $S \times S$ mit Randverteilungen $\mu(dx)$ und $\nu(dy)$. Beispielsweise ist das Produkt-Maß $\mu \otimes \nu$ eine Kopplung von μ und ν. Die *Kantorovich-Rubinstein-Dualität* besagt, dass sich die Kantorovich-Distanz auch als

$$d_K(\mu, \nu) \; = \; \sup_{\substack{f:S\to\mathbb{R} \text{ mit} \\ |f(x) - f(y)| \le d(x,y)}} \left| \int f \, d\mu - \int f \, d\nu \right|,$$

darstellen lässt, siehe z.B. VILLANI: OPTIMAL TRANSPORT. Die Variationsdistanz ist ein Spezialfall der Kantorovich-Distanz: Wählt man auf S die Metrik $d(x, y) = I_{\{x \neq y\}}$, dann gilt $d_K(\mu, \nu) = \|\mu - \nu\|_{\text{TV}}$. Die Kantorovich-Distanz wird häufig auch als *Kantorovich-Rubinstein-* oder L^1*-Wasserstein-Distanz* bezeichnet.

Offensichtlich folgt aus der Konvergenz in Variationsdistanz die schwache Konvergenz, aus der wiederum die vage Konvergenz folgt:

$$\|\mu_n - \mu\|_{\text{TV}} \to 0 \quad \Longrightarrow \quad \mu_n \xrightarrow{w} \mu \quad \Longrightarrow \quad \mu_n \to \mu \quad \text{vage}.$$

Auch aus der Konvergenz in Kantorovich-Distanz folgt schwache Konvergenz:

$$d_K(\mu_n, \mu) \to 0 \quad \Longrightarrow \quad \mu_n \xrightarrow{w} \mu.$$

Ist S beschränkt bzgl. der Metrik d, dann ist die schwache Konvergenz sogar äquivalent zur Konvergenz in Kantorovich-Distanz, siehe VILLANI: OPTIMAL TRANSPORT.

Die folgenden Beispiele verdeutlichen die unterschiedlichen Konvergenzbegriffe:

Beispiele (Vergleich der Konvergenzbegriffe).

(1). ***Diracmaße:*** Aus $x_n \to x$ folgt keine Konvergenz von δ_{x_n} gegen δ_x in Variationsnorm, denn $\|\delta_{x_n} - \delta_x\|_{\mathrm{TV}} = 1$ für $x_n \neq x$. Hingegen gilt

$$d_K(\delta_x, \delta_{x_n}) \;=\; d(x, x_n) \;\to\; 0.$$

(2). ***Degeneration/Diracfolge:*** Auf $S = \mathbb{R}^1$ konvergiert die Folge $\mu_n := N(0, \frac{1}{n})$ von Normalverteilungen mit degenerierender Varianz schwach gegen das Dirac-Maß δ_0, denn mit dem Satz von Lebesgue folgt für $f \in C_b(\mathbb{R})$:

$$\begin{aligned}
\int f\, d\mu_n \quad &= \quad \int f(x)\, \frac{1}{\sqrt{2\pi/n}}\, e^{-\frac{x^2}{2/n}}\, dx \\
&\overset{y=\sqrt{n}x}{=} \quad \int f\left(\frac{y}{\sqrt{n}}\right) \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}}\, dy \\
&\overset{Lebesgue}{\longrightarrow} \quad f(0) \cdot \underbrace{\int \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}}\, dy}_{=1} \;=\; \int f\, d\delta_0.
\end{aligned}$$

In diesem Beispiel gilt auch Konvergenz in Kantorovich-Distanz, denn

$$d_K(\mu_n, \delta_0) \;=\; \int |x|\, N(0, 1/n)(dx) \;\leq\; \sqrt{1/n} \;\to\; 0.$$

Hingegen gilt wiederum $\|\mu_n - \delta_0\|_{\mathrm{TV}} = 1$ für alle n.

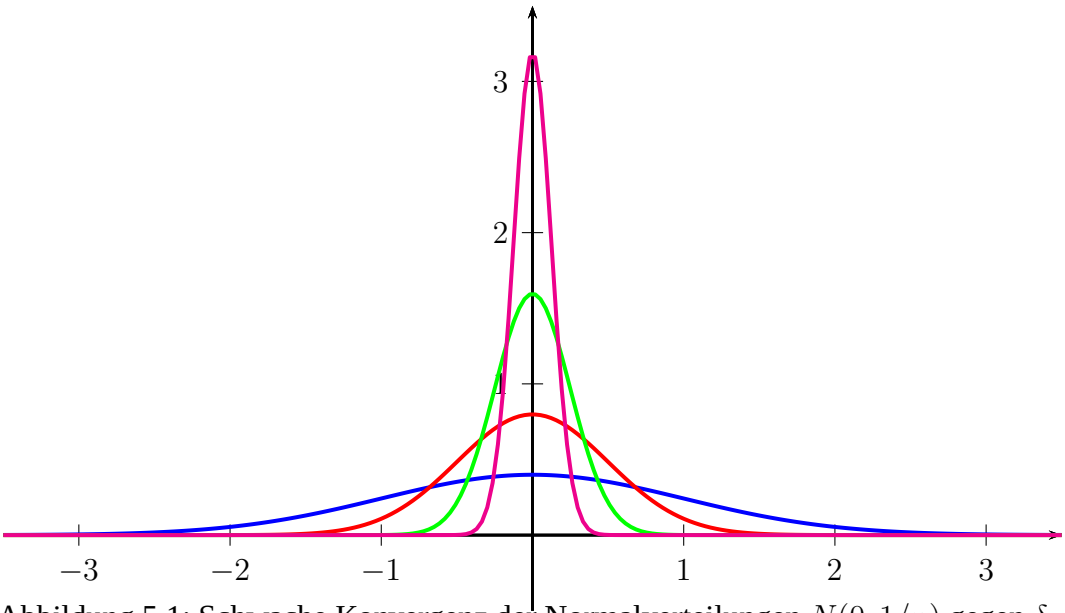


Abbildung 5.1: Schwache Konvergenz der Normalverteilungen $N(0, 1/n)$ gegen δ_0.

(3). ***Schwache vs. vage Konvergenz:*** Die Folge $\mu_n = N(0,n)$ konvergiert vage gegen das Nullmaß μ mit $\mu[A] = 0$ für alle A. In der Tat gilt für $f \in C(\mathbb{R})$ mit $f(x) = 0$ für $x \notin [-K, K]$:

$$\left|\int f\, d\mu_n\right| \quad = \quad \left|\int_{-K}^{K} f(x) \cdot \frac{1}{\sqrt{2\pi n}} e^{-x^2/2n} dx\right| \quad \leq \quad \frac{2K}{\sqrt{2\pi n}} \cdot \sup|f| \quad \stackrel{n\to\infty}{\longrightarrow} \quad 0.$$

Es gilt aber keine schwache Konvergenz, da

$$\int 1\, d\mu_n \quad = \quad \mu_n[\mathbb{R}] \quad = \quad 1 \quad \not\to \quad 0.$$

Die Masse wandert in diesem Fall ins Unendliche ab.

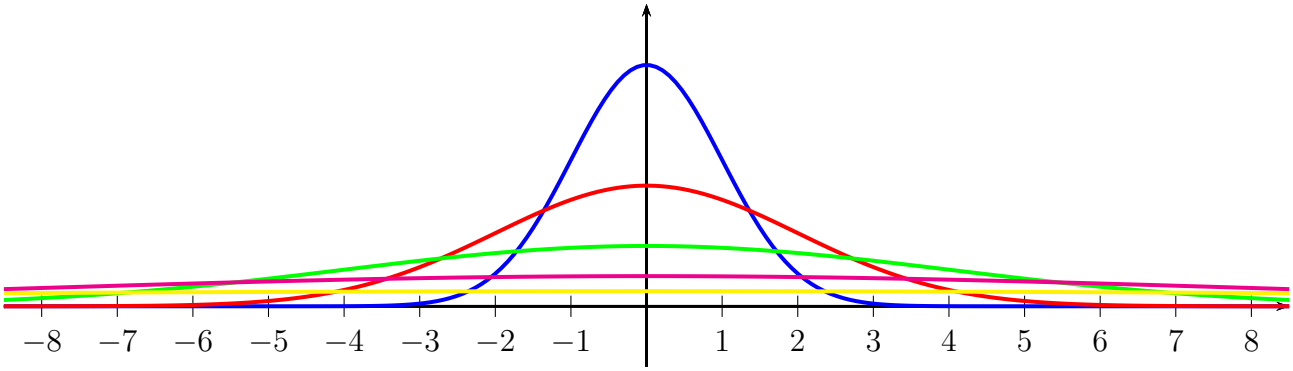


Abbildung 5.2: Vage Konvergenz der Normalverteilungen $N(0,n)$ gegen das Nullmaß.

(4). ***Diskrete Approximation von Wahrscheinlichkeitsverteilungen:*** Eine gegebene Wahrscheinlichkeitsverteilung können wir auf verschiedene Arten durch diskrete Wahrscheinlichkeitsverteilungen, also Konvexkombinationen von Diracmaßen approximieren:

(a) **Klassische numerische Approximation:** Sei μ ein absolutstetiges Wahrscheinlichkeitsmaß auf $[0,1]$ mit positiver stetiger Dichtefunktion proportional zu $g(x)$, und sei

$$\mu_n \quad := \quad \sum_{i=1}^{n} w_n^{(i)} \delta_{\frac{i}{n}} \qquad \text{mit} \qquad w_n^{(i)} \quad = \quad \frac{g(\frac{i}{n})}{\sum_{j=1}^{n} g(\frac{j}{n})}.$$

Dann konvergiert μ_n schwach gegen μ, denn

$$\int f\, d\mu_n \quad = \quad \sum_{i=1}^{n} w_n^{(i)}\, f\left(\frac{i}{n}\right) \quad = \quad \frac{\frac{1}{n}\sum_{i=1}^{n} f\left(\frac{i}{n}\right)\, g\left(\frac{i}{n}\right)}{\frac{1}{n}\sum_{i=1}^{n} g\left(\frac{i}{n}\right)}$$

$$\stackrel{n\nearrow\infty}{\longrightarrow} \quad \frac{\int_0^1 fg\, dx}{\int_0^1 g\, dx} \quad = \quad \int f\, d\mu \qquad \text{für alle } f \in C([0,1]).$$

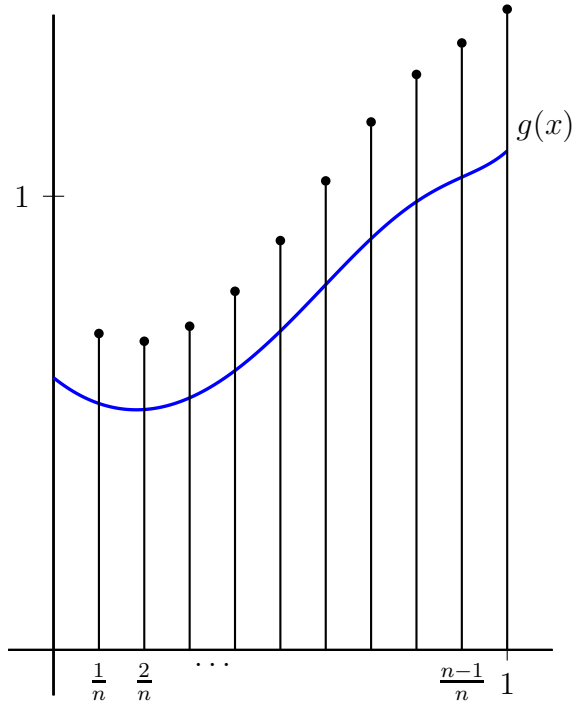


Abbildung 5.3: Stützstellen und Gewichte einer deterministischen Approximation von μ.

Die Stützstellen i/n und die Gewichte $w_n^{(i)}$ können natürlich auch auf andere Art gewählt werden, z.B. kann die hier verwendete naive Approximation des Integrals durch eine andere deterministische Quadraturformel ersetzt werden.

(b) **Monte-Carlo-Approximation**: Sei μ nun ein *beliebiges* Wahrscheinlichkeitsmaß auf $\mathcal{B}(\mathbb{R})$. Sind $X_1, X_2, ... : \Omega \to S$ unabhängige Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ mit Verteilung μ, dann konvergieren die **empirischen Verteilungen**

$$\widehat{\mu}_n(\omega) \quad := \quad \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}$$

nach dem Satz von Glivenko-Cantelli (Satz 4.15) für P-fast alle ω schwach gegen μ. Allgemeiner gilt die fast sichere schwache Konvergenz der empirischen Verteilungen sogar für unabhängige, identisch verteilte Zufallsvariablen X_i mit Werten in einem beliebigen vollständigen metrischen Raum S, siehe Satz 6.10.

5.1.2 Konvergenz der Verteilungen von Zufallsvariablen

Im Gegensatz zu anderen Konvergenzbegriffen für eine Folge $(Y_n)_{n\in\mathbb{N}}$ von Zufallsvariablen bezieht sich die Verteilungskonvergenz nur auf die Verteilungen der Y_n. Insbesondere können die

Zufallsvariablen Y_n und der Grenzwert Y alle auf unterschiedlichen Wahrscheinlichkeitsräumen definiert sein.

Beispiel (Wartezeiten). Die Wartezeit T_p auf den ersten Erfolg bei unabhängigen Ereignissen mit Erfolgswahrscheinlichkeit $p \in (0,1)$ ist geometrisch verteilt:

$$P[T_p > k] \quad = \quad (1-p)^k \qquad \text{für alle } k \in \mathbb{N}.$$

Sei nun eine Intensität $\lambda > 0$ gegeben. Um kontinuierliche Wartezeiten zu approximieren, betrachten wir unabhängige Ereignisse, die zu den Zeitpunkten i/n, $n \in \mathbb{N}$, mit Wahrscheinlichkeit λ/n stattfinden. Dann ist $\widetilde{T}_n := \frac{1}{n} T_{\lambda/n}$ die Wartezeit bis zum ersten Eintreten eines Ereignisses. Für $n \to \infty$ gilt

$$P\left[\widetilde{T}_n > c\right] \;=\; P\left[T_{\lambda/n} > nc\right] \;=\; (1-\lambda/n)^{\lfloor nc \rfloor} \quad \overset{n \nearrow \infty}{\longrightarrow} \quad e^{-\lambda c} \qquad \text{für alle } c \geq 0.$$

Also konvergiert $\widetilde{T}_n$ in Verteilung gegen eine $\mathrm{Exp}(\lambda)$-verteilte Zufallsvariable.

Wir untersuchen nun den Zusammenhang der schwachen Konvergenz der Verteilungen mit anderen Konvergenzarten in dem Fall, dass die Folge Y_n und Y auf einem *gemeinsamen Wahrscheinlichkeitsraum* $(\Omega, \mathcal{A}, P)$ definiert sind:

Satz 5.2 (Stochastische Konvergenz impliziert Konvergenz in Verteilung). *Seien Y_n $(n \in \mathbb{N})$ und Y Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit Werten in einem metrischen Raum S. Konvergiert die Folge Y_n P-fast sicher oder P-stochastisch gegen Y, dann konvergiert Y_n auch in Verteilung gegen Y.*

Beweis. Sei $f : S \to \mathbb{R}$ Lipschitz-stetig und beschränkt. Konvergiert Y_n fast sicher gegen Y, dann konvergiert auch $f(Y_n)$ fast sicher gegen $f(Y)$. Nach dem Satz von Lebesgue folgt

$$E[f(Y_n)] \quad \longrightarrow \quad E[f(Y)].$$

Konvergiert Y_n stochastisch gegen Y, dann konvergiert auch $f(Y_n)$ stochastisch gegen $f(Y)$, denn für $\varepsilon > 0$ gilt

$$P[|f(Y_n) - f(Y)| \geq \varepsilon] \;\leq\; P[d(Y_n, Y) \geq \varepsilon / L],$$

wobei L eine Lipschitz-Konstante für f ist. Also hat jede Teilfolge $f(Y_{n_k})$ von $f(Y_n)$ eine fast sicher gegen $f(Y)$ konvergente Teilfolge $f(Y_{n_{k_l}})$, und wie zuvor folgt

$$E[f(Y_{n_{k_l}})] \quad \longrightarrow \quad E[f(Y)].$$

Damit haben wir gezeigt, dass jede Teilfolge der Folge $E[f(Y_n)]$ der Erwartungswerte eine gegen $E[f(Y)]$ konvergente Teilfolge, und somit es gilt erneut

$$E[f(Y_n)] \quad \longrightarrow \quad E[f(Y)].$$

□

Wir beweisen nun eine partielle Umkehrung der Aussage aus Satz 5.2 im Fall $S = \mathbb{R}$:

Satz 5.3 (Skorokhod - Darstellung und Charakterisierung der schwachen Konvergenz). *Seien μ_n, μ Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Verteilungsfunktionen F_n bzw. F. Dann sind äquivalent:*

(1). Die Folge $(\mu_n)_{n\in\mathbb{N}}$ konvergiert schwach gegen μ.

(2). $F_n(c) \to F(c)$ für alle Stetigkeitsstellen c von F.

(3). Es existieren Zufallsvariablen G_n, G auf $(\Omega, \mathcal{A}, P) = ((0,1), \mathcal{B}((0,1)), \mathcal{U}_{(0,1)})$ mit Verteilungen μ_n bzw. μ, sodass $G_n \to G$ P-fast sicher.

Beweis. Die Implikation „(3) $\Rightarrow$ (1)“ folgt aus Satz 5.2.

„(1) $\Rightarrow$ (2)“ folgt aus dem Portemanteau-Theorem: Ist F stetig bei c, dann gilt $\mu[\{c\}] = 0$, und damit

$$F_n(c) \quad = \quad \int I_{(-\infty,c]}\, d\mu_n \quad \longrightarrow \quad = \quad \int I_{(-\infty,c]}\, d\mu \quad = \quad F(c). \tag{5.1.1}$$

„(2) $\Rightarrow$ (3)“: Für $u \in (0,1)$ betrachten wir die minimalen und maximalen u-Quantile

$$\underline{G}(u) := \inf\{x \in \mathbb{R} : F(x) \geq u\}, \qquad \text{und} \qquad \overline{G}(u) := \inf\{x \in \mathbb{R} : F(x) > u\}$$

der Verteilung μ, siehe Abschnitt 1.4. Entsprechend seien $\underline{G}_n$ und $\overline{G}_n$ die minimalen und maximalen u-Quantile der Verteilung μ_n. Analog zum Beweis von Satz 1.22 zeigt man, dass $\underline{G}$ und $\overline{G}$ bzw. $\underline{G}_n$ und $\overline{G}_n$ unter der Gleichverteilung $P = \mathcal{U}_{(0,1)}$ Zufallsvariablen mit Verteilung μ bzw. μ_n sind. Wir zeigen, dass aus (2) folgt:

Behauptung: $\lim \underline{G}_n \;=\; \lim \overline{G}_n \;=\; \underline{G} \;=\; \overline{G}$ P-fast sicher.

Damit ist dann auch die Implikation „(2) $\Rightarrow$ (3)“ bewiesen. Den Beweis der Behauptung führen wir in mehreren Schritten durch:

(a) Offensichtlich gilt $\underline{G} \leq \overline{G}$ und $\underline{G}_n \leq \overline{G}_n$ für alle $n \in \mathbb{N}$.

(b) Weiterhin gilt $\underline{G} = \overline{G}$ und $\underline{G}_n = \overline{G}_n$ P-fast sicher, denn

$$P[\underline{G} \neq \overline{G}] \;=\; P\left[\bigcup_{c\in\mathbb{Q}} \{\underline{G} \leq c < \overline{G}\}\right] \;\leq\; \sum_{c\in\mathbb{Q}} (\underbrace{P[\{\underline{G} \leq c\}]}_{=F(c)} - \underbrace{P[\{\overline{G} \leq c\}]}_{=F(c)}) \;=\; 0\,,$$

und entsprechend folgt $P[\underline{G}_n \neq \overline{G}_n]$ für jedes $n \in \mathbb{N}$.

(c) Wir zeigen nun

$$\limsup \overline{G}_n(u) \quad\leq\quad \overline{G}(u) \qquad \text{und} \qquad \liminf\, \underline{G}_n(u) \quad\geq\quad \underline{G}(u) \tag{5.1.2}$$

für $u \in (0,1)$. Zum Beweis der ersten Aussage genügt es zu zeigen, dass

$$\limsup \overline{G}_n(u) \quad\leq\quad c \qquad \text{für alle } c > \overline{G}(u) \quad \text{mit } \mu[\{c\}] = 0 \tag{5.1.3}$$

gilt, denn es existieren höchstens abzählbar viele c mit $\mu[\{c\}] \neq 0$. Für $c > \overline{G}(u)$ mit $\mu[\{c\}] = 0$ gilt aber nach Definition von $\overline{G}$ und nach (2):

$$u \quad<\quad F(c) \quad=\quad \lim_{n\to\infty} F_n(c).$$

Also existiert ein $n_0 \in \mathbb{N}$, so dass

$$F_n(c) \;>\; u, \qquad \text{und damit} \qquad \overline{G}_n(u) \;\leq\; c \tag{5.1.4}$$

für alle $n \geq n_0$ gilt. Es folgt $\limsup \overline{G}_n(u) \leq c$. Damit haben wir die erste Aussage in (5.1.2) bewiesen. Die zweite Aussage zeigt man auf ähnliche Weise.

(d) Aus (a), (b) und (c) folgt nun P-fast sicher:

$$\limsup \underline{G}_n \overset{(a)}{\leq} \limsup \overline{G}_n \overset{(c)}{\leq} \overline{G} \overset{(b)}{=} \underline{G} \overset{(3)}{\leq} \liminf \overline{G}_n \overset{(a)}{\leq} \liminf \overline{G}_n,$$

also $\lim \underline{G}_n \;=\; \lim \overline{G}_n \;=\; \underline{G} \;=\; \overline{G}$.

□

5.1.3 Existenz schwach konvergenter Teilfolgen

Ein wesentlicher Schritt, um den oben skizzierten Beweis des Zentralen Grenzwertsatzes zu vervollständigen, ist es, zu zeigen, dass die Verteilungen der standardisierten Summen von unabhängigen, identisch verteilten, quadratintegrierbaren Zufallsvariablen eine schwach konvergente Teilfolge haben.

Auf einer *endlichen* Menge S mit d Elementen können wir eine Folge von Wahrscheinlichkeitsverteilungen als beschränkte Folge in $\mathbb{R}^d$ auffassen. Daher existiert stets eine konvergente Teilfolge – der Grenzwert ist wieder eine Wahrscheinlichkeitsverteilung auf S. Für unendliche Mengen S gilt eine entsprechende Aussage im Allgemeinen nicht. Wir formulieren nun ein hinreichendes (und unter schwachen zusätzlichen Voraussetzungen auch notwendiges) Kriterium für die Existenz schwach konvergenter Teilfolgen für Folgen von Wahrscheinlichkeitsmaßen auf einem allgemeinen metrischen Raum S.

Definition (Straffheit von Folgen von Wahrscheinlichkeitsmaßen). *Eine Folge $\mu_n \in WV(S)$ heißt **straff**, falls zu jedem $\varepsilon > 0$ eine kompakte Menge $K \subseteq S$ existiert mit*

$$\mu_n[K] \;\geq\; 1 - \varepsilon \qquad \textit{für alle } \; n \in \mathbb{N}.$$

Eine straffe Folge von Wahrscheinlichkeitsmaßen ist also gleichmäßig auf Kompakta konzentriert. Die Masse kann daher für $n \to \infty$ nicht ins Unendliche abwandern.

Beispiel (Normalverteilungen). Die Folge $\mu_n = N(m_n, \sigma_n^2), m_n \in \mathbb{R},\ \sigma_n > 0$, ist genau dann straff, wenn die Folgen m_n und σ_n der Mittelwerte und Standardabweichungen beschränkt sind.

Satz 5.4 (Prokhorov). *Jede straffe Folge $\mu_n \in WV(S)$ hat eine schwach konvergente Teilfolge.*

Bemerkung. (1). Ist S selbst kompakt, dann ist der Raum $WV(S)$ aller Wahrscheinlichkeitsverteilungen auf S nach dem Satz von Prokhorov sogar **kompakt** bezüglich der schwachen Topologie, d.h. **jede** Folge $\mu_n \in WV(S)$ hat eine schwach konvergente Teilfolge.

(2). Insbesondere ist der Raum $WV(\overline{\mathbb{R}})$ aller Wahrscheinlichkeitsverteilungen auf $[-\infty, \infty]$ **kompakt**. Hieraus folgt, dass jede Folge $\mu_n \in WV(\mathbb{R})$ eine vag konvergente Teilfolge hat. Der Limes ist jedoch i.A. kein Wahrscheinlichkeitsmaß auf $\mathbb{R}$, da die Masse ins unendliche abwandern kann.

(3). UMKEHRUNG EINFÜGEN

Wir beweisen den Satz von Prokhorov hier nur für $S = \mathbb{R}$ (Satz von Helly-Bray). In diesem Fall können wir kompakte Mengen ersetzen durch kompakte Intervalle von der Form $[-c, c]$, da jede kompakte Menge in $\mathbb{R}$ in einem solchen Intervall enthalten ist. Einen Beweis im allgemeinen Fall findet man zum Beispiel in *Billingsley: Convergence of probability measures*.

Beweis im Fall $S = \mathbb{R}$. Sei μ_n $(n \in \mathbb{N})$ eine straffe Folge von Wahrscheinlichkeitsmaßen auf $\mathbb{R}$. Um die Existenz einer schwach konvergenten Teilfolge zu zeigen, betrachten wir die Folge der Verteilungsfunktionen F_n. Wir zeigen die Aussage in mehreren Schritten:

(1). *Es existiert eine Teilfolge $(F_{n_k})_{k\in\mathbb{N}}$, sodass $F_{n_k}(x)$ für alle $x \in \mathbb{Q}$ konvergiert:*

Zum Beweis verwenden wir ein Diagonalverfahren: Sei $x_1, x_2, ...$ eine Abzählung von $\mathbb{Q}$. Wegen $0 \le F_n \le 1$ existiert eine Teilfolge $(F_{n_k^{(1)}})_{k\in\mathbb{N}}$, für die $F_{n_k^{(1)}}(x_1)$ konvergiert. Ebenso existiert eine Teilfolge $(F_{n_k^{(2)}})_{k\in\mathbb{N}}$ von $(F_{n_k^{(1)}})_{k\in\mathbb{N}}$, für die $F_{n_k^{(2)}}(x_2)$ konvergiert, usw. Die Diagonalfolge $F_{n_k}(x) := F_{n_k^{(k)}}(x)$ konvergiert dann für alle $x \in \mathbb{Q}$.
Für $x \in \mathbb{Q}$ setzen wir $\overline{F}(x) := \lim_{k\to\infty} F_{n_k}(x)$. Nach (1) existiert der Grenzwert, außerdem ist die Funktion $\overline{F} : \mathbb{Q} \to [0,1]$Der Limes existiert nach 1. für $x \in \mathbb{Q}$ und die Funktion $\overline{F} : \mathbb{Q} \to [0,1]$ monoton wachsend, da die Funktionen F_{n_k} monoton wachsend sind.

(2). *Stetige Fortsetzung von $\overline{F}$ auf $[0,1]$:* Für $x \in \mathbb{R}$ setzen wir

$$F(x) := \inf\{\overline{F}(y) \mid y \in \mathbb{Q}, y > x\}.$$

Die folgenden Eigenschaften der Funktion F prüft man leicht nach:

(a) Die Funktion F ist rechtsstetig, monoton wachsend, und es gilt $0 \le F \le 1$.

(b) $F_{n_k}(x) \to F(x)$ für alle $x \in \mathbb{R}$, an denen F stetig ist.

(3). Aus (a)folgt, dass durch

$$\mu[(a,b]] := F(b) - F(a), \qquad -\infty < a \le b < \infty,$$

ein positives Maß auf $\mathbb{R}$ definiert wird mit

$$\mu[\mathbb{R}] = \lim_{c\to\infty} \mu[(-c,c]] \ \in [0,1].$$

Wir zeigen nun, dass μ eine *Wahrscheinlichkeitsverteilung* auf $\mathbb{R}$ ist, falls die Folge $(\mu_n)_{n\in\mathbb{N}}$ *straff* ist. Es gilt nämlich:

$$\mu[(-c,c]] \ = \ F(c) - F(-c) \ = \ \lim_{k\to\infty} (F_{n_k}(c) - F_{n_k}(-c)) \ = \ \lim_{k\to\infty} \mu_{n_k}[(-c,c]] \quad (5.1.5)$$

für fast alle c. Aus der Straffheit von $(\mu_n)_{n\in\mathbb{N}}$ folgt, dass zu jedem $\varepsilon > 0$ ein $c(\varepsilon) \in \mathbb{R}$ existiert mit

$$\mu_{n_k}[(-c,c]] \ge 1 - \varepsilon \qquad \text{für alle } k.$$

Aus (5.1.5) folgt dann $\mu[(-c,c]] \ge 1 - \varepsilon$, falls c groß genug ist, und damit für $\varepsilon \searrow 0$:

$$\mu[\mathbb{R}] \ge 1, \qquad \text{also} \quad \mu(\mathbb{R}) = 1.$$

(4). Aus (b) folgt nun nach Satz 5.3, dass die Folge $(\mu_{n_k})_{k\in\mathbb{N}}$ schwach gegen μ konvergiert.

□

5.1.4 Schwache Konvergenz über charakteristische Funktionen

Unter Verwendung der Existenz schwach konvergenter Teilfolgen einer straffen Folge von Wahrscheinlichkeitsverteilungen zeigen wir nun, dass eine Folge von Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$ genau dann schwach konvergiert, wenn die charakteristischen Funktionen gegen eine Grenzfunktion konvergieren, die bei 0 stetig ist. Dazu bemerken wir zunächst, dass eine Wahrscheinlichkeitsverteilung nach Lévy's Inversionsformel (Satz 4.14) *eindeutig* durch ihre charakteristische Funktion ϕ festgelegt ist.

Satz 5.5 (Stetigkeitssatz, Konvergenzsatz von Lévy). *Seien $(\mu_n)_{n\in\mathbb{N}}$ Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit charakteristischen Funktionen*

$$\phi_n(t) \quad = \quad \int e^{itx}\, \mu_n(dx).$$

Dann gilt:

(1). Konvergiert μ_n schwach gegen eine Wahrscheinlichkeitsverteilung μ, dann konvergieren auch die charakteristischen Funktionen:

$$\phi_n(t) \quad \to \quad \phi(t) \quad := \quad \int e^{itx}\, \mu(dx) \qquad \textit{für alle } \ t \in \mathbb{R}.$$

(2). Konvergiert umgekehrt $\phi_n(t)$ für alle $t \in \mathbb{R}$ gegen einen Limes $\phi(t)$, und ist ϕ stetig bei $t = 0$, dann ist ϕ die charakteristische Funktion einer Wahrscheinlichkeitsverteilung μ, und μ_n konvergiert schwach gegen μ.

Bemerkung. (1). Die Stetigkeit von ϕ bei 0 ist wesentlich. Zum Beispiel ist die Folge $\mu_n = N(0, n)$ nicht schwach konvergent, aber die charakteristischen Funktionen konvergieren punktweise:

$$\phi_n(t) = e^{-\frac{t^2}{2n}} \overset{n\uparrow\infty}{\to} \begin{cases} 0 & \text{falls } t \neq 0 \\ 1 & \text{falls } t = 0 \end{cases}.$$

(2). Eine Aussage wie im Satz gilt auch für Wahrscheinlichkeitsverteilungen auf $\mathbb{R}^d$. Hier definiert man die charakteristische Funktion $\phi : \mathbb{R}^d \to \mathbb{C}$ durch

$$\phi(t) \quad = \quad \int_{\mathbb{R}^d} e^{it\cdot x}\, \mu(dx), \qquad t \in \mathbb{R}^d.$$

Beweis. Der erste Teil der Aussage folgt unmittelbar aus $e^{itx} \quad = \quad \cos(tx) + i\sin(tx)$, denn Kosinus und Sinus sind beschränkte stetige Funktionen.
Der Beweis des zweiten Teils der Aussage erfolgt nun in mehreren Schritten. Wir nehmen an, dass die charakteristischen Funktionen $\phi_n(t)$ punktweise gegen eine bei 0 stetige Grenzfunktion $\phi(t)$ konvergieren.

(1). *Relative Kompaktheit: Jede Teilfolge von $(\mu_n)_{n\in\mathbb{N}}$ hat eine schwach konvergente Teilfolge.*

Dies ist der zentrale Schritt im Beweis. Nach dem Satz von Helly-Bray genügt es zu zeigen, dass μ_n $(n \in \mathbb{N})$ unter den Voraussetzungen straff ist. Dazu schätzen wir die Wahrscheinlichkeiten $\mu_n[|x| \geq c]$ mithilfe der charakteristischen Funktionen ab. Da die Funktion $f(u) = 1 - \frac{\sin u}{u}$ für $u \neq 0$ strikt positiv ist mit $\lim_{|u|\to\infty} f(u) = 1$, existiert eine Konstante $a > 0$ mit $f(u) \geq a$ für alle $|u| \geq 1$. Damit erhalten wir für $\varepsilon > 0$:

$$\begin{aligned} & \mu_n\left[|x| \geq \frac{1}{\varepsilon}\right] \\ = \quad & \mu_n\left[\{x \in \mathbb{R} \mid |\varepsilon x| \geq 1\}\right] \quad \leq \quad \frac{1}{a}\int \underbrace{\left(1 - \frac{\sin \varepsilon x}{\varepsilon x}\right)}_{=\frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon}(1-\cos(xt))dt} \mu_n(dx) \\ \overset{Fubini}{=} \quad & \frac{1}{2a\varepsilon}\int_{-\varepsilon}^{\varepsilon}(1 - \mathrm{Re}(\phi_n(t)))dt \underset{Lebesgue}{\overset{n\nearrow\infty}{\longrightarrow}} \frac{1}{2a\varepsilon}\cdot\int_{-\varepsilon}^{\varepsilon}(1 - \mathrm{Re}(\phi(t)))dt. \end{aligned} \tag{5.1.6}$$

Sei nun $\delta > 0$ vorgegeben. Ist ε hinreichend klein, dann gilt wegen der vorausgesetzten Stetigkeit von ϕ bei 0:

$$|1 - \mathrm{Re}(\phi(t))| \quad = \quad |\mathrm{Re}(\phi(0) - \phi(t))| \quad \leq \quad \frac{\delta a}{2} \qquad \text{für alle } t \in [-\varepsilon, \varepsilon].$$

Also können wir die rechte Seite von (5.1.6) durch $\delta/2$ abschätzen, und somit existiert ein $n_0 \in \mathbb{N}$ mit

$$\mu_n\left[|x| \geq \frac{1}{\varepsilon}\right] \quad \leq \quad \delta \qquad \text{für alle } n \geq n_0. \tag{5.1.7}$$

Diese Aussage gilt natürlich auch, falls wir ε noch kleiner wählen. Zudem gilt (5.1.7) auch für alle $n < n_0$, falls ε klein genug ist. Also ist μ_n $(n \in \mathbb{N})$ straff.

(2). *Der Grenzwert **jeder** schwach konvergenten Teilfolge von $(\mu_n)_{n\in\mathbb{N}}$ hat die charakteristische Funktion ϕ.*

Zum Beweis sei $(\mu_{n_k})_{k\in\mathbb{N}}$ eine Teilfolge von $(\mu_n)_{n\in\mathbb{N}}$ und μ eine Wahrscheinlichkeitsverteilung mit $\mu_{n_k} \xrightarrow{w} \mu$. Dann gilt nach dem ersten Teil der Aussage des Satzes:

$$\phi_\mu(t) \quad = \quad \lim_{k\to\infty} \phi_{n_k}(t) \quad = \quad \phi(t) \qquad \text{für alle } \ t \in \mathbb{R}.$$

(3). *Schwache Konvergenz von $(\phi_n)_{n\in\mathbb{N}}$.*

Nach dem Inversionssatz existiert höchstens eine Wahrscheinlichkeitsverteilung μ mit charakteristischer Funktion ϕ. Also konvergieren nach (2) alle schwach konvergenten Teilfolgen von $(\mu_n)_{n\in\mathbb{N}}$ gegen denselben Limes μ. Hieraus folgt aber, zusammen mit (1), dass

$(\mu_n)_{n\in\mathbb{N}}$ schwach gegen μ konvergiert, denn für $f \in \mathcal{C}_b(S)$ hat jede Teilfolge von $\int f\, d\mu_n$ eine gegen $\int f\, d\mu$ konvergente Teilfolge, und somit gilt $\int f\, d\mu_n \to \int f\, d\mu$.

□

5.2 Der Zentrale Grenzwertsatz

Wir können nun den zu Beginn dieses Kapitels skizzierten Beweis des Zentralen Grenzwertsatzes (engl. *Central Limit Theorem*) vervollständigen. Wir zeigen zunächst, dass ein zentraler Grenzwertsatz für Summen beliebiger unabhängiger, identisch verteilter Zufallsvariablen mit endlicher Varianz gilt. Diese Aussage wurde zuerst 1900 von Lyapunov bewiesen, der damit den Satz von de Moivre/Laplace (1733) deutlich verallgemeinern konnte. Am Ende dieses Abschnitts beweisen wir eine noch allgemeinere Version des Zentralen Grenzwertsatzes, die auf Lindeberg und Feller zurückgeht.

5.2.1 Zentraler Grenzwertsatz für Summen von i.i.d. Zufallsvariablen

Satz 5.6 (Zentraler Grenzwertsatz – 1. Version). *Seien $X_1, X_2, ... \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ unabhängige, identisch verteilte Zufallsvariablen mit Varianz σ^2 und sei*

$$S_n \;=\; X_1 + ... + X_n \;.$$

Dann konvergieren die Verteilungen der standardisierten Summen

$$\hat{S}_n \;=\; \frac{S_n - E[S_n]}{\sqrt{n}} \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - E[X_i])$$

schwach gegen $N(0, \sigma^2)$.

Bemerkung. (1). Alternativ kann man die standardisierten Summen auf Varianz 1 normieren, und erhält

$$\frac{S_n - E[S_n]}{\sigma \cdot \sqrt{n}} \quad \xrightarrow{\mathcal{D}} \quad Z,$$

wobei Z eine standardnormalverteilte Zufallsvariable ist.

(2). Die Voraussetzung $X_i \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ ist wesentlich. Bei unendlicher Varianz der X_i können sich andere Grenzverteilungen für die geeignet renormierten Summen $\frac{S_n - a_n}{b_n}$ $(a_n \in \mathbb{R}, b_n > 0)$ ergeben. Als Grenzverteilungen können i.A. die sogenannten stabilen Verteilungen auftreten, siehe dazu z.B. Satz 5.11 unten.

(3). Im Fall $\sigma^2 = 0$ gilt die Aussage auch. Hierbei interpretieren wir das Diracmaß δ_m als degenerierte Normalverteilung $N(m, 0)$.

Wir beweisen nun den Zentralen Grenzwertsatz in der oben stehenden Form:

Beweis. O.B.d.A. sei $E[X_i] = 0$, ansonsten betrachten wir die zentrierten Zufallsvariablen $\widetilde{X}_i := X_i - E[X_i]$. Nach dem Konvergenzsatz von Lévy genügt es zu zeigen, dass die charakteristischen Funktionen der standardisierten Summen $\widehat{S}_n$ punktweise gegen die charakteristische Funktion der Normalverteilung $N(0, \sigma^2)$ konvergieren, d.h.

$$\phi_{\hat{S}_n}(t) \;\to\; \phi_{N(0,\sigma^2)}(t) \;=\; e^{-\frac{\sigma^2 t^2}{2}} \qquad \forall\, t \in \mathbb{R}. \tag{5.2.1}$$

Da die Zufallsvariablen X_i unabhängig, identisch verteilt und zentriert sind, gilt für $t \in \mathbb{R}$:

$$\phi_{\hat{S}_n}(t) \quad \overset{E[S_n]=0}{=} \quad \phi_{S_n}\left(\frac{t}{\sqrt{n}}\right) \quad \overset{X_i \text{ iid}}{=} \quad \left(\phi_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

Aus $X_1 \in \mathcal{L}^2$ folgt $\phi_{X_1} \in C^2(\mathbb{R})$, und

$$\phi_{X_1}(t) \quad = \quad E[e^{itX_1}] \quad = \quad 1 + itE[X_1] + \frac{(it)^2}{2}E[X_1^2] + o(t^2) \quad = \quad 1 - \frac{t^2\sigma^2}{2} + o(t^2),$$

wobei o für eine Funktion $o : \mathbb{R}^+ \to \mathbb{C}$ mit $\lim_{\varepsilon \downarrow 0} \frac{|o(\varepsilon)|}{\varepsilon} = 0$ steht. Damit erhalten wir:

$$\phi_{\hat{S}_n}(t) \quad = \quad \left(1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n.$$

Wir vermuten, dass dieser Ausdruck für $n \to \infty$ gegen $e^{-\frac{t^2\sigma^2}{2}}$ strebt. Dies kann man beweisen, indem man den Logarithmus nimmt, und die Taylorapproximation $\log(1+w) = w + o(|w|)$ verwendet. Da die charakteristische Funktion komplexwertig ist, muss dazu allerdings der Hauptzweig der komplexen Logarithmusfunktion verwendet werden.
Wir zeigen stattdessen die Konvergenz ohne Verwendung von Aussagen aus der Funktionentheorie: Für komplexe Zahlen $z_i, w_i \in \mathbb{C}$ mit $|z_i|, |w_i| \leq 1$ gilt nach der Dreiecksungleichung

$$\begin{aligned}\left|\prod_{i=1}^n z_i - \prod_{i=1}^n w_i\right| &= |(z_1 - w_1)z_2 z_3 \cdots z_n + w_1(z_2 - w_2)z_3 z_4 \cdots z_n + \ldots + w_1 \cdots w_{n-1}(z_n - w_n)| \\ &\leq \sum_{i=1}^n |z_i - w_i|.\end{aligned}$$

Damit erhalten wir:

$$\begin{aligned}\left|\phi_{\widehat{S}_n}(t) - \exp\left(-\frac{t^2\sigma^2}{2}\right)\right| &= \left|\left(1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n - \exp\left(-\frac{t^2\sigma^2}{2n}\right)^n\right| \\ &\leq n \cdot \left|1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{t^2}{n}\right) - \exp\left(-\frac{t^2\sigma^2}{2n}\right)\right|.\end{aligned}$$

Da die rechte Seite für $n \to \infty$ gegen 0 konvergiert, folgt (5.2.1) und damit die Behauptung. □

Beispiel. (1). Sind $X_1, X_2, \ldots$ unabhängig mit $P[X_i = 1] = p$ und $P[X_i = 0] = 1 - p$, dann ist $S_n = \sum_{i=1}^{n} X_i$ binomialverteilt mit Parametern n und p. Die Aussage des Zentralen Grenzwertsatzes folgt in diesem Fall aus dem Satz von de Moivre/Laplace.

(2). Sind die Zufallsvariablen X_i unabhängig und Poissonverteilt mit Parameter $\lambda > 0$, dann ist $S_n = \sum_{i=1}^{n} X_i$ Poissonverteilt mit Parameter $n\lambda$. Der Zentrale Grenzwertsatz liefert in diesem Fall eine Normalapproximation für Poissonverteilungen mit großer Intensität (Übung).

(3). Sind $X_1, X_2, \ldots$ unabhängige, $N(m, \sigma^2)$-verteilte Zufallsvariablen, dann gilt

$$\widehat{S}_n \quad = \quad \frac{X_1 + X_2 + \ldots + X_n - nm}{\sqrt{n}} \quad \sim \quad N(0, \sigma^2)$$

für alle $n \in \mathbb{N}$ (und nicht nur asymptotisch!).

Warum tritt die Normalverteilung im Limes auf? Wie schon im letzten Beispiel bemerkt, gilt

$$X_i \sim N(0, \sigma^2) \text{ unabhängig} \quad \Rightarrow \quad \frac{X_1 + \ldots + X_n}{\sqrt{n}} \sim N(0, \sigma^2).$$

Die zentrierten Normalverteilungen sind also „invariant“ unter der *Reskalierungstransformation* aus dem zentralen Grenzwertsatz. Man kann sich leicht plausibel machen, dass eine Grenzverteilung der standardisierten Summen unabhängiger quadratintegrierbarer Zufallsvariablen eine entsprechende Invarianzeigenschaft haben muss. Tatsächlich sind die zentrierten Normalverteilungen die einzigen nichtdegenerierten Wahrscheinlichkeitsverteilungen mit dieser Invarianz. Aus dem Zentralen Grenzwertsatz folgt sogar:

Korollar 5.7. *Sei μ eine Wahrscheinlichkeitsverteilung auf $\mathbb{R}$ mit $\int x^2 \mu(dx) < \infty$. Gilt*

$$X, Y \sim \mu \text{ unabhängig} \quad \Rightarrow \quad \frac{X + Y}{\sqrt{2}} \sim \mu, \tag{5.2.2}$$

dann ist μ eine zentrierte Normalverteilung.

Bemerkung. Die Aussage gilt auch ohne die Voraussetzung $\int x^2 \mu(dx) < \infty$; der Beweis ist aber aufwändiger, siehe z.B. BREIMAN: PROBABILITY.

Beweis. Seien $X_1, X_2, \ldots$ unabhängige Zufallsvariablen mit Verteilung μ. Aus der Voraussetzung (5.2.2) folgt $E[X_i] = \int x \, \mu(dx) = 0$ für alle $i \in \mathbb{N}$, und durch Induktion:

$$\frac{X_1 + \ldots + X_n}{\sqrt{n}} \quad \sim \quad \mu \qquad \text{für } n = 2^k, k \in \mathbb{N}.$$

Wegen $\int x^2 \mu(dx) < \infty$ sind die X_i quadratintegrierbar. Durch Anwenden des zentralen Grenzwertsatzes auf die standardisierten Summen folgt, dass μ eine zentrierte Normalverteilung ist. □

5.2.2 Normal- und Poisson-Approximationen

Die Normalverteilungsasymptotik der standardisierten Summen wird häufig verwendet, um Wahrscheinlichkeiten näherungsweise zu berechnen. Wir betrachten zunächst ein typisches Beispiel:

Beispiel (Versicherungsgesellschaft mit n Verträgen). Eine Versicherungsgesellschaft habe mit n Kunden Verträge abgeschlossen. Beim Eintreten des Schadenfalls für Vertrag i muss die Leistung $X_i \geq 0$ gezahlt werden. Wir nehmen an, dass gilt:

$$X_i \in \mathcal{L}^2 \text{ i.i.d.} \quad \text{mit} \quad E[X_i] = m,\ \mathrm{Var}[X_i] = \sigma^2.$$

Die Prämie pro Vertrag betrage $\Pi = m + \lambda\sigma^2$, wobei m die erwartete Leistung ist und $\lambda\sigma^2$ mit $\lambda > 0$ einem Risikozuschlag entspricht. Die Einnahmen nach einer Zeitperiode betragen dann $n \cdot \Pi$, die Ausgaben $S_n = X_1 + ... + X_n$. Wir wollen die Wahrscheinlichkeit des Ruinereignisses

$$S_n > k + n\Pi,$$

berechnen, wobei k das Anfangskapital bezeichnet. Hierbei nehmen wir implizit an, dass nicht verzinst wird, und die Abrechnung nur am Schluß einer Zeitperiode erfolgt. Wenn die standardisierten Schadenssummen mithilfe einer ZGS-Näherung approximiert werden, ergibt sich

$$\begin{aligned} P[\text{Ruin}] &= P[S_n > k + n\Pi] = P[S_n - E[S_n] > k + n\lambda\sigma^2] \\ &= P\left[\frac{S_n - E[S_n]}{\sigma\sqrt{n}} > \frac{k}{\sigma\sqrt{n}} + \lambda\sigma\sqrt{n}\right] \\ &\approx P\left[Z > \frac{k}{\sigma\sqrt{n}} + \lambda\sigma\sqrt{n}\right], \end{aligned}$$

wobei Z eine standardnormalverteilte Zufallsvariable ist. Der Ausdruck auf der rechten Seite geht für $n \to \infty$ gegen 0. Eine große Anzahl von Verträgen sollte also eine kleine Ruinwahrscheinlichkeit implizieren. Für $n = 2000$, $\sigma = 60$ und $\lambda = 0,05\%$ ergibt sich beispielsweise:

$$\begin{aligned} k = 0 &: P[\text{Ruin}] \approx 9\%, \\ k = 1500 &: P[\text{Ruin}] \approx 3\%. \end{aligned}$$

Nach einer solchen Überschlagsrechnung sollte man das verwendete Modell und die Approximationsschritte einer kritischen Analyse unterziehen. In unserem Fall stellen sich unmittelbar mehrere Fragen:

(1). Wir haben die ZGS-Näherung verwendet, obwohl die auftretenden Schranken für die standardisierten Summen von n abhängen. Ist das in diesem Fall zulässig?

(2). Ist die Quadratintegrierbarkeit der X_i eine sinnvolle Modellannahme, und was ergibt sich andernfalls?

(3). In einem realistischen Modell kann man nicht davon ausgehen, dass die X_i identisch verteilt sind. Gilt trotzdem ein Zentraler Grenzwertsatz?

(4). Ist die Unabhängigkeitsannahme gerechtfertigt?

Wir werden nun auf die ersten drei Fragen näher eingehen. Das folgende Beispiel zeigt, dass man in der Tat vorsichtig sein sollte, wenn man von n abhängige Quantile von standardisierten Summen durch entsprechende Quantile von Normalverteilungen ersetzt:

Beispiel (**Eine zu naive ZGS-Approximation).** Seien $X_i, i \in \mathbb{N}$, unabhängige, identisch verteilte Zufallsvariablen mit $E[X_i] = 0$ und $\mathrm{Var}[X_i] = 1$, und sei $a > 0$. Mit einer ZGS-Approximation erhalten wir für große n:

$$\begin{aligned} P\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq a\right] &= P\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \geq a\sqrt{n}\right] \\ &\approx \frac{1}{\sqrt{2\pi}}\int_{a\sqrt{n}}^{\infty} e^{-\frac{x^2}{2}}dx \\ &= e^{-\frac{na^2}{2}}\cdot\frac{1}{\sqrt{2\pi}}\int_0^{\infty} e^{-a\sqrt{n}y-\frac{y^2}{2}}dy \qquad \left(x = a\sqrt{n}+y\right) \\ &= e^{-\frac{na^2}{2}}\cdot\frac{1}{\sqrt{2\pi n}}\int_0^{\infty} e^{-az-\frac{z^2}{2n}}\,dz \qquad \left(z = \sqrt{n}y\right) \\ &\sim \frac{1}{\sqrt{2\pi a^2 n}}\cdot\exp\left(-\frac{na^2}{2}\right) \end{aligned}$$

Dies ist aber **nicht** die korrekte Asymptotik für $n \to \infty$. Auf der exponentiellen Skala gilt nämlich

$$P\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq a\right] \simeq \exp\left(-nI(a)\right),$$

wobei $I(a)$ die Ratenfunktion aus dem Satz von Chernoff ist. Diese ist im Allgemeinen von $na^2/2$ verschieden. Die ZGS-Approximation ist hier nicht anwendbar, da $a\sqrt{n}$ von n abhängt!

Dass die Näherung im Versicherungsbeispiel von oben trotzdem mit Einschränkungen anwendbar ist, wenn die Zufallsvariablen X_i dritte Momente haben, garantiert die folgende *Abschätzung der Konvergenzgeschwindigkeit im Zentralen Grenzwertsatz:*

Satz 5.8 (Berry-Esséen). *Seien $X_i \in \mathcal{L}^3$ i.i.d. Zufallsvariablen, $Z \sim N(0,1)$, und seien*

$$\begin{aligned} F_n(x) &:= P\left[\frac{S_n - E[S_n]}{\sigma\sqrt{n}} \leq x\right], \\ \Phi(x) &:= P[Z \leq x]. \end{aligned}$$

Dann gilt folgende Abschätzung:

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

Den Beweis dieser Aussage findet man etwa im Buch PROBABILITY THEORY von R. Durrett (4.10).
Für die Normalapproximation der Binomialverteilung $\mathrm{Bin}(n,p)$ ergibt sich beispielsweise

$$\frac{3 \cdot E[|X_1 - E[X_1]|^3]}{\sigma^3\sqrt{n}} = \frac{3 \cdot ((1-p)^2 + p^2)}{\sqrt{np(1-p)}}.$$

Für $p \to 0$ oder $p \to 1$ divergiert die rechte Seite. Wir erhalten also möglicherweise einen hohen Approximationsfehler für p nahe 0 oder 1. In diesen Fällen empfiehlt sich in der Tat die Verwendung der Poisson-Approximation anstelle des zentralen Grenzwertsatzes. Der folgende Satz quantifiziert den Fehler der Poisson-Approximation in der totalen Variationsdistanz:

Satz 5.9 (Poisson-Approximation der Binomialverteilung). *Für $p \in [0,1]$ und $n \in \mathbb{N}$ gilt*

$$\|\mathrm{Bin}(n,p) - \mathrm{Poisson}(np)\|_{TV} \leq np^2.$$

Der Beweis basiert auf einem Kopplungsargument:

Definition (Kopplung zweier Wahrscheinlichkeitsmaße). *Eine **Kopplung** zweier Wahrscheinlichkeitsmaße μ und ν auf meßbaren Räumen $(S,\mathcal{S})$ und $(T,\mathcal{T})$ ist gegeben durch ein Wahrscheinlichkeitsmaß η auf dem Produktraum $(S\times T, \mathcal{S}\otimes\mathcal{T})$ mit Randverteilungen μ und ν, bzw. durch Zufallsvariablen $X : \Omega \to S$ und $Y : \Omega \to T$ mit Verteilungen $X \sim \mu$ und $Y \sim \nu$, die auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definiert sind.*

Beweis. Um die Variationsdistanz abzuschätzen, konstruieren wir explizit eine Kopplung der Verteilungen $\mu := \mathrm{Bin}(n,p)$ und $\nu := \mathrm{Poisson}(np)$, die durch Zufallsvariablen S_n, T_n mit $S_n \sim \mu$, $T_n \sim \nu$, und $P[S_n \neq T_n] \leq np^2$ realisiert wird. Daraus folgt die Behauptung wegen

$$\begin{aligned} \|\mu - \nu\| &= \sup_{A\subseteq\mathbb{Z}_+} |\mu[A] - \nu[A]| = \sup_{A\subseteq\mathbb{Z}_+} |P[S_n \in A] - P[T_n \in A]| \\ &\leq P[S_n \neq T_n] \leq np^2. \end{aligned}$$

Da Bin(n,p) und Poisson(np) jeweils die Verteilung einer Summe von n unabhängigen Bernoulli- bzw. Poisson-verteilten Zufallsvariablen zum Parameter p sind, konstruieren wir zunächst eine Kopplung der beiden letzteren Wahrscheinlichkeitsmaße. Dazu verwenden wir eine Zufallsvariable Z mit Verteilung

$$Z \;=\; \begin{cases} 0 & \text{mit Wahrscheinlichkeit } 1-p, \\ k & \text{mit Wahrscheinlichkeit } p^k e^{-k}/k! \text{ für } k \in \mathbb{N}, \\ -1 & \text{mit Wahrscheinlichkeit } e^{-p} - (1-p). \end{cases}$$

Die Zufallsvariablen $X := I_{\{Z \neq 0\}}$ und $Y := \max(Z, 0)$ sind dann Bernoulli(p) bzw. Poisson(p)-verteilt, und es gilt

$$P[X \neq Y] \;=\; P[Z \notin \{0,1\}] \;=\; 1 - (1 - p + pe^{-p} \;=\; p\,(1 - e^{-p}) \;\leq\; p^2.$$

Um eine Kopplung von Bin(n,p) und Poisson(np) zu konstruieren, verwenden wir nun n unabhängige Kopien dieses Modells, d.h. wir setzen für $i = 1, \ldots, n$:

$$X_i := I_{\{Z_i \neq 0\}} \text{ und } Y_i := \max(Z_i, 0) \qquad \text{mit } Z_1, \ldots, Z_n \text{ unabhängig } \sim Z.$$

Dann sind $S_n := X_1 + \ldots + X_n$ und $T_n := Y_1 + \ldots + Y_n$ binomialverteilt mit Parametern (n,p) bzw. Poisson-verteilt mit Parameter np, und es gilt

$$P[S_n \neq T_n] \quad = \quad P[\exists\, i = 1, \ldots, n \,:\, X_i \neq Y_i] \quad \leq \quad \sum_{i=1}^{n} P[X_i \neq Y_i] \quad \leq \quad np^2.$$

□

Satz 5.9 zeigt, dass die Poisson-Approximation eine gute Näherung liefert, falls p deutlich kleiner als $\sqrt{n}$ ist. Ist p größer als $1 - \sqrt{n}$, dann kann man die Poisson-Approximation auf die mit Parametern $(n, 1-p)$ binomialverteilte Zufallsvariable $n - S_n$ mit $S_n \sim \mathrm{Bin}(n,p)$ anwenden. Geht p für $n \to \infty$ schneller als $1/\sqrt{n}$, aber langsamer als $1/n$ gegen Null, dann sind sowohl die Poisson- als auch die Normalapproximation der Binomialverteilung anwendbar.

5.2.3 Heavy Tails, Konvergenz gegen α-stabile Verteilungen

Als nächstes betrachten wir ein Beispiel, welches zeigt, dass die Voraussetzung der Quadratintegrierbarkeit der Zufallsvariablen essentiell für den zentralen Grenzwertsatz ist:

Seien $\alpha \in (1,2), r \in (0,\infty)$, und seien $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ unabhängige identisch verteilte absolutstetige Zufallsvariablen, deren Dichtefunktion

$$f_{X_i}(x) \quad = \quad |x|^{-\alpha-1} \qquad \text{für alle } |x| \geq r$$

erfüllt. Da die Dichte für $|x| \to \infty$ nur langsam abfällt, sind die Zufallsvariablen nicht quadratintegrierbar; sie sind aber integrierbar. Daher ergibt sich ein anderes asymptotisches Verhalten der charakteristischen Funktionen für $t \to 0$:

Lemma 5.10. *Für $t \to 0$ gilt*

$$\phi_{X_i}(t) \quad = \quad 1 + imt - c|t|^\alpha + O(t^2)$$

mit $m = E[X_i]$ und $c = \int_{\mathbb{R}} (1 - \cos u)|u|^{-\alpha-1}\, du \in (0, \infty)$.

Beweis. Sei $t \neq 0$. Wegen $e^{iu} - 1 - iu = O(u^2)$ und $\cos u - 1 = O(u^2)$ erhalten wir

$$\begin{aligned}
\phi_{X_i}(t) - 1 - imt &= \int_{-\infty}^{\infty} (e^{itx} - 1 - itx) f(x)\, dx \\
&= \int_{-\infty}^{\infty} (e^{iu} - 1 - iu) f\left(\frac{u}{t}\right) \frac{1}{|t|}\, du \\
&= \frac{1}{|t|} \int_{-tr}^{tr} (e^{iu} - 1 - iu) f\left(\frac{u}{t}\right)\, du + |t^\alpha| \int_{[-tr,tr]^C} (\cos u - 1)|u|^{-\alpha-1}\, du \\
&= -c|t|^\alpha + O(t^2).
\end{aligned}$$

□

Für die zentrierten Summen $S_n = \sum_{i=1}^n (X_i - m)$ folgt nach dem Lemma:

$$\phi_{S_n}(t) \quad = \quad (1 - c|t|^\alpha + O(t^2))^n.$$

Um Konvergenz der charakteristischen Funktionen zu erhalten, müssen wir X_n nun mit $n^{-1/\alpha}$ statt $n^{-1/2}$ reskalieren:

$$\begin{aligned}
\phi_{n^{-1/\alpha} S_n}(t) &= \phi_{S_n}(n^{-1/\alpha} t) \quad = \quad (1 - c|t|^\alpha n^{-1} + O(n^{-2/\alpha}))^n \\
&\to \exp(-c|t|^\alpha) \qquad \text{für } n \to \infty.
\end{aligned}$$

Nach dem Konvergenzsatz von Lévy folgt:

Satz 5.11. *Für $n \to \infty$ gilt*

$$n^{-1/\alpha} S_n \quad \xrightarrow{\mathcal{D}} \quad \mu_{c,\alpha},$$

wobei $\mu_{c,\alpha}$ die Wahrscheinlichkeitsverteilung mit charakteristischer Funktion

$$\phi_{c,\alpha}(t) \quad = \quad \exp(-c|t|^\alpha)$$

ist.

Definition. *Seien $\alpha \in (0,2]$ und $m \in \mathbb{R}$. Die Wahrscheinlichkeitsverteilungen mit charakteristischer Funktion*

$$\phi(t) \quad = \quad \exp(imt - c|t|^\alpha),$$

*$c \in (0,\infty)$, heißen **symmetrische α-stabile Verteilungen** mit Mittelwert m.*

Die Dichten der α-stabilen Verteilungen sind für $\alpha \neq 1,2$ nicht explizit berechenbar, fallen aber für $|x| \to \infty$ wie $|x|^{-\alpha-1}$ ab. Für $\alpha = 1$ erhält man die Cauchyverteilungen, für $\alpha = 2$ die Normalverteilungen. Satz 5.11 ist ein Spezialfall eines allgemeineren Grenzwertsatzes für Summen von Zufallsvariablen mit polynomiellen Tails, siehe z.B. BREIMAN, THEOREM 9.34.

5.2.4 Der Satz von Lindeberg-Feller

Wir wollen nun die Annahme fallen lassen, dass die Summanden X_i identisch verteilt sind, und zeigen, dass trotzdem ein zentraler Grenzwertsatz gilt. Sei

$$\widehat{S}_n \quad = \quad Y_{n,1} + Y_{n,2} + ... + Y_{n,n} \qquad \text{mit } Y_{n,i} \in \mathcal{L}^2(\Omega, \mathcal{A}, P).$$

Die Zufallsvariablen $Y_{n,i}$ können etwa kleine Störungen oder Messfehler beschreiben. Setzen wir

$$Y_{n,i} = \frac{X_i - E[X_i]}{\sqrt{n}} \quad \text{mit } X_i \in \mathcal{L}^2 \text{ unabhängig}, \tag{5.2.3}$$

so erhalten wir das Setup von oben.

Satz 5.12 (ZGS von Lindeberg-Feller). *Sei $\sigma \in (0,\infty)$. Es gelte:*

(i) $Y_{n,i}$ $(1 \leq i \leq n)$ *sind unabhängig für jedes $n \in \mathbb{N}$ mit* $E[Y_{n,i}] = 0$,

(ii) $\mathrm{Var}[\widehat{S}_n] = \sum_{i=1}^n \mathrm{Var}[Y_{n,i}] \overset{n\uparrow\infty}{\longrightarrow} \sigma^2$,

(iii) $\gamma_{n,\varepsilon} := \sum_{i=1}^n E[Y_{n,i}^2;\ |Y_{n,i}| > \varepsilon] \overset{n\uparrow\infty}{\longrightarrow} 0 \quad \forall\, \varepsilon > 0.$

Dann konvergiert die Verteilung von $\widehat{S}_n$ schwach gegen $N(0,\sigma^2)$.

Der Satz zeigt, dass die Summe vieler kleiner unabhängiger Störungen unter geeigneten Voraussetzungen ungefähr normalverteilt ist. Dies rechtfertigt bis zu einem gewissen Grad, dass Zufallsgrößen mit unbekannter Verteilung, die durch Überlagerung vieler kleiner Effekte entstehen, häufig durch normalverteilte Zufallsvariablen modelliert werden.

Bemerkung. (1). Der Zentrale Grenzwertsatz von oben ist ein Spezialfall des Satzes von Lindeberg-Feller: Sind $X_i \in \mathcal{L}^2$ i.i.d. Zufallsvariablen mit $E[X_i] = m$ und $\mathrm{Var}[X_i] = \sigma^2$, und definieren wir $Y_{n,i}$ wie in (5.2.3), dann gilt:

$$\mathrm{Var}[\widehat{S}_n] \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[X_i] \;=\; \mathrm{Var}[X_1] \;=\; \sigma^2, \qquad \text{für alle } n \in \mathbb{N},$$

und, für $\varepsilon > 0$

$$\begin{aligned}\gamma_{n,\varepsilon} &= \sum_{i=1}^{n} E\left[Y_{n,i}^2; |Y_{n,i}| > \varepsilon\right] \;=\; \frac{1}{n}\sum_{i=1}^{n} E\left[|X_i - m|^2; |X_i - m| > \varepsilon\sqrt{n}\right] \\ &= E\left[|X_1 - m|^2; |X_1 - m| > \varepsilon\sqrt{n}\right] \;\to\; 0 \qquad \text{für } n \to \infty,\end{aligned}$$

da X_1 quadratintegrierbar ist.

(2). Die Bedingung (iii) ist insbesondere erfüllt, wenn die *Lyapunovbedingung*

$$\sum_{i=1}^{n} E[|Y_{n,i}|^p] \overset{n\to\infty}{\longrightarrow} 0 \quad \text{für ein } p > 2 \text{ gilt,}$$

denn für $\varepsilon > 0$ ist $E[Y_{n,i}^2; |Y_{n,i}| \geq \varepsilon] \leq E[|Y_{n,i}|^p]/\varepsilon^{p-2}$.

Wir beweisen nun den Satz von Lindeberg-Feller: Der Beweis basiert wieder auf einer Analyse der Asymptotik der charakteristischen Funktionen. Dazu zeigen wir zunächst einige asymptotische Abschätzungen:

Beweis. (a) ***Vorüberlegungen:*** Sei $t \in \mathbb{R}$ fest.

(I) *Taylorapproximation für $\phi_{n,i}(t) := E[e^{itY_{n,i}}]$:*

Aus den verschiedenen Abschätzungen des Taylorrestglieds erhält man

$$e^{ix} = 1 + ix - \frac{x^2}{2} + R(x) \quad \text{mit} \quad |R(x)| \leq \min\left(\frac{|x|^3}{6}\,,\, x^2\right). \tag{5.2.4}$$

Damit ergibt sich

$$\phi_{n,i}(t) \;=\; 1 + itE[Y_{n,i}] - \frac{t^2}{2}E[Y_{n,i}^2] + E[R(tY_{n,i})] \;=\; 1 - \frac{t^2\sigma_{n,i}^2}{2} + R_{n,i},$$

wobei für $R_{n,i} := E[R(tY_{n,i})]$ die Abschätzung

$$|R_{n,i}| \;\leq\; E\left[\min\left(\frac{|tY_{n,i}|^3}{6}, t^2Y_{n,i}^2\right)\right] \tag{5.2.5}$$

gilt.

(II) Wir zeigen $\sum_{i=1}^n |R_{n,i}| \to 0$ *für* $n \to \infty$:

Für $\varepsilon > 0$ gilt nach (5.2.5):

$$|R_{n,i}| \leq \frac{1}{6} \cdot E\left[|tY_{n,i}|^3; |Y_{n,i}| \leq \varepsilon\right] + E[|tY_{n,i}|^2; |Y_{n,i}| > \varepsilon].$$

Mit $E\left[|tY_{n,i}|^3; |Y_{n,i}| \leq \varepsilon\right] \leq |t|^3 \varepsilon \cdot \sigma_{n,i}^2$ erhalten wir

$$\sum_{i=1}^n |R_{n,i}| \quad \leq \quad \frac{|t|^3 \varepsilon}{6} \sum_{i=1}^n \sigma_{n,i}^2 + t^2 \gamma_{n,\varepsilon},$$

und somit nach Voraussetzung (ii) und (iii)

$$\limsup_{n\to\infty} \sum_{i=1}^n |R_{n,i}| \quad \leq \quad \frac{\sigma^2 |t|^3}{6} \varepsilon \, .$$

Die Behauptung folgt für $\varepsilon \to 0$.

(III) Wir zeigen $\sup_{1\leq i\leq n} \sigma_{n,i}^2 \to 0$ *für* $n \to \infty$:

Für $\varepsilon > 0$ und $1 \leq i \leq n$ gilt

$$\sigma_{n,i}^2 \quad = \quad E[Y_{n,i}^2; |Y_{n,i}| \leq \varepsilon] + E[Y_{n,i}^2; |Y_{n,i}| > \varepsilon] \quad \leq \quad \varepsilon^2 + \gamma_{n,\varepsilon}.$$

Wegen $\gamma_{n,\varepsilon} \to 0$ für $n \to \infty$ ergibt sich

$$\limsup_{n\to\infty} \sup_{1\leq i\leq n} \sigma_{n,i}^2 \leq \varepsilon^2.$$

Die Behauptung folgt wieder für $\varepsilon \to 0$.

(b) ***Hauptteil des Beweises:*** Zu zeigen ist

$$\phi_{\hat{S}_n}(t) = \prod_{i=1}^n \phi_{n,i}(t) \overset{n\to\infty}{\longrightarrow} \exp\left(-\frac{t^2\sigma^2}{2}\right), \tag{5.2.6}$$

die Aussage folgt dann aus dem Konvergenzsatz von Lévy.

Wir zeigen:

$$\left|\prod_{i=1}^n \phi_{n,i}(t) - \prod_{i=1}^n \left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right)\right| \overset{n\to\infty}{\longrightarrow} 0, \qquad \text{und} \tag{5.2.7}$$

$$\prod_{i=1}^n \left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right) \overset{n\to\infty}{\longrightarrow} e^{-\frac{t^2\sigma^2}{2}}. \tag{5.2.8}$$

Daraus folgt (5.2.6), und damit die Behauptung.

Beweis von (5.2.7): Wie oben gezeigt, gilt für $z_i, w_i \in \mathbb{C}$ mit $|z_i|, |w_i| \leq 1$:

$$\left| \prod_{i=1}^{n} z_i - \prod_{i=1}^{n} w_i \right| \leq \sum_{i=1}^{n} |z_i - w_i|.$$

Zudem gilt $|\phi_{n,i}(t)| \leq 1$, und nach der 3. Vorüberlegung existiert ein $n_0 \in \mathbb{N}$ mit

$$1 - \frac{t^2 \sigma_{n,i}^2}{2} \in (0,1) \qquad \text{für alle } n \geq n_0 \text{ und } 1 \leq i \leq n. \tag{5.2.9}$$

Damit erhalten wir für $n \geq n_0$:

$$\left| \prod_{i=1}^{n} \phi_{n,i}(t) - \prod_{i=1}^{n} \left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right) \right| \;\leq\; \sum_{i=1}^{n} \left| \phi_{n,i}(t) - \left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right) \right| \;=\; \sum_{i=1}^{n} |R_{n,i}|$$

Die rechte Seite konvergiert nach der 2. Vorüberlegung gegen 0.

Beweis von (5.2.8): Wegen (5.2.9) erhalten wir

$$\begin{aligned} \log\left(\prod_{i=1}^{n} \left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right) \right) &= \sum_{i=1}^{n} \log\left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right) \\ &= -\sum_{i=1}^{n} \frac{t^2\sigma_{n,i}^2}{2} + \sum_{i=1}^{n} \widetilde{R}_{n,i}, \end{aligned}$$

wobei $|\widetilde{R}_{n,i}| \leq C \cdot \left(t^2\sigma_{n,i}^2\right)^2$ mit $\widetilde{C} \in (0,\infty)$. Die rechte Seite konvergiert nach Voraussetzung (ii) für $n \to \infty$ gegen $-\frac{t^2\sigma^2}{2}$, denn

$$\sum_{i=1}^{n} |\widetilde{R}_{n,i}| \;\leq\; Ct^4 \cdot \sum_{i=1}^{n} \sigma_{n,i}^4 \;\leq\; Ct^4 \cdot \sum_{i=1}^{n} \sigma_{n,i}^2 \cdot \sup_{1 \leq i \leq n} \sigma_{n,i}^2 \;\to\; 0$$

nach der 3. Vorüberlegung.

□

Bemerkung (Zentrale Grenzwertsätze für Summen abhängiger Zufallsvariablen). In allen Fällen haben wir bisher angenommen, dass die Zufallsvariablen X_i unabhängig sind. Tatsächlich hat man zentrale Grenzwertsätze auch für viele große Modellklassen mit Abhängigkeit gezeigt, beispielsweise für Martingale, additive Funktionale von Markovketten, Skalierungslimiten von Teilchensystemen, unterschiedliche Folgen von Parameterschätzern in der Statistik, usw. Wir werden darauf in weiterführenden Vorlesungen zurückkommen.

%sectionMultivariate Normalverteilungen

5.3 Multivariate Normalverteilungen und ZGS im $\mathbb{R}^d$

Multivariate Normalverteilungen haben wir bereits in Abschnitt 3.4.2 eingeführt. Mithilfe von charakteristischen Funktionen können wir eine etwas allgemeinere Definition geben, die auch degenerierte Normalverteilungen (zum Beispiel Dirac-Maße) einschließt:

Definition (Normalverteilung im $\mathbb{R}^d$). *Sei $m \in \mathbb{R}^d$, und sei $C \in \mathbb{R}^{d\times d}$ eine symmetrische, nicht-negativ definite Matrix. Die eindeutige Wahrscheinlichkeitsverteilung $N(m, C)$ im $\mathbb{R}^d$ mit charakteristischer Funktion $\phi(t) = \exp\left(-\frac{1}{2}t \cdot Ct + it \cdot m\right)$ heißt **Normalverteilung** mit Mittelwertvektor m und Kovarianzmatrix C.*

Die Existenz und Konstruktion einer Zufallsvariable mit Verteilung $N(m, C)$ ergibt sich aus der folgenden Bemerkung (3).

Bemerkung (Charakterisierungen und Transformationen von Normalverteilungen). Die folgenden Aussagen beweist man mithilfe von charakteristischen Funktionen:

(1). Ein Zufallsvektor $X : \Omega \to \mathbb{R}^d$ ist genau dann multivariat normalverteilt, wenn jede Linearkombination $\sum_{i=1}^d t_i X_i$ der Komponenten mit $t_1, \ldots, t_d \in \mathbb{R}$ normalverteilt ist. Genauer ist $X \sim N(m, C)$ äquivalent zu

$$t \cdot X \ \sim \ N(t \cdot m, t \cdot Ct) \qquad \text{für alle } t \in \mathbb{R}^d.$$

(2). Ist $X \sim N(m, C)$, dann gilt

$$AX + b \sim N(Am + b, ACA^T) \qquad \text{für alle } b \in \mathbb{R}^k \text{ und } A \in \mathbb{R}^{k\times d}, k \in \mathbb{N}.$$

(3). Sind $Z_1, \ldots, Z_d$ unabhängige, standardnormalverteilte Zufallsvariablen, und ist σ eine reelle $d \times d$-Matrix mit $C = \sigma\sigma^T$, dann hat der Zufallsvektor $\sigma Z + m$ mit $Z = (Z_1, \ldots, Z_d)^T$ die Verteilung $N(m, C)$.

(4). Im Fall $\det C \neq 0$ ist die Verteilung $N(m, C)$ absolutstetig bzgl. des d-dimensionalen Lebesgue-Maßes mit Dichte

$$f(y) \ = \ \frac{1}{\sqrt{(2\pi)^d |\det C|}} \exp\left(-\frac{1}{2}(y - m) \cdot C^{-1}(y - m)\right) .$$

Beispiel (χ^2-Verteilungen). Wir berechnen die Verteilung vom Quadrat des Abstandes vom Ursprung eines standardnormalverteilten Zufallsvektors im $\mathbb{R}^d$:

$$Z \ = \ (Z_1, ..., Z_d) \ \sim \ N(0, I_d), \quad \|Z\|^2 \ = \ \sum_{i=1}^d Z_i^2.$$

Wegen $f_{|Z_i|}(x) = \frac{2}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} \cdot I_{(0,\infty)}(x)$ folgt durch Anwenden des Dichtetransformationssatzes:

$$f_{Z_i^2}(y) = \sqrt{\frac{2}{\pi}}\, e^{-\frac{y}{2}} \cdot I_{(0,\infty)}(y) \cdot \frac{1}{2\sqrt{y}}\,,$$

d.h. Z_i^2 ist $\Gamma(\frac{1}{2}, \frac{1}{2})$-verteilt. Da die Zufallsvariablen $Z_i^2,\ 1 \le i \le d$, unabhängig sind, folgt:

$$\|Z\|^2 \;=\; \sum_{i=1}^{d} Z_i^2 \;\sim\; \Gamma\left(\frac{1}{2}, \frac{d}{2}\right)\,.$$

Definition (χ^2-Verteilung). *Die Gamma-Verteilung mit Parametern $1/2$ und $d/2$ heißt auch **Chiquadrat-Verteilung** $\chi^2(d)$ **mit** d **Freiheitsgraden**.*

5.3.1 Multivariater zentraler Grenzwertsatz

Auch im $\mathbb{R}^d$ gilt ein zentraler Grenzwertsatz:

Satz 5.13 (Multivariater zentraler Grenzwertsatz). *Seien $X_1, X_2, ... : \Omega \to \mathbb{R}^d$ unabhängige, identisch verteilte, quadratintegrierbare Zufallsvektoren auf $(\Omega, \mathcal{A}, P)$, und sei $S_n = X_1 + \ldots + X_n$. Dann gilt*

$$\frac{S_n - E[S_n]}{\sqrt{n}} \;\xrightarrow{\mathcal{D}}\; Z \;\sim\; N(0, C),$$

wobei $C_{jk} = \mathrm{Cov}[X_{1,j}, X_{1,k}]$ die Kovarianzmatrix der Zufallsvektoren X_i ist.

Der Beweis basiert auf folgender Charakterisierung der Verteilungskonvergenz von Zufallsvektoren:

Lemma 5.14 (Cramér-Wold Device). *Für Zufallsvariablen $Y, Y_1, Y_2, ... : \Omega \to \mathbb{R}^d$ gilt:*

$$Y_n \;\xrightarrow{\mathcal{D}}\; Y \quad\Leftrightarrow\quad p \cdot Y_n \;\xrightarrow{\mathcal{D}}\; p \cdot Y \qquad \forall\, p \in \mathbb{R}^d.$$

Beweisskizze. Die Richtung „⇒“ ist klar, da $Y \mapsto p \cdot Y$ stetig ist. Umgekehrt gilt:

$$p \cdot Y_n \xrightarrow{\mathcal{D}} p \cdot Y \;\Rightarrow\; E[\exp(ip \cdot Y_n)] \;\to\; E[\exp(ip \cdot Y)] \quad \forall\, p \in \mathbb{R}^d.$$

Mit einem ähnlichen Beweis wie im $\mathbb{R}^1$ folgt dann aus der Konvergenz der charakteristischen Funktionen die schwache Konvergenz $Y_n \xrightarrow{\mathcal{D}} Y$. Um die relative Kompaktheit zu zeigen (Satz von Helly-Bray), verwendet man dabei im $\mathbb{R}^d$ die multivariaten Verteilungsfunktionen

$$F_n(x_1, ..., x_d) := P[Y_{n,1} \le x_1, ..., Y_{n,d} \le x_d], \qquad (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

□

Wir beweisen nun den zentralen Grenzwertsatz im $\mathbb{R}^d$:

Beweis. Für $p \in \mathbb{R}^d$ gilt nach dem eindimensionalen zentralen Grenzwertsatz:

$$\begin{aligned} p \cdot \left(\frac{S_n - E[S_n]}{\sqrt{n}} \right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (p \cdot X_i - E[p \cdot X_i]) \\ &\xrightarrow{\mathcal{D}} N\left(0, \mathrm{Var}[p \cdot X_1]\right) = N(0, p \cdot Cp), \end{aligned}$$

da

$$\mathrm{Var}[p \cdot X_1] = \mathrm{Cov}\left[\sum_k p_k X_{1,k} \, , \, \sum_l p_k X_{1,l} \right] = \sum_{k,l} p_k p_l C_{kl} = p \cdot Cp.$$

Ist Y ein $N(0, C)$-verteilter Zufallsvektor, dann ist $N(0, p \cdot Cp)$ die Verteilung von $p \cdot Y$. Mithilfe der Cramér-Wold Device folgt also

$$(S_n - E[S_n]) / \sqrt{n} \quad \xrightarrow{\mathcal{D}} \quad Y.$$

□

5.3.2 Gauß-Prozesse

Sei $I = \mathbb{Z}_+$ oder $I = [0, \infty)$. Ein stochastischer Prozeß $(X_t)_{t \in I}$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ heißt **Gauß-Prozeß**, falls die gemeinsamen Verteilungen von endlichen vielen der Zufallsvariablen X_t ($t \in I$) multivariate Normalverteilungen sind.

Die wichtigste Beispielklasse von zeitdiskreten Gauß-Prozessen sind *autoregressive Prozesse*, die zur Modellierung von Zeitreihen eingesetzt werden. Seien X_0 und $Z_n, n \in \mathbb{N}$, unabhängige reellwertige Zufallsvariablen mit $Z_n \sim N(0,1)$ für alle n. Der durch das „stochastische Bewegungsgesetz“

$$X_n = \underbrace{\alpha X_{n-1}}_{\substack{\text{lineares} \\ \text{Bewegungsgesetz}}} + \underbrace{\varepsilon Z_n}_{\substack{\text{zufällige Störung,} \\ \text{Rauschen}}}, \qquad n \in \mathbb{N}, \tag{5.3.1}$$

definierte stochastische Prozess $(X_n)_{n=0,1,2,\ldots}$ heißt **autoregressiver Prozess AR(1)** mit Parametern $\varepsilon, \alpha \in \mathbb{R}$. Im allgemeineren autoregressiven Modell **AR(p)**, $p \in \mathbb{N}$, mit Parametern $\varepsilon, \alpha_1, \ldots, \alpha_p \in \mathbb{R}$ lautet das Bewegungsgesetz

$$X_n = \sum_{i=1}^{p} \alpha_i X_{n-i} + \varepsilon Z_n, \qquad n \geq p,$$

wobei die Rauschterme Z_n unabhängig von den Startwerten $X_0, X_1, \ldots, X_{p-1}$ sind.

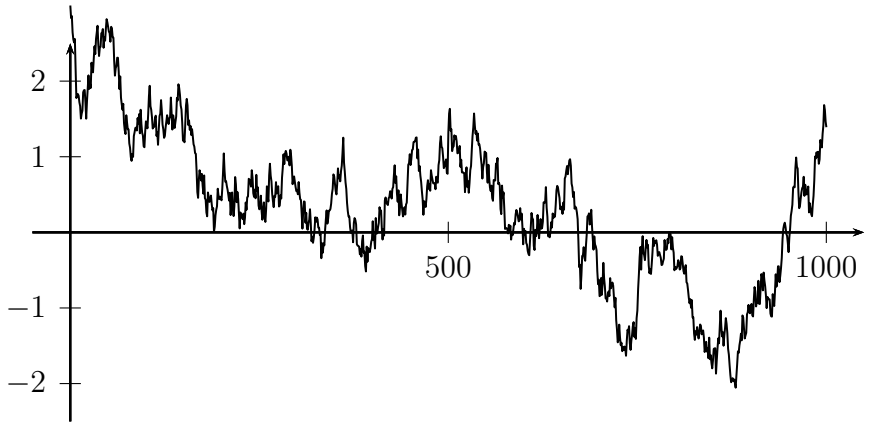


Abbildung 5.4: Trajektorie eines AR(1)-Prozesses mit Parametern $\alpha = 0.8$ und $\varepsilon^2 = 1.5$.

Satz 5.15 (Autoregressive Prozesse als Gaußprozesse). *Ist die gemeinsame Verteilung der Startwerte $X_0, X_1, \ldots, X_{p-1}$ eine multivariate Normalverteilung, dann ist der AR(p)-Prozess ein Gauß-Prozeß.*

Beweis. Mit Induktion folgt für alle $n \geq p$, dass Z_n unabhängig von $X_0, X_1, \ldots, X_{n-1}$ ist, und dass die Zufallsvektoren $(X_0, X_1, \ldots, X_{n-1}, Z_n)$ und $(X_0, X_1, \ldots, X_n)$ multivariat normalverteilt sind. □

Der AR(1)-Prozess ist eine *Markovkette* mit Übergangswahrscheinlichkeiten $p(x, \cdot) = N(\alpha x, \varepsilon^2)$, siehe unten. Das folgende Korollar fasst einige grundlegende Eigenschaften des AR(1) Modells zusammen.

Lemma 5.16 (Gleichgewicht und exponentieller Abfall der Korrelationen). *Für den AR(1)-Prozess mit Parametern ε, α und $m \in \mathbb{R}, \sigma > 0$ gilt:*

(1). $X_{n-1} \sim N(m, \sigma^2) \quad \Longrightarrow \quad X_n \sim N(\alpha m, \alpha^2\sigma^2 + \varepsilon^2)$.

(2). Für $|\alpha| < 1$ ist die Verteilung $\mu = N(0, \frac{\varepsilon^2}{1-\alpha^2})$ ein Gleichgewicht, d.h.

$$X_0 \sim \mu \quad \Longrightarrow \quad X_n \sim \mu \qquad \forall n \in \mathbb{N}.$$

Bei Startverteilung $P \circ X_0^{-1} = \mu$ gilt:

$$\mathrm{Cov}[X_n, X_{n-k}] \quad = \quad \alpha^k \cdot \frac{\varepsilon^2}{1-\alpha^2} \qquad \text{für alle } 0 \leq k \leq n.$$

Beweis. Gilt $X_{n-1} \sim N(m, \sigma^2)$, dann ist (X_{n-1}, Z_n) bivariat normalverteilt, also ist auch die Linearkombination $X_n = aX_{n-1} + \varepsilon Z_n$ normalverteilt. Der Erwartungswert und die Varianz von

X_n ergeben sich aus (5.3.1). Der Beweis der übrigen Aussagen wird dem Leser als Übungsaufgabe überlassen. □

Auch ein allgemeiner AR(p)-Prozess lässt sich als Markovkette interpretieren, wenn man statt der Zufallsvariablen X_n die zeitliche Entwicklung der Zufallsvektoren

$$\widetilde{X}_n \;:=\; (X_n - p + 1, X_n - p + 2, \ldots, X_n)$$

betrachtet. Man kann dann ähnliche Aussagen wie in Lemma 5.16 herleiten.

5.3.3 Vom Random Walk zur Brownschen Bewegung

Sei $S_n = X_1 + ... + X_n$, wobei die X_i unabhängige Zufallsvariablen in $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ mit

$$E[X_i] = 0 \qquad \text{und} \qquad \mathrm{Var}[X_i] = 1$$

sind. Beispielsweise ist S_n ein klassischer Random Walk. Um einen stochastischen Prozess in kontinuierlicher Zeit zu erhalten, interpolieren wir $n \mapsto S_n$ linear. Anschließend reskalieren wir in Raum und Zeit, und setzen für $m \in \mathbb{N}$:

$$\widetilde{S}_t^{(m)} \;:=\; \frac{1}{\sqrt{m}}\, S_{mt}, \qquad t \in \mathbb{R}_+.$$

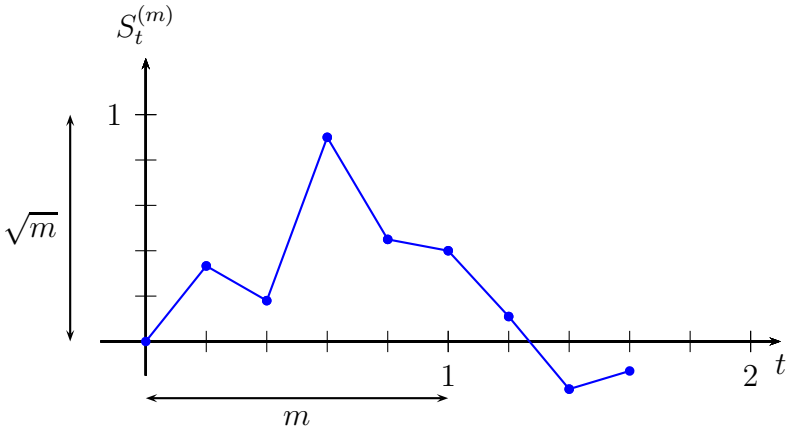


Abbildung 5.5: Raum-Zeit-Reskalierung eines Random Walks.

Aus dem Zentralen Grenzwertsatz folgt für $m \to \infty$:

$$\widetilde{S}_t^{(m)} \quad = \quad \sqrt{t}\,\frac{1}{\sqrt{mt}}\, S_{mt} \quad \xrightarrow{\mathcal{D}} \quad N(0,t) \qquad \text{für jedes feste } t \in \mathbb{R}_+,$$

d.h. die eindimensionalen Randverteilungen der Prozesse $\widetilde{S}^{(m)} = (\widetilde{S}_t^{(m)})_{t \geq 0}$ konvergieren.

Allgemeiner zeigt man mithilfe des multivariaten zentralen Grenzwertsatzes, dass auch endlich dimensionale Randverteilungen schwach konvergieren. Für $0 = t_0 < t_1 < \ldots < t_k$ mit $k \in \mathbb{N}$ und $mt_i \in \mathbb{Z}_+$ sind die Inkremente $\widetilde{S}_{t_1}^{(m)} - \widetilde{S}_{t_0}^{(m)}$, $\widetilde{S}_{t_2}^{(m)} - \widetilde{S}_{t_1}^{(m)}, \ldots, \widetilde{S}_{t_k}^{(m)} - \widetilde{S}_{t_{k-1}}^{(m)}$ unabhängige Zufallsvariablen, die sich als Summen der Zufallsvariablen X_i über disjunkte Bereiche darstellen lassen. Aus dem eindimensionalen zentralen Grenzwertsatz folgt

$$\widetilde{S}_t^{(m)} - \widetilde{S}_r^{(m)} \quad \xrightarrow{\mathcal{D}} \quad N(0, t-r) \qquad \text{für alle } 0 \le r \le t,$$

und mit dem multivariaten ZGS ergibt sich die Konvergenz der gemeinsamen Verteilungen der Inkremente gegen ein Produkt von eindimensionalen Normalverteilungen:

$$\left(\widetilde{S}_{t_1}^{(m)} - \widetilde{S}_{t_0}^{(m)}, \widetilde{S}_{t_2}^{(m)} - \widetilde{S}_{t_1}^{(m)}, \ldots, \widetilde{S}_{t_k}^{(m)} - \widetilde{S}_{t_{k-1}}^{(m)}\right) \quad \xrightarrow{\mathcal{D}} \quad \prod_{i=1}^{k} N(0, t_i - t_{i-1}). \tag{5.3.2}$$

Dies motiviert die folgende Definition, die auf N. Wiener zurückgeht:

Definition (Brownsche Bewegung). *Ein stochastischer Prozess $B_t : \Omega \to \mathbb{R}$, $t \in [0, \infty)$, auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, heißt **Brownsche Bewegung** , falls gilt:*

(1). Für jede Partition $0 = t_0 < t_1 < \ldots < t_k$ mit $k \in \mathbb{N}$ sind die Inkremente $B_{t_{i+1}} - B_{t_i}$ unabhängige Zufallsvariablen mit Verteilung

$$B_{t_{i+1}} - B_{t_i} \quad \sim \quad N(0, t_{i+1} - t_i).$$

(2). Die Funktion $t \mapsto B_t(\omega)$ ist für P-fast alle ω stetig.

Hierbei ist ein zeitstetiger stochastischer Prozess einfach eine Kollektion von Zufallsvariablen B_t, $t \in \mathbb{R}_+$, die auf einem gemeinsamen Wahrscheinlichkeitsraum definiert sind. Eine Brownsche Bewegung ist ein zeitstetiger Gauß-Prozess.

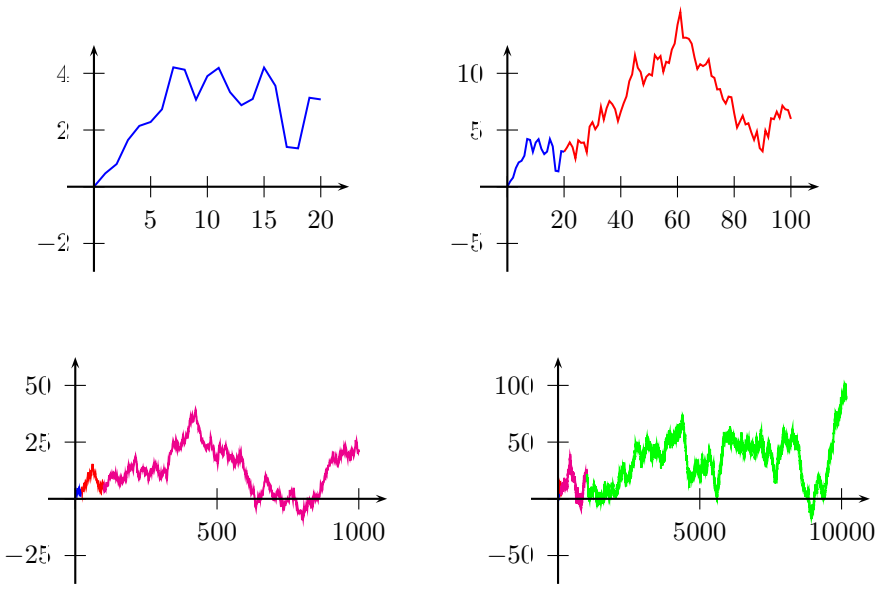


Abbildung 5.6: Vom Random Walk zur Brownschen Bewegung.

Die Existenz einer Brownschen Bewegung wird in der Vorlesung »Stochastische Prozesse« gezeigt. Aus (5.3.2) folgt, dass die endlichdimensionalen Randverteilungen der reskalierten Random Walks gegen die entsprechenden Randverteilungen einer Brownschen Bewegung $(B_t)_{t\in[0,\infty)}$ mit Startwert $B_0 = 0$ konvergieren:

$$\left(\widetilde{S}_{t_1}^{(m)}, \widetilde{S}_{t_2}^{(m)}, ..., \widetilde{S}_{t_k}^{(m)}\right) \quad \xrightarrow{\mathcal{D}} \quad (B_{t_1}, ..., B_{t_k}), \qquad \text{für alle } 0 \le t_1 < t_2 < \ldots < t_k, k \in \mathbb{N}.$$

Eine noch allgemeinere Aussage liefert ein **funktionaler zentralen Grenzwertsatz** auf dem Banachraum $C([0,1],\mathbb{R})$ (*Invarianzprinzip von Donsker*): Der gesamte stochastische Prozess $(\widetilde{S}_t^{(m)})_{0\le t\le 1}$ konvergiert in Verteilung gegen eine Brownsche Bewegung $(B_t)_{0\le t\le 1}$. Mehr dazu in den weiterführenden Vorlesungen »Stochastische Prozesse« und »Grundzüge der stochastischen Analysis«.

5.4 Schätzer und Tests in Gauß-Modellen

5.4.1 Parameterschätzung im Gauß-Modell

Angenommen, wir beobachten reellwertige Messwerte (Stichproben, Daten), die von einer unbekannten Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}$ stammen. Ziel der Statistik ist es, Rückschlüsse auf die zugrundeliegende Verteilung aus den Daten zu erhalten. Im einfachsten Modell (Gauß-

modell) nimmt man an, dass die Daten unabhängige Stichproben von einer Normalverteilung mit unbekanntem Mittelwert und/oder Varianz sind:

$$\mu = N(m, v), \quad m, v \text{ unbekannt.}$$

Eine partielle Rechtfertigung für die Normalverteilungsannahme liefert der zentrale Grenzwertsatz. Letztendlich muss man aber in jedem Fall überprüfen, ob eine solche Annahme gerechtfertigt ist.Ein erstes Ziel ist es nun, den Wert von m auf der Basis von n unabhängigen Stichproben $X_1(\omega) = x_1, \ldots, X_n(\omega) = x_n$ zu schätzen, und zu quantifizieren.

Problemstellung: Schätzung des Erwartungswerts

- Schätze m auf der Basis von n unabhängigen Stichproben $X_1(\omega), ..., X_n(\omega)$ von μ.
- Herleitung von Konfidenzintervallen.

Im mathematischen Modell interpretieren wir die Beobachtungswerte als Realisierungen von unabhängigen Zufallsvariablen $X_1, \ldots, X_n$. Da wir die tatsächliche Verteilung nicht kennen, untersuchen wir alle in Betracht gezogenen Verteilungen simultan:

$$X_1, \ldots, X_n \sim N(m, v) \quad \text{unabhängig unter } P_{m,v}. \tag{5.4.1}$$

Ein naheliegender Schätzer für m ist der *empirische Mittelwert*

$$\overline{X}_n(\omega) := \frac{X_1(\omega) + ... + X_n(\omega)}{n} .$$

Wir haben oben bereits gezeigt, dass dieser Schätzer *erwartungstreu (unbiassed)* und *konsistent* ist, d.h. für alle m, v gilt:

$$E_{m,v}[\overline{X}_n] = m$$

und

$$\overline{X}_n \to m \quad P_{m,v}\text{-stochastisch für } n \to \infty.$$

Wie wir den Schätzfehler quantifizieren hängt davon ab, ob wir die Varianz kennen.

Schätzung von m bei bekannter Varianz v.
Um den Schätzfehler zu kontrollieren, berechnen wir die Verteilung von $\overline{X}_n$:

$$\begin{aligned} X_i \sim N(m, v) \text{ unabh.} \quad &\Rightarrow \quad X_1 + ... + X_n \sim N(nm, nv) \\ &\Rightarrow \quad \overline{X}_n \sim N(m, \frac{v}{n}) \\ &\Rightarrow \quad \frac{\overline{X}_n - m}{\sqrt{v/n}} \sim N(0, 1) \end{aligned}$$

Bezeichnet Φ die Verteilungsfunktion der Standardnormalverteilung, dann erhalten wir

$$P_{m,v}\left[|\overline{X}_n - m| < q\sqrt{\frac{v}{n}}\right] = N(0,1)(-q,q) = 2\left(\Phi(q) - \frac{1}{2}\right) \quad \text{für alle } m \in \mathbb{R}.$$

Satz 5.17. *Im Gaußmodell (5.4.1) mit bekannter Varianz v ist das zufällige Intervall*

$$\left(\overline{X}_n - \Phi^{-1}(\alpha)\sqrt{\frac{v}{n}}\,,\ \overline{X}_n + \Phi^{-1}(\alpha)\sqrt{\frac{v}{n}}\right)$$

*ein $(2\alpha-1)\cdot 100\%$ **Konfidenzintervall** für m, d.h.*

$$P_{m,v}[m \in \textit{Intervall}] \geq 2\alpha - 1 \qquad \textit{für alle } m \in \mathbb{R}.$$

Man beachte, dass die Länge des Konfidenzintervalls in diesem Fall nicht von den beobachteten Stichproben abhängt!

Schätzung von m bei unbekannter Varianz v. In Anwendungen ist meistens die Varianz unbekannt. In diesem Fall können wir das Intervall oben nicht verwenden, da es von der unbekannten Varianz v abhängt. Stattdessen schätzen wir m und v simultan, und konstruieren ein Konfidenzintervall für m mithilfe beider Schätzwerte. Erwartungstreue Schätzer für m und v sind

$$\overline{X}_n \ = \ \frac{1}{n}\sum_{i=1}^n X_i \quad \text{und} \quad V_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2 \ .$$

Um ein Konfidenzintervall für m zu erhalten, bestimmen wir mithilfe des Transformationssatzes die gemeinsame Verteilung von $\overline{X}_n$ und V_n:

Lemma 5.18. *$\overline{X}_n$ und V_n sind unabhängig unter $P_{m,v}$ mit Verteilung*

$$\overline{X}_n \ \sim \ N\left(m, \frac{v}{n}\right) \quad , \quad \frac{n-1}{v}V_n \sim \chi^2(n-1) \ .$$

Beweis. Wir führen eine lineare Koordinatentransformation $Y = OX$ durch, wobei O eine orthogonale $n \times n$-Matrix vom Typ

$$O = \begin{pmatrix} \frac{1}{\sqrt{n}} \cdots \frac{1}{\sqrt{n}} \\ \text{beliebig} \end{pmatrix}$$

ist. Eine solche Matrix erhalten wir durch Ergänzen des normierten Vektors $(\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}})$ zu einer Orthonormalbasis des $\mathbb{R}^n$. In den neuen Koordinaten gilt:

$$\begin{aligned} \overline{X}_n \qquad &= \qquad \frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{\sqrt{n}}Y_1, \quad \text{und} \\ (n-1)V_n \qquad &= \qquad \sum_{i=1}^n (X_i - \overline{X}_n)^2 = \sum_{i=1}^n X_i^2 - n\overline{X}_n^2 = ||X||^2_{\mathbb{R}^n} - n\overline{X}_n^2 \\ &\overset{O \text{ orthogonal}}{=} \quad ||Y||^2_{\mathbb{R}^n} - Y_1^2 = \sum_{i=2}^n Y_i^2 \ . \end{aligned}$$

Da die Zufallsvariablen X_i $(1 \leq i \leq n)$ unabhängig und $N(m, v)$-verteilt sind, ist der Zufallsvektor $X = (X_1, ..., X_n)$ multivariat normalverteilt mit Mittel $(m, \ldots, m)$ und Kovarianzmatrix $v \cdot I_n$. Nach dem Transformationssatz folgt

$$Y \sim N\left(O \begin{pmatrix} m \\ \vdots \\ m \end{pmatrix}, v \cdot OI_nO^T\right) = N\left(\begin{pmatrix} m\sqrt{n} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, v \cdot I_n\right).$$

Also sind $Y_1, ..., Y_n$ unabhängige Zufallsvariablen mit Verteilungen

$$Y_1 \sim N(m\sqrt{n}, v) \quad , \quad Y_i \sim N(0, v) \quad \text{für } i \geq 2.$$

Es folgt, dass

$$\overline{X}_n = \frac{Y_1}{\sqrt{n}} \quad \text{und} \quad \frac{n-1}{v} V_n = \sum_{i=2}^{n} \left(\frac{Y_i}{\sqrt{v}}\right)^2$$

unabhängige Zufallsvariablen mit Verteilungen $N(m, \frac{v}{n})$ bzw. $\chi^2(n-1)$ sind. □

Bei bekannter Varianz v hatten wir Konfidenzintervalle für m vom Typ $\overline{X}_n \pm q \cdot \sqrt{\frac{v}{n}}$ erhalten, wobei q ein geeignetes Quantil der Standardnormalverteilung ist. Daher liegt es nahe, zu versuchen, bei unbekannter Varianz Konfidenzintervalle vom Typ $\overline{X}_n \pm q \cdot \sqrt{\frac{V_n}{n}}$ herzuleiten. Es gilt:

$$P_{m,v}\left[|\overline{X}_n - m| \geq q\sqrt{\frac{V_n}{n}}\right] = P_{m,v}[|T_{n-1}| \geq q] \qquad \text{mit}$$

$$T_{n-1} := \frac{\sqrt{n} \cdot (\overline{X}_n - m)}{\sqrt{V_n}} .$$

Die Zufallsvariable T_{n-1} heißt **Studentsche t-Statistik mit $n-1$ Freiheitsgraden**.[1] Unsere Überlegungen zeigen, dass wir aus Quantilen der Studentschen t-Statistik Konfidenzintervalle für das Gaußmodell herleiten können. Wir müssen nur noch die Verteilung von T_n berechnen:

Satz 5.19 (Student[2])**.** *Die Verteilung von T_n ist absolutstetig mit Dichte*

$$f_{T_n}(t) \quad = \quad B\left(\frac{1}{2}, \frac{n}{2}\right)^{-1} \cdot n^{-1/2} \cdot \left(1 + \frac{t^2}{2}\right)^{-n/2} \qquad (t \in \mathbb{R}).$$

[1] In der Statistik bezeichnet man eine messbare Funktion der Beobachtungsdaten als Statistik - ein (Punkt-) Schätzer ist eine Statistik, die zum Schätzen eines unbekannten Parameters verwendet wird, ein Konfidenzintervall nennt man auch Intervallschätzer.

[2] Synonym von W. S. Gosset, der als Angestellter der Guiness-Brauerei nicht publizieren durfte.

***»Studentsche** t**-Verteilung mit** n **Freiheitsgraden«**. Hierbei ist*

$$B\left(\frac{1}{2},\frac{n}{2}\right)=\frac{1}{\sqrt{n}}\int_{-\infty}^{\infty}(1+s^2)^{-\frac{n}{2}}ds$$

*die **Eulersche Beta-Funktion**, die als Normierungsfaktor auftritt.*
Insbesondere ist das zufällige Intervall

$$\overline{X}_n \pm q\cdot\sqrt{\frac{V_n}{n}}$$

ein $100\cdot(1-2\alpha)\%$ *Konfidenzintervall für* m*, falls*

$$q=F_{T_{n-1}}^{-1}(1-\alpha)$$

ein $(1-\alpha)$*-Quantil der* t*-Verteilung mit* $n-1$ *Freiheitsgraden ist.*

Beweis. Direkt oder mithilfe des Transformationssatzes zeigt man: Sind Z und Y unabhängige Zufallsvariablen mit Verteilungen $N(0,1)$ bzw. $\chi^2(n-1)$, dann ist $Z/\sqrt{\frac{1}{n-1}Y}$ absolutstetig mit dichte $f_{T_{n-1}}$.
Der Satz folgt dann nach Lemma 5.18 mit

$$Z:=\frac{\overline{X}_n-m}{\sqrt{v/n}} \quad \text{und} \quad Y:=\frac{n-1}{v}V_n\ .$$

□

Bemerkung (Nichtparametrische und Verteilungsunabhängige Konfidenzintervalle). In Anwendungen ist es oft unklar, ob eine Normalverteilungsannahme an die Beobachtungswerte gerechtfertigt ist. Zudem können einzelne größere Ausreißer in den Daten (z.B. aufgrund von Messfehlern) das Stichprobenmittel relativ stark beeinflussen. Der Stichprobenmedian ist dagegen in den meisten Fällen ein deutlich stabilerer Schätzwert für den Median der zugrundeliegenden Verteilung, und die in Abschnitt 2.1 hergeleiteten, auf Ordnungsstatistiken basierenden, Konfidenzintervalle für den Median und andere Quantile werden ebenfalls in der Regel weniger stark durch Ausreißer beeinflusst. Zudem gelten diese Konfidenzintervalle simultan für alle stetigen Verteilungen. Ist man sich daher nicht sicher, ob eine Normalverteilungsannahme aufgrund der Daten gerechtfertigt ist, empfiehlt es sich, auf die stabileren Ordnungsintervalle zurückzugreifen.

Beispiel. (NOCH EINZUFÜGEN)

5.4.2 Hypothesentests

In Anwendungen werden statistische Aussagen häufig nicht über Konfidenzintervalle, sondern als Hypothesentest formuliert. Mathematisch passiert dabei nichts wirklich Neues – es handelt sich nur um eine durch praktische Erwägungen motivierte Umformulierung derselben Resultate: Angenommen, wir haben n unabhängige reellwertige Stichproben $X_1, ..., X_n$ von einer unbekannten Verteilung vorliegen und wir gehen davon aus, daß die zugrundeliegende Verteilung aus einer Familie μ_θ $(\theta \in \Theta)$ von Wahrscheinlichkeitsverteilungen kommt, z.B. der Familie aller Normalverteilungen $\mu_{m,v}, \theta = (m, v) \in \mathbb{R} \times \mathbb{R}_+$. Die gemeinsame Verteilung von $X_1, \ldots, X_n$ ist dann das Produktmaß $\mu_\theta^n = \bigotimes_{i=1}^{n} \mu_\theta$. Sei nun Θ_0 eine Teilmenge des Parameterbereichs. Wir wollen entscheiden zwischen der

$$\textit{Nullhypothese } H_0\text{:} \qquad \text{»}\theta \in \Theta_0\text{«}$$

und der

$$\textit{Alternative } H_1\text{:} \qquad \text{»}\theta \notin \Theta_0\text{«}$$

Ein **Hypothesentest** für ein solches Problem ist bestimmt durch eine messbare Teilmenge $C \subseteq \mathbb{R}^n$ (den **Verwerfungsbereich**) mit zugehöriger Entscheidungsregel:

$$\text{Akzeptiere } H_0 \iff (X_1, ..., X_n) \notin C.$$

Beispiel (t-Test). Seien $X_1, X_2, \ldots, X_n$ unabhängige Stichproben von einer Normalverteilung mit unbekanntem Parameter $(m, v) \in \Theta = \mathbb{R} \times \mathbb{R}^+$. Wir wollen testen, ob der Mittelwert der Verteilung einen bestimmten Wert m_0 hat:

$$\textit{Nullhypothese } H_0\text{:} \qquad \text{»}m = m_0\text{«}\,, \qquad \Theta_0 = \{m_0\} \times \mathbb{R}^+\,.$$

Ein solches Problem tritt z.B. in der Qualitätskontrolle auf, wenn man überprüfen möchte, ob ein Sollwert m_0 angenommen wird. Eine andere Anwendung ist der Vergleich zweier Verfahren, wobei X_i die Differenz der mit beiden Verfahren erhaltenen Messwerte ist. Die Nullhypothese mit $m_0 = 0$ besagt hier, daß kein signifikanter Unterschied zwischen den Verfahren besteht.

Im *t–Test* für obiges Testproblem wird die Nullhypothese akzeptiert, falls der Betrag der *Student-schen t-Statistik* unterhalb einer angemessen zu wählenden Konstanten c liegt, bzw. verworfen, falls

$$|T_{n-1}| \;=\; \left| \frac{\sqrt{n} \cdot (\overline{X}_n - m_0)}{\sqrt{V_n}} \right| \;>\; c$$

gilt.

Seien nun allgemein $X_1, X_2, \ldots$ unter P_θ unabhängige Zufallsvariablen mit Verteilung μ_θ. Bei einem Hypothesentest können zwei Arten von Fehlern auftreten:

Fehler 1. Art: *H_0 wird verworfen, obwohl wahr.* Die Wahrscheinlichkeit dafür beträgt:

$$P_\theta[(X_1, ..., X_n) \in C] = \mu_\theta^n[C] \,, \qquad \theta \in \Theta_0.$$

Fehler 2. Art: *H_0 wird akzeptiert, obwohl falsch.* Die Wahrscheinlichkeit beträgt:

$$P_\theta[(X_1, ..., X_n) \notin C] = \mu_\theta^n[C^C] \,, \qquad \theta \in \Theta \setminus \Theta_0.$$

Obwohl das allgemeine Testproblem im Prinzip symmetrisch in H_0 und H_1 ist, interpretiert man beide Fehler i.a. unterschiedlich. Die Nullhypothese beschreibt in der Regel den Normalfall, die Alternative eine Abweichung oder einen zu beobachtenden Effekt. Da ein Test Kritiker überzeugen soll, sollte die Wahrscheinlichkeit für den Fehler 1. Art (Effekt prognostiziert, obgleich nicht vorhanden) unterhalb einer vorgegebenen (kleinen) Schranke α liegen. Die Wahrscheinlichkeit

$$\mu_\theta^n[C] \,, \qquad \theta \in \Theta \setminus \Theta_0 \,,$$

daß kein Fehler 2. Art auftritt, sollte unter dieser Voraussetzung möglichst groß sein.

Definition. *Die Funktion*

$$G(\theta) \;=\; P_\theta[(X_1, ..., X_n) \in C] \;=\; \mu_\theta^n[C]$$

*heißt **Gütefunktion** des Tests. Der Test hat **Niveau** α, falls*

$$G(\theta) \;\leq\; \alpha \quad \textit{für alle } \theta \in \Theta_0$$

*gilt. Die Funktion $G(\theta)$ mit $\theta \in \Theta_1$ heißt **Macht** des Tests.*

Aus Satz 5.19 und der Symmetrie der Studentschen t-Verteilung folgt unmittelbar:

Korollar 5.20. *Der Studentsche t-Test hat Niveau α falls c ein $(1 - \frac{\alpha}{2})$-Quantil der Studentschen t-Verteilung mit $n - 1$ Freiheitsgraden ist.*

Allgemeiner gilt:

Satz 5.21 (Korrespondenz Konfidenzintervalle $\leftrightarrow$ Hypothesentests). *Für einen reellwertigen Parameter $\gamma = c(\theta)$, ein Irrtumsniveau $\alpha \in (0, 1)$, und messbare Abbildungen (Statistiken) $\hat{\gamma}, \varepsilon : \mathbb{R}^n \to \mathbb{R}$ sind äquivalent:*

(i) Das Intervall

$$[\hat{\gamma}(X_1, \ldots, X_n) - \varepsilon(X_1, \ldots, X_n)\,,\, \hat{\gamma}(X_1, \ldots, X_n) + \varepsilon(X_1, \ldots, X_n)]$$

ist ein $(1-\alpha) \cdot 100\,\%$ *Konfidenzintervall für* γ*.*

(ii) Für jedes $\gamma_0 \in \mathbb{R}$ *ist*

$$C \;=\; \{(x_1, ..., x_n) \,:\, |\hat{\gamma}(x_1, \ldots, x_n) - \gamma_0| > \varepsilon(x_1, \ldots, x_n)\}$$

der Verwerfungsbereich eines Test der Nullhypothese $\gamma = \gamma_0$ *zum Niveau* α*.*

Beweis. Das Intervall ist genau dann ein Konfidenzintervall für γ zum Irrtumsniveau α, wenn

$$P_\theta\,[|\hat{\gamma}(X_1, \ldots, X_n) - c(\theta)| > \varepsilon(X_1, ..., X_n)] \;\leq\; \alpha \qquad \forall\, \theta \in \Theta$$

gilt, also wenn der entsprechende Test der Nullhypothesen $c(\theta) = \gamma_0$ für jedes γ_0 Niveau α hat. □

Kapitel 6

Entropie und große Abweichungen

Um Wahrscheinlichkeiten seltener Ereignisse zu untersuchen, geht man häufig zu einer neuen absolutstetigen Wahrscheinlichkeitsverteilung über, bezüglich der das relevante Ereignis nicht mehr selten ist. Der Maßwechsel geschieht dabei typischerweise mit einer exponentiellen Dichte. Auf diese Weise erhält man unter anderem asymptotische Aussagen über die Wahrscheinlichkeiten großer Abweichungen. Eine zentrale Rolle spielt dabei der Begriff der relativen Entropie, die die statistische Unterscheidbarkeit zweier Wahrscheinlichkeitsverteilungen misst. Anwendungen liegen in der Asymptotik von Likelihood basierten Schätz- und Testverfahren, und der asymptotischen Effizienz von Importance Sampling Schätzern.

Asymptotische Aussagen über Wahrscheinlichkeiten, die wir in diesem Abschnitt herleiten werden, betreffen meistens nur die exponentielle Abfallrate. Subexponentiell fallend oder wachsende Faktoren werden ignoriert. Wir führen einen entsprechenden Äquivalenzbegriff für Folgen auf der exponentiellen Skala ein:

Definition (Asymptotische exponentielle Äquivalenz von Folgen). *Zwei Folgen $(a_n)_{n\in\mathbb{N}}$ und $(b_n)_{n\in\mathbb{N}}$ von positiven reellen Zahlen heißen **asymptotisch exponentiell äquivalent ($a_n \simeq b_n$)**, falls*

$$\frac{1}{n}\log\frac{a_n}{b_n} \;=\; \frac{1}{n}\left(\log a_n - \log b_n\right) \;\longrightarrow\; 0 \qquad \textit{für } n\to\infty.$$

Beispielsweise gilt $n^{-k}\exp(-cn) \simeq \exp(-cn)$ für alle $k, c \in \mathbb{R}$.

Um exponentielle Äquivalenz zu zeigen werden wir häufig separat eine Abschätzung nach oben („obere Schranke") und eine Abschätzung nach unten („untere Schranke") beweisen. Entsprechend schreiben wir „$a_n \preceq b_n$", falls

$$\limsup_{n\to\infty}\; \frac{1}{n}\log\frac{a_n}{b_n} \quad\leq\quad 0$$

ist. Beispielsweise gilt für reelle Zahlen c, d, k, l:

$$n^{-k} \exp(-cn) \preceq n^{-l} \exp(-dn) \qquad \Longleftrightarrow \qquad c \geq d.$$

6.1 Exponentielle Familien und der Satz von Cramér

In diesem Abschnitt wollen wir die exponentielle Abfallrate der Wahrscheinlichkeiten großer Abweichungen vom Gesetz der großen Zahlen identifizieren. Um die Chernoff-Abschätzung durch eine asymptotische untere Schranke zu vervollständigen, verwenden wir einen Maßwechsel zu einer Verteilung aus einer exponentiellen Familie.

6.1.1 Exponentielle Familien

Sei ν ein positives Maß auf $(S, \mathcal{S})$, $U : S \to \mathbb{R}^d$ eine messbare Funktion, und

$$Z(t) \quad = \quad \int e^{t \cdot U} \, d\nu, \qquad t \in \mathbb{R}^d,$$

die momentenerzeugende Funktion von U mit Definitionsbereich

$$\Theta \quad = \quad \{t \in \mathbb{R}^d \, : \, Z(t) < \infty\}.$$

Für $t \in \Theta$ sei $\Lambda(t) = \log Z(t)$ die kumulantenerzeugende Funktion.

Definition (Exponentielle Familie). *Die Familie der Wahrscheinlichkeitsverteilungen*

$$\mu_t(dx) \quad := \quad \frac{1}{Z(t)} \, e^{t \cdot U(x)} \, \nu(dx) \quad = \quad e^{t \cdot U(x) - \Lambda(t)} \, \nu(dx), \qquad t \in \Theta,$$

heißt ***exponentielle Familie zu*** ν ***und*** U.

Bemerkung (Boltzmann-Verteilung). In der statistischen Physik treten exponentielle Familien als Gleichgewichtsverteilungen auf. Beispielsweise ist die Verteilung im thermodynamischen Gleichgewicht in einem abgeschlossenen System bei inverser Temperatur $\beta = 1/T$ gleich μ_β, wobei ν die Gleichverteilung bzw. das Lebesguemaß auf dem Zustandsraum, und $U(x) = -H(x)$ die negative Energie des Zustandes x ist. Die Normierungskonstante $Z(\beta)$ heißt in der statistischen Physik *Partitionsfunktion*.

Wir betrachten nun einige elementare Beispiele von exponentiellen Familien:

Beispiel. (1). **Exponential- und Gammaverteilungen.** Ist ν die Exponentialverteilung mit Parameter $\lambda > 0$, und $U(x) = -x$, dann ist $M(t)$ für $t > -\lambda$ endlich, und es gilt

$$\mu_t \quad = \quad \mathrm{Exp}(\lambda + t) \qquad \text{für alle } t > -\lambda.$$

Die exponentielle Familie besteht also aus allen Exponentialverteilungen.
Ist $\nu = \Gamma(\alpha, \lambda)$ eine Gammaverteilung, dann gilt entsprechend $\mu_t = \Gamma(\alpha, \lambda + t)$.

(2). **Bernoulli-, Binomial- und Poisson-Verteilungen.** Ist ν die Bernoulli-Verteilung mit Parameter p und $U(k) = k$, dann gilt $\mu_t(1) = p_t$ mit

$$p_t \quad = \quad \frac{e^t p}{e^t p + 1 - p} \quad = \quad \frac{p}{p + (1-p)e^{-t}},$$

d.h. μ_t ist die Bernoulliverteilung mit Parameter p_t. Entsprechend gilt für $U(k) = k$:

$$\begin{aligned} \nu \;&=\; \mathrm{Bin}(n,p) \quad \Rightarrow \quad \mu_t \;=\; \mathrm{Bin}(n,p_t), \qquad \text{und} \\ \nu \;&=\; \mathrm{Poisson}(\lambda) \quad \Rightarrow \quad \mu_t \;=\; \mathrm{Poisson}(\lambda e^t). \end{aligned}$$

Die exponentielle Familie besteht also jeweils aus allen Bernoulli-Verteilungen, Binomialverteilungen mit festem n, bzw. Poisson-Verteilungen.

(3). **Normalverteilungen.** Ist $\nu = N(m, C)$ eine d-dimensionale Normalverteilung, und $U(x) = x$, dann gilt $\mu_t = N(m + Ct, C)$ für $t \in \mathbb{R}^d$. Im nichtdegenerierten Fall enthält die exponentielle Familie also alle Normalverteilungen mit fester Kovarianzmatrix C. Für $d = 1, \nu = N(m, \sigma^2)$, und

$$U(x) \quad = \quad -\frac{(x-m)^2}{2}$$

erhält man

$$\mu_t \quad = \quad N\left(m, \left(\frac{1}{\sigma^2} + \frac{1}{t}\right)^{-1}\right) \qquad \text{für } t > 0,$$

d.h. die exponentielle Familie besteht aus Normalverteilungen mit festem Mittelwert m. Entsprechend kann man die Familie der eindimensionalen Normalverteilungen als zweiparametrige exponentielle Familie bzgl. einer Referenz-Normalverteilung interpretieren.

Wir beschränken uns nun auf den Fall $d = 1$. Sei $(\mu_t)_{t\in\Theta}$ eine einparametrige exponentielle Familie zu ν und U, und sei $\overset{\circ}{\Theta} = \Theta \setminus \partial\Theta$ der offene Kern des Definitionsbereichs.

Lemma 6.1 (Eigenschaften exponentieller Familien).

(1). Es gilt $Z \in C^\infty(\overset{\circ}{\Theta})$. Für $t \in \overset{\circ}{\Theta}$ existieren die Erwartungswerte und Varianzen

$$m(t) \quad = \quad \int U \, d\mu_t \qquad \textit{bzw.} \qquad v(t) \quad = \quad \mathrm{Var}_{\mu_t}[U],$$

und es gilt $m(t) = \Lambda'(t)$ und $v(t) = \Lambda''(t)$.

(2). Die Funktion m ist auf $\overset{\circ}{\Theta}$ beliebig oft differenzierbar und monoton wachsend. Ist U nicht μ-fast überall konstant, dann ist m sogar strikt monoton wachsend. Im Fall $\Theta = \mathbb{R}$ gilt zudem

$$\lim_{t\to\infty} m(t) \quad = \quad \operatorname{esssup} U \quad = \quad \inf\{a \in \mathbb{R} \,:\, \nu[U > a] = 0\}, \quad \textit{und} \tag{6.1.1}$$

$$\lim_{t\to-\infty} m(t) \quad = \quad \operatorname{essinf} U \quad = \quad \sup\{a \in \mathbb{R} \,:\, \nu[U < a] = 0\}, \tag{6.1.2}$$

d.h. $m : \mathbb{R} \to (\operatorname{essinf} U, \operatorname{esssup} U)$ ist bijektiv.

Beweis. (1). Sei $t \in \overset{\circ}{\Theta}$. Wir betrachten die momentenerzeugende Funktion

$$M(s) \quad = \quad \int e^{sU} \, d\mu_t$$

der Verteilung μ_t. Wegen $t \in \overset{\circ}{\Theta}$ gilt

$$M(s) \quad = \quad \int \frac{1}{Z(t)} e^{(s+t)U} \, d\nu \quad = \quad Z(s+t)/Z(t) \quad < \quad \infty \tag{6.1.3}$$

für alle s in einer Umgebung $(-\varepsilon, \varepsilon)$ der 0, also $M \in C^\infty(-\varepsilon, \varepsilon)$. Wegen (6.1.3) folgt $Z \in C^\infty(t-\varepsilon, t+\varepsilon)$,

$$\begin{aligned} \int U \, d\mu_t \quad &= \quad M'(0) \quad = \quad \frac{Z'(t)}{Z(t)} \quad = \quad \Lambda'(t), \qquad \text{und} \\ \mathrm{Var}_{\mu_t}[U] \quad &= \quad (\log M)''(0) \quad = \quad \Lambda''(t). \end{aligned}$$

(2). Aus (1) folgt $m = \Lambda' \in C^\infty(\overset{\circ}{\Theta})$ und $m' = \Lambda'' = v$. Also ist m monoton wachsend, und strikt monoton wachsend, falls $\mathrm{Var}_{\mu_t}[U] > 0$. Diese Bedingung ist immer erfüllt, wenn U nicht ν-fast überall konstant ist.

Für $a \in (\operatorname{essinf} U, \operatorname{esssup} U)$ folgt mit monotoner Konvergenz:

$$\mu_t[U < a] \quad = \quad \frac{\int e^{tU} \cdot I_{\{U<a\}} \, d\nu}{\int e^{tU} \, d\nu} \quad \le \quad \frac{\int e^{t(U-a)} \cdot I_{\{U<a\}} \, d\nu}{\int e^{t(U-a)} \cdot I_{\{U\ge a\}} \, d\nu} \quad \to \quad 0$$

für $t \to \infty$, also $\lim_{t\to\infty} \mu_t[U > a] = 1$. Hieraus folgt

$$\liminf_{t\to\infty} m(t) \quad \ge \quad a \cdot \liminf_{t\to\infty} \mu_t[U > a] \quad = \quad a \qquad \text{für alle } a < \operatorname{esssup} U,$$

also (6.1.1). Die Aussage (6.1.2) zeigt man analog.

□

Eine weitere fundamentale Eigenschaft exponentieller Familien werden wir in Satz 6.6 beweisen: Die Verteilungen aus der exponentiellen Familie *minimieren die relative Entropie bzgl. ν unter Nebenbedingungen an den Erwartungswert von U*.

Beispiel (Ising-Modell). Das Ising-Modell wurde 1925 in der Dissertation von Ernst Ising mit der Absicht eingeführt, Phasenübergänge von ferromagnetischen Materialien in einem vereinfachten mathematischen Modell nachzuweisen. Heute spielt das Modell eine wichtige Rolle als einfach zu formulierendes, aber schwer zu analysierendes grundlegendes mathematisches Modell, das auch in unterschiedlichen Anwendungsbereichen wie z.B. der Bildverarbeitung eingesetzt wird.

Sei $S = \{-1, 1\}^V$, wobei V die Knotenmenge eines endlichen Graphen (V, E) ist, z.B.

$$V \quad = \quad \{-k, -k+1, \ldots, k-1, k\}^d \subseteq \mathbb{Z}^d, \quad d, k \in \mathbb{N}.$$

Ein Element $\sigma = (\sigma_i : i \in V)$ aus S interpretieren wir physikalisch als Konfiguration von Spins $\sigma_i \in \{-1, 1\}$ an den Knoten $i \in V$, wobei $\sigma_i = +1$ für einen Spin in Aufwärtsrichtung und $\sigma_i = -1$ für einen Spin in Abwärtsrichtung steht. Die Energie einer Konfiguration σ durch

$$H(\sigma) \quad = \quad \sum_{(i,j)\in E} |\sigma_i - \sigma_j|^2 + h \sum_{i \in V} \sigma_i$$

gegeben, wobei die erste Summe über alle Kanten des Graphen läuft, und der zweite Term die Wechselwirkung mit einem äußeren Magnetfeld mit Stärke $h \in \mathbb{R}$ beschreibt. Der erste Term bewirkt, dass sich benachbarte Spins vorzugsweise gleich ausrichten. Als Gleichgewichtsverteilung bei inverser Temperatur $\beta = 1/T$ ergibt sich die Verteilung $\mu_{\beta,h}$ auf S mit Gewichten

$$\mu_{\beta,h}(\sigma) \quad \propto \quad \exp\left(-\beta \sum_{(i,j)\in E} |\sigma_i - \sigma_j|^2 - \beta h \sum_{i \in V} \sigma_i\right).$$

Die folgende Grafik zeigt Stichproben von der Verteilung $\mu_{\beta,h}$ auf einem ?×? Gitter V für verschiedene Werte von β und h.

GRAFIK EINFÜGEN

Für $\beta = 0$ (d.h. bei unendlicher Temperatur) ergibt sich eine Gleichverteilung. Für $\beta \to \infty$ (Temperatur $\to 0$) konzentriert sich die Verteilung dagegen auf den energieminimierenden Konfigurationen. Dieses sind für $h = 0$ die beiden konstanten Konfigurationen $\sigma_i \equiv +1$ und $\sigma_i \equiv -1$, für $h \neq 0$ hat dagegen nur eine dieser Konfigurationen minimale Energie.

6.1.2 Große Abweichungen vom Gesetz der großen Zahlen

Sei ν eine Wahrscheinlichkeitsverteilung auf einem messbaren Raum $(S, \mathcal{S})$, $U : S \to \mathbb{R}$ eine messbare Funktion, und sei $(X_i)_{i\in\mathbb{N}}$ eine Folge unabhängiger Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit Verteilung ν. Wir setzen voraus:

Annahmen:

(1). Alle exponentiellen Momente der Zufallsvariablen $U(X_i)$ existieren, d.h.

$$\Lambda(t) \quad = \quad \log \int e^{tU} d\nu \quad < \quad \infty \qquad \text{für alle } t \in \mathbb{R}.$$

(2). U ist nicht ν-fast sicher konstant.

Sei $a \in \mathbb{R}$ fest. Wir möchten nun die Asymptotik der Wahrscheinlichkeiten

$$\theta_n \quad = \quad P[S_n \geq an], \qquad S_n = \sum_{i=1}^{n} U(X_i),$$

für $n \to \infty$ genauer untersuchen. Nach dem Gesetz der großen Zahlen gilt:

$$\frac{S_n}{n} \quad \longrightarrow \quad m \; = \; \int U \, d\nu \qquad P\text{-fast sicher.}$$

Für $a > m$ ist das Ereignis $\{S_n \geq an\}$ also eine große Abweichung vom typischen Verhalten. Der Satz von Chernoff liefert eine obere Schranke der Wahrscheinlichkeiten θ_n. Um eine asymptotische untere Schranke zu erhalten, führen wir eine Maßtransformation durch. Es gilt

$$\theta_n \; = \; \nu^n[A_n] \qquad \text{mit} \qquad A_n \; = \; \left\{ x \in S^n \; : \; \sum_{i=1}^{n} U(x_i) \geq an \right\}. \tag{6.1.4}$$

Wir wollen zu einer Verteilung übergehen, bzgl. der das Ereignis A_n nicht mehr selten, sondern typisch ist. Dazu betrachten wir die Produktmaße $\mu_t^n, t \in \mathbb{R}$, wobei μ_t die Verteilung aus der exponentiellen Familie mit relativer Dichte

$$\frac{d\mu_t}{d\nu}(x) \quad = \quad \exp\left(tU(x) - \Lambda(t)\right)$$

ist. Die relative Dichte von μ_t^n bzgl. ν^n ist dann

$$w_t^n(x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^{n} \frac{d\mu_t}{d\nu}(x_i) \quad = \quad \exp\left(t \sum_{i=1}^{n} U(x_i) - n\Lambda(t) \right). \tag{6.1.5}$$

Man beachte, dass $(\mu_t^n)_{t\in\mathbb{R}}$ wieder eine exponentielle Familie ist. Es gilt

$$w_t^n(X_1,\ldots,X_n) \quad = \quad \exp(tS_n - n\Lambda(t)).$$

Wir wollen uns nun überlegen, wie wir den Parameter t in angemessener Weise wählen können. Wenn wir t zu klein wählen, dann hat das Ereignis A_n für große n nur eine geringe Wahrscheinlichkeit bzgl. μ_t^n. Wählen wir umgekehrt t sehr groß, dann liegt die Wahrscheinlichkeit $\mu_t^n[A_n]$ für große n nahe bei 1. In beiden Fällen sind Abschätzungen für $\mu_t^n[A_n]$ daher nur bedingt aussagekräftig. Um eine präzisere Aussage zu erhalten, sollten wir t so groß wählen, dass das Ereignis A_n „gerade typisch wird“. Der Erwartungswert

$$m(t) \quad = \quad \int U\, d\mu_t, \qquad t\in\mathbb{R},$$

ist nach Lemma 6.1 strikt monoton wachsend und stetig in t. Wählen wir t^* mit

$$m(t^*) \quad = \quad a,$$

dann gilt nach dem Gesetz der großen Zahlen

$$\lim_{n\to\infty} \mu_{t^*}^n \left[\left\{x\in S^n \,:\, \frac{1}{n}\sum_{i=1}^n U(x_i)\in(a-\varepsilon,a+\varepsilon)\right\}\right] \quad = \quad 1 \qquad \text{für alle } \varepsilon>0,$$

und nach dem zentralen Grenzwertsatz

$$\lim_{n\to\infty} \mu_{t^*}^n \left[\left\{x\in \mathbb{R}^n \,:\, \frac{1}{n}\sum_{i=1}^n U(x_i)\geq a\right\}\right] \quad = \quad \frac{1}{2},$$

d.h. t^* ist gerade der gesuchte „Schwellenwert“.

Die Umsetzung unserer Überlegungen führt zu einer ersten Aussage über die Asymptotik der Wahrscheinlichkeiten großer Abweichungen vom Gesetz der großen Zahlen auf der exponentiellen Skala:

Satz 6.2 (Cramér). *Unter den Annahmen von oben gilt*

$$\lim_{n\to\infty}\frac{1}{n}\log P\left[\frac{S_n}{n}\geq a\right] \quad = \quad -I(a) \qquad \textit{für alle } a\in(m,\operatorname{esssup} U),$$

wobei die Ratenfunktion

$$I(a) \quad = \quad \sup_{t\in\mathbb{R}}\,(ta-\Lambda(t))$$

die Legendre-Transformation von Λ ist.

In der am Anfang dieses Kapitels eingeführten Notation besagt der Satz von Cramér, dass

$$\theta_n \;=\; P[S_n/n \geq a] \;\simeq\; \exp(-n \cdot I(a))$$

gilt, wobei subexponentiell wachsende Faktoren vernachlässigt werden. Er besagt *nicht*, dass die Folgen (θ_n) und $(\exp(-n \cdot I(a)))$ asymptotisch äquivalent sind!

Zum Beweis einer solchen Aussage zeigt man in der Regel separat eine obere Schranke der Form $\theta_n \preceq \exp(-n \cdot I(a))$ und die untere Schranke $\theta_n \succeq \exp(-n \cdot I(a))$:

Beweis. Der Beweis setzt sich zusammen aus einer nicht-asymptotischen Abschätzung der Wahrscheinlichkeiten

$$\theta_n \quad = \quad P[S_n \geq an] \quad = \quad \nu^n[A_n], \qquad A_n \;=\; \left\{ x \in S^n \;:\; \sum_{i=1}^n U(x_i) \geq an \right\},$$

nach oben, und einer asymptotischen Abschätzung der Wahrscheinlichkeiten nach unten.

(1). *Obere Schranke.* Die nicht-asymptotische obere Schranke

$$\frac{1}{n} \log \theta_n \leq -I(a) \qquad \text{für alle } n \in \mathbb{N}$$

liefert der Satz von Chernoff (Satz 4.13). Zur Illustration schreiben wir das Hauptargument aus dem Beweis von oben noch einmal so auf, dass der Zusammenhang mit einer Maßtransformation verdeutlicht wird: Für $t > 0$ gilt nach (6.1.5):

$$\begin{aligned} \theta_n \quad = \quad \nu^n[A_n] \;&=\; \int_{A_n} \frac{1}{w_t^n}\, d\mu_t^n \\ &=\; \int_{A_n} \exp\left(-t \sum_{i=1}^n U(x_i) + \Lambda(t) n \right)\, d\mu_t^n \\ &\leq\; e^{-(ta - \Lambda(t))n} \cdot \mu_t^n[A_n] \\ &\leq\; e^{-(ta-\Lambda(t))n}. \end{aligned}$$

Hieraus folgt die Behauptung wie im Beweis von Satz 4.13 durch Optimieren der Abschätzung in t.

(2). *Untere Schranke.* Wir zeigen nun die asymptotische untere Schranke

$$\liminf_{n\to\infty} \frac{1}{n} \log \nu^n[A_n] \quad \geq \quad -I(a). \tag{6.1.6}$$

Zusammen mit der oberen Schranke folgt dann

$$\lim_{n\to\infty} \frac{1}{n} \log \nu^n[A_n] \quad = \quad -I(a),$$

d.h. die obere Schranke ist asymptotisch „scharf“. Zum Beweis von (6.1.6) gehen wir zu der Verteilung $\mu_{t^*}^n$ zum Schwellenwert $t^* = m^{-1}(a)$ über. Nach Lemma 6.1 ist $m : \mathbb{R} \to (\operatorname{essinf} U, \operatorname{esssup} U)$ bijektiv, also existiert $m^{-1}(a) > 0$ für $a \in (m, \operatorname{esssup} U)$. Für $\varepsilon > 0$ sei

$$A_{n,\varepsilon} \quad = \quad \left\{ x \in S^n \,:\, a \le \frac{1}{n} \sum_{i=1}^n U(x_i) \le a + \varepsilon \right\}.$$

Ähnlich wie bei der oberen Schranke erhalten wir

$$\begin{aligned} \nu^n[A_n] \quad \ge \quad \nu^n[A_{n,\varepsilon}] \quad &= \quad \int_{A_{n,\varepsilon}} \exp\left(-t^* \sum_{i=1}^n U(x_i) + \Lambda(t^*) n \right) \, d\mu_{t^*}^n \\ &\ge \quad e^{-(t^*(a+\varepsilon) - \Lambda(t^*))n} \, \mu_{t^*}^n[A_{n,\varepsilon}] \\ &\ge \quad e^{-I(a)n} e^{-t^*\varepsilon n} \, \mu_{t^*}^n[A_{n,\varepsilon}] \end{aligned} \tag{6.1.7}$$

Wegen $\int U \, d\mu_{t^*} = m(t^*) \,=\, a$ gilt nach dem zentralen Grenzwertsatz:

$$\begin{aligned} \mu_{t^*}^n[A_{n,\varepsilon}] \quad &= \quad \mu_{t^*}^n \left[0 \le \frac{1}{\sqrt{n}} \sum_{i=1}^n (U(x_i) - a) \le \varepsilon \sqrt{n} \right] \\ &\overset{n\to\infty}{\longrightarrow} \quad N(0, \operatorname{Var}_{\mu_{t^*}}[U]) \, [\, [0, \infty) \,] \quad = \quad \frac{1}{2}, \end{aligned} \tag{6.1.8}$$

d.h. die große Abweichung ist typisch unter $\mu_{t^*}^n$. Für die Wahrscheinlichkeiten bzgl. ν^n ergibt sich dann nach (6.1.7):

$$\liminf \frac{1}{n} \log \nu^n[A_n] \ge -I(a) - t^*\varepsilon.$$

Die Behauptung folgt für $\varepsilon \searrow 0$.

□

Bemerkung. Ähnliche Aussagen über die Asymptotik von Wahrscheinlichkeiten großer Abweichungen wurden auch in vielen Modellen mit Abhängigkeit bewiesen. Sie spielen unter anderem in der mathematischen statistischen Mechanik eine wichtige Rolle.

6.2 Entropie und relative Entropie

Wir definieren nun die Entropie einer diskreten Wahrscheinlichkeitsverteilung und die relative Entropie zweier beliebiger Wahrscheinlichkeitsmaße. Mithilfe des Gesetzes der großen Zahlen können wir statistische Interpretationen dieser Größen geben. Insbesondere mißt die relative

Entropie die Unterscheidbarkeit zweier Wahrscheinlichkeitsmaße durch Folgen von unabhängigen Stichproben. In Abschnitt 6.3 werden wir zeigen, dass daraus eine allgemeine untere Schranke für die exponentielle Abfallrate der Wahrscheinlichkeiten seltener Ereignisse in Produktmodellen folgt.

6.2.1 Entropie und Information

Wir bemerken zunächst, dass die auf $[0, \infty)$ definierte Funktion

$$u(x) := \begin{cases} x \log x & \text{für } x > 0 \\ 0 & \text{für } x = 0 \end{cases}$$

stetig und strikt konvex ist mit $u'(x) = 1 + \log x$ und $u''(x) = 1/x$ für $x > 0$. Insbesondere gilt

$$u(x) \leq 0 \qquad \text{für alle } x \in [0, 1], \tag{6.2.1}$$

$$u(x) \geq x - 1 \qquad \text{für alle } x \geq 0, \tag{6.2.2}$$

und $u(1/e) = -1/e$ ist das absolute Minimum der Funktion u.

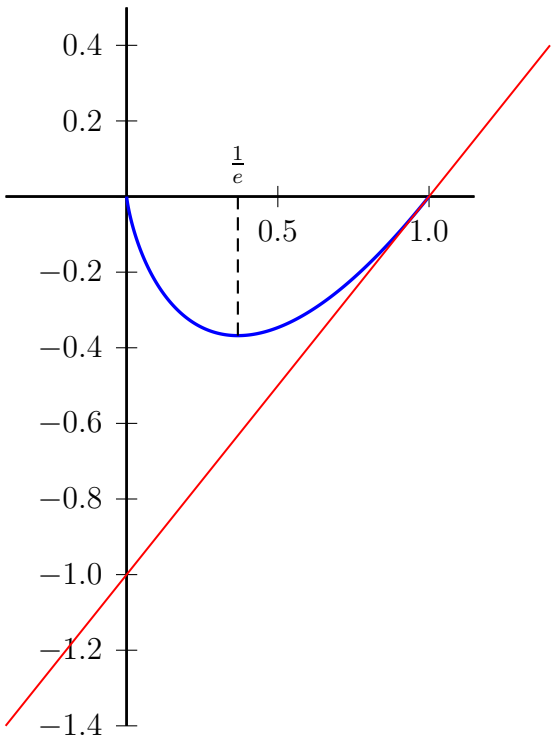


Abbildung 6.1: Graph der Funktion $u(x)$ (blau) und ihrer unteren Schranke $x - 1$ (rot)

Sei nun S eine abzählbare Menge, und $\mu = (\mu(x))_{x \in S}$ eine Wahrscheinlichkeitsverteilung auf S.

Definition (Entropie einer diskreten Wahrscheinlichkeitsverteilung). *Die Größe*

$$H(\mu) \quad := \quad -\sum_{\substack{x\in S\\ \mu(x)\neq 0}} \mu(x)\log\mu(x) \quad = \quad -\sum_{x\in S} u(\mu(x)) \quad \in [0,\infty]$$

*heißt **Entropie** der Wahrscheinlichkeitsverteilung μ.*

Anschaulich können wir $-\log\mu(x)$ interpretieren als Maß für die »Überraschung« bzw. den »Informationsgewinn«, falls eine Stichprobe von der Verteilung μ den Wert x hat. Die »Überraschung« ist umso größer, je unwahrscheinlicher x ist. Die Entropie $H(\mu)$ ist dann die »mittlere Überraschung« bzw. der »mittlere Informationsgewinn« beim Ziehen einer Stichprobe von ν. Eine wichtige Eigenschaft der Entropie, die auch die Wahl des Logarithmus erklärt, ist:

Satz 6.3 (Faktorisierungseigenschaft). *Für beliebige diskrete Wahrscheinlichkeitsverteilungen ν und μ gilt:*

$$H(\nu\otimes\mu) \;=\; H(\nu)+H(\mu).$$

Der mittlere Informationszuwachs in einem aus zwei unabhängigen Experimenten zusammengesetzten Zufallsexperiment ist also die Summe der einzelnen mittleren Informationszuwächse.

Beweis. Nach Definition der Entropie gilt:

$$\begin{aligned} H(\nu\otimes\mu) &= \sum_{\substack{x,y\\ \nu(x)\mu(y)\neq 0}} \nu(x)\mu(y)\log(\nu(x)\mu(y)) \\ &= -\sum_{x:\nu(x)\neq 0} \nu(x)\log(\nu(x)) - \sum_{y:\mu(y)\neq 0} \mu(y)\log(\mu(y)) \\ &= H(\nu)+H(\mu). \end{aligned}$$

□

Wir bestimmen nun auf einer gegebenen abzählbaren Menge S die Wahrscheinlichkeitsverteilungen mit minimaler bzw. maximaler Entropie.

Entropieminima

Nach (6.2.1) ist die Entropie stets nicht-negativ. Zudem gilt:

$$H(\mu)=0 \quad\Longleftrightarrow\quad \mu(x)\in\{0,1\} \quad \forall\, x\in S \quad\Longleftrightarrow\quad \mu \text{ ist ein Dirac-Maß.}$$

Die Dirac-Maße sind also die Entropieminima. Ist das Zufallsexperiment deterministisch, d.h. μ ein Diracmaß, dann tritt bei Ziehen einer Stichprobe von μ keine Überraschung bzw. kein Informationszuwachs auf.

Entropiemaximum

Ist S endlich, dann gilt für alle Wahrscheinlichkeitsverteilungen μ auf S:

$$H(\mu) \leq \log(|S|) = H(\mathcal{U}_S),$$

wobei $\mathcal{U}_S$ die Gleichverteilung auf S ist. Nach der Jensenschen Ungleichung gilt nämlich

$$\begin{aligned} -\sum_{x\in S} u(\mu(x)) &= -|S| \cdot \int u(\mu(x))\, \mathcal{U}_S(dx) \\ &\leq -|S| \cdot u\left(\int \mu(x)\, \mathcal{U}_S(dx)\right) \\ &= -|S| \cdot u\left(1/|S|\right) \quad = \quad \log|S| \end{aligned}$$

mit Gleichheit genau dann, wenn μ die Gleichverteilung ist.

Die Gleichverteilung maximiert also die Entropie auf einem endlichen Zustandsraum. Anschaulich können wir die Gleichverteilung als eine »völlig zufällige« Verteilung auffassen – d.h. wir verwenden die Gleichverteilung als Modell, wenn wir keinen Grund haben, einen der Zustände zu bevorzugen. Die Entropie ist in diesem Sinne ein Maß für die *»Zufälligkeit«* (bzw. *»Unordnung«*) der Wahrscheinlichkeitsverteilung μ.

Ist S abzählbar unendlich, dann gibt es Wahrscheinlichkeitsverteilungen auf S mit unendlicher Entropie.

Als nächstes geben wir eine statistische Interpretation der Entropie. Sei μ eine Wahrscheinlichkeitsverteilung auf einer abzählbaren Menge S. Die Wahrscheinlichkeit einer Folge von Ausgängen $x_1, \ldots, x_n$ bei Entnehmen einer Stichprobe aus n unabhängigen Zufallsgrößen mit Verteilung μ beträgt

$$p_n(x_1, ..., x_n) = \prod_{i=1}^{n} \mu(x_i).$$

Der gemittelte Informationszuwachs durch Auswertung der Werte $x_1, \ldots, x_n$ ist also

$$-\frac{1}{n} \log p_n(x_1, ..., x_n).$$

Mithilfe des Gesetzes der großen Zahlen können wir die Asymptotik dieser Größen für $n \to \infty$ untersuchen:

Satz 6.4 (Entropie als asymptotische Informationszuwachsrate). *Seien $X_1, X_2, \ldots : \Omega \to S$ unter P unabhängige Zufallsvariablen mit Verteilung μ. Dann gilt P-fast sicher :*

$$-\frac{1}{n} \log p_n(X_1, \ldots, X_n) \quad \longrightarrow \quad H(\mu) \qquad \textit{für } n \to \infty.$$

Beweis. Mit Wahrscheinlichkeit 1 gilt $0 < \mu(X_i) \leq 1$, also $-\log \mu(X_i) \in [0, \infty)$ für alle i. Nach Korollar 4.11 folgt fast sicher

$$-\frac{1}{n} \log p_n(X_1, \ldots, X_n) \quad = \quad -\frac{1}{n} \sum_{i=1}^{n} \log \mu(X_i) \quad \stackrel{n\to\infty}{\longrightarrow} \quad -\int \log \mu \, d\mu \quad = \quad H(\mu).$$

□

In der zu Beginn dieses Kapitels eingeführten Notation zur Äquivalenz auf der exponentiellen Skala besagt die Aussage des Satzes, dass fast sicher

$$p_n(X_1, \ldots, X_n) \quad \simeq \quad e^{-nH(\mu)} \qquad \text{gilt.}$$

6.2.2 Relative Entropie und statistische Unterscheidbarkeit

Die Entropie ist nur für diskrete Wahrscheinlichkeitsmaße definiert. Eine Übertragung der Definition auf absolutstetige Wahrscheinlichkeitsmaße auf $\mathbb{R}^d$ ist möglich, indem man

$$H(\mu) \; = \; -\int_{\mathbb{R}^d} f \, \log f \, dx \qquad \text{mit} \qquad f = d\mu/dx$$

setzt. Allerdings kann die so definierte Entropie sowohl positive als auch negative Werte annehmen. Wir führen jetzt den allgemeineren Begriff der *relativen Entropie* zweier Wahrscheinlichkeitsmaße auf einem beliebigen meßbaren Raum $(S, \mathcal{S})$ ein:

Definition (Relative Entropie zweier Wahrscheinlichkeitsmaße). *Seien μ und ν Wahrscheinlichkeitsmaße auf $(S, \mathcal{S})$. Die durch*

$$H(\mu \mid \nu) \quad := \quad \begin{cases} \int \log w \, d\mu \; = \; \int w \log w \, d\nu & \textit{falls } \mu \ll \nu \textit{ mit Dichte } w, \\ \infty \qquad \textit{sonst.} \end{cases} \tag{6.2.3}$$

*definierte Größe $H(\mu \mid \nu) \in [0, \infty]$ heißt **relative Entropie** (oder **Kullback-Leibler Information**) von μ bzgl. ν.*

Um eine anschauliche Interpretation der relativen Entropie zu geben, nehmen wir an, dass μ und ν Wahrscheinlichkeitsmaße auf $S = \mathbb{R}^d$ oder einem diskreten Raum mit Dichten (bzw. Massenfunktionen) $f, g > 0$ sind. Die relative Dichte w von μ bzgl. ν ist dann

$$w(x) \quad = \quad \frac{d\mu}{d\nu}(x) \quad = \quad \frac{f(x)}{g(x)} \qquad \text{für } \nu\text{-fast alle } x \in S,$$

und

$$H(\mu \mid \nu) \quad = \quad \int \log \frac{f}{g} \, d\mu \quad = \quad \int (-\log g(x) - (-\log f(x))) \, \mu(dx).$$

Wir können $-\log g(x)$ und $-\log f(x)$ als Maß für die Überraschung (den Informationsgewinn) bei Eintreten von x interpretieren, falls ν bzw. μ das zugrundeliegende Modell ist. Wenn wir also ν als Modell annehmen, aber tatsächlich μ die zugrundeliegende Verteilung ist, dann erhöht sich die Überraschung (der Informationszuwachs) bei Ziehen einer Stichprobe im Vergleich zum korrekten Modell im Mittel um $H(\mu \mid \nu)$.

Bemerkung (Entropie als Spezialfall der relativen Entropie). Ist ν das Zählmaß auf einer abzählbaren Menge S, dann gilt

$$H(\mu \mid \nu) \quad = \quad \sum_{\mu(x)\neq 0} \mu(x) \log \mu(x) \quad = \quad -H(\mu). \tag{6.2.4}$$

Ist S endlich, und ν die Gleichverteilung (also das normierte Zählmaß) auf S, dann folgt entsprechend

$$H(\mu \mid \nu) \quad = \quad \sum_{\mu(x)\neq 0} \mu(x) \log(\mu(x) \cdot |S|) \quad = \quad \log |S| - H(\mu).$$

Aussagen über die relative Entropie liefern also als Spezialfall entsprechende Aussagen für die Entropie (wobei sich aber das Vorzeichen umkehrt!)

Das folgende Lemma fasst elementare Eigenschaften der relativen Entropie zusammen:

Lemma 6.5 (Eigenschaften der relativen Entropie).

(1). Es gilt $H(\mu \mid \nu) \geq 0$ mit Gleichheit genau dann, wenn $\mu = \nu$.

(2). $H(\mu_1 \otimes \ldots \otimes \mu_n \mid \nu_1 \otimes \ldots \otimes \nu_n) \quad = \quad \sum_{i=1}^{n} H(\mu_i \mid \nu_i).$

Beweis. Sei $\mu \ll \nu$ mit relativer Dichte w. Wegen $x \log x \geq x - 1$ folgt

$$H(\mu \mid \nu) \quad = \quad \int w \log w \, d\nu \quad \geq \quad \int (w-1) \, d\nu \quad = \quad \int w \, d\nu - 1 \quad = \quad 0.$$

Gleichheit gilt genau dann, wenn w ν-fast sicher gleich 1, also $\mu = \nu$ ist.

Der Beweis der zweiten Aussage ist eine Übungsaufgabe. □

Beispiele (Relative Entropie von Bernoulli- und Binomialverteilungen).

(1). Für die Bernoulli-Verteilungen mit $\nu_p(1) = p$ und $\nu_p(0) = 1 - p$ gilt:

$$H(\nu_a \mid \nu_p) \quad = \quad a \log\left(\frac{a}{p}\right) + (1-a) \log\left(\frac{1-a}{1-p}\right) \qquad \text{für alle } a, p \in (0,1).$$

(2). Für Normalverteilungen mit Mittelwerten $m, \widetilde{m} \in \mathbb{R}$ und Varianzen $v, \widetilde{v} > 0$ gilt

$$\begin{aligned} H(N(\widetilde{m},\widetilde{v}) \mid N(m,v)) &= \frac{1}{2}\left(\log\left(\frac{v}{\widetilde{v}}\right) + \frac{\widetilde{v}}{v} - 1 + \frac{(\widetilde{m}-m)^2}{v}\right), \qquad \text{also insbesondere} \\ H(N(\widetilde{m},v) \mid N(m,v)) &= \frac{(\widetilde{m}-m)^2}{2v}. \end{aligned}$$

Die Beipiele zeigen, dass die relative Entropie im Allgemeinen *nicht symmetrisch* ist.

Minima der relativen Entropie unter Nebenbedingungen.

Nach Lemma 6.5 ist $H(\mu|\nu)$ minimal, wenn $\mu = \nu$ gilt. Als Minimierer der relativen Entropie unter Nebenbedingungen an μ ergeben sich Verteilungen aus exponentiellen Familien:

Sei ν ein festes Wahrscheinlichkeitsmaß auf $(S, \mathcal{S})$ und $U : S \to \mathbb{R}$ eine meßbare Funktion mit endlicher momentenerzeugender Funktion $Z(t) = \int \exp(tU)\, d\nu,\ t \in \mathbb{R}$.

Satz 6.6 (Exponentielle Familien als Minimierer der relativen Entropie). *Für jedes feste $t \geq 0$ minimiert das Maß*

$$\mu_t(dx) \quad = \quad \frac{1}{Z(t)} \exp\left(tU(x)\right) \nu(dx)$$

die relative Entropie bzgl. ν unter allen Wahrscheinlichkeitsmaßen μ mit $\int U\, d\mu \geq m(t)$, wobei $m(t) = \int U\, d\mu_t$ der Erwartungswert von U bzgl. μ_t ist. Genauer gilt:

$$H(\mu_t \mid \nu) \;=\; t \cdot m(t) - \log Z(t) \;\leq\; H(\mu \mid \nu) \tag{6.2.5}$$

für jedes Wahrscheinlichkeitsmaß μ auf $(S, \mathcal{S})$ mit $\int U\, d\mu \geq m(t)$.

Beweis. Sei μ ein Wahrscheinlichkeitsmaß mit $H(\mu \mid \nu) < \infty$ und $\int U\, d\mu \geq m(t)$. Dann gilt $\mu \ll \nu$ und

$$\begin{aligned} H(\mu \mid \nu) &= \int \log \frac{d\mu}{d\nu}\, d\mu \quad = \quad \int \log \frac{d\mu}{d\mu_t}\, d\mu + \int \log \frac{d\mu_t}{d\nu}\, d\mu \\ &= H(\mu \mid \mu_t) + \left(t \int U\, d\mu - \log Z(t)\right) \\ &\geq tm(t) - \log Z(t). \end{aligned}$$

Für $\mu = \mu_t$ ergibt sich Gleichheit. □

Bemerkung (Variationsproblem). Sei $w = d\mu/d\nu$ die relative Dichte von μ bzgl. ν. Das im Satz betrachtete Variationsproblem hat dann die Form

$$H(\mu \mid \nu) \quad = \quad \int w \log w\, d\nu \;\stackrel{!}{=}\; \min$$

unter der Nebenbedingung

$$\int U \, d\mu \quad = \quad \int Uw \, d\nu \quad \geq \quad m(t),$$

und kann auch formal durch klassische Variationsrechnung gelöst werden.

Wir geben nun eine statistische Interpretation der relativen Entropie. Dazu nehmen wir wieder an, dass μ und ν Wahrscheinlichkeitsverteilungen auf $S = \mathbb{R}^d$ oder einem diskreten Raum mit Dichten (bzw. Massenfunktionen) $f, g > 0$, und relativer Dichte $w = f/g$ sind. Die Dichte bzw. Massenfunktion

$$L_n(\nu; x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^{n} g(x_i)$$

der Verteilung n unabhängiger Stichproben $X_1, \ldots, X_n$ von ν bezeichnet man in der Statistik auch als **Likelihood** der Verteilung ν bzgl. der Daten $(x_1, \ldots, x_n)$.

Wie kann man anhand von unabhängigen Stichproben erkennen, welche der beiden Verteilungen ν und μ in einem Zufallsexperiment vorliegt? Dazu betrachten wir den **Likelihood-Quotienten**

$$w_n(x_1, \ldots, x_n) \quad := \quad \frac{L_n(\mu; x_1, \ldots, x_n)}{L_n(\nu; x_1, \ldots, x_n)} \quad = \quad \frac{\prod_{i=1}^n f(x_i)}{\prod_{i=1}^n g(x_i)} \quad = \quad \prod_{i=1}^{n} w(x_i).$$

Analog zu Satz 6.4 erhalten wir die folgende (allgemeinere) Aussage:

Satz 6.7 (Relative Entropie als statistische Unterscheidbarkeit; Shannon-McMillan). *Seien $X_1, X_2, \ldots : \Omega \to S$ unabhängige Zufallsvariablen unter P_ν bzw. P_μ mit Verteilung ν bzw. μ. Dann gilt für $n \to \infty$:*

$$\frac{1}{n} \log w_n(X_1, \ldots, X_n) \quad \longrightarrow \quad H(\mu \mid \nu) \qquad P_\mu\text{-fast sicher, und} \tag{6.2.6}$$

$$\frac{1}{n} \log w_n(X_1, \ldots, X_n) \quad \longrightarrow \quad -H(\nu \mid \mu) \qquad P_\nu\text{-fast sicher.} \tag{6.2.7}$$

Beweis. (1). Für $n \to \infty$ gilt nach dem Gesetz der großen Zahlen

$$\frac{1}{n} \log w_n(X_1, \ldots, X_n) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \log w(X_i) \quad \longrightarrow \quad \int \log w \, d\mu \qquad P_\mu\text{-fast sicher.}$$

Das Gesetz der großen Zahlen ist anwendbar, da

$$\int (\log w)^- \, d\mu \quad = \quad \int (w \log w)^- \, d\nu \quad \leq \quad \frac{1}{e} \quad < \quad \infty.$$

(2). Da ν absolutstetig bzgl. μ mit Dichte $1/w$ ist, gilt entsprechend

$$\begin{aligned} \frac{1}{n}\log w_n(X_1,\ldots,X_n) &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{1}{w(X_i)} \\ &\xrightarrow{\text{GGZ}} -\int \log\frac{1}{w}\,d\nu = -H(\nu\mid\mu) \qquad P_\nu\text{-fast sicher.} \end{aligned}$$

□

Der Satz von Shannon-Mc Millan zeigt, dass sich die Produktdichte (der Likelihood-Quotient) asymptotisch auf der exponentiellen Skala (d.h. unter Vernachlässigung subexponentiell wachsender Faktoren) folgendermaßen verhält:

$$w_n(X_1,\ldots,X_n) \simeq \begin{cases} e^{nH(\mu\mid\nu)} & P_\mu\text{-fast sicher,} \\ e^{-nH(\nu\mid\mu)} & P_\nu\text{-fast sicher.} \end{cases}$$

Damit erhalten wir eine statistische Interpretation der relativen Entropie als natürlichen *(nicht-symmetrischen!)* Abstandsbegriff für Wahrscheinlichkeitsmaße. Wir werden diese statistische Interpretation in Satz 6.9 noch weiter präzisieren.

6.2.3 Entropie von Markovketten

Sei $p(x,y)$ $(x,y \in S)$ eine stochastische Matrix auf einer endlichen Menge S mit Gleichgewichtsverteilung $\nu \in WV(S)$, d.h. für alle $y \in S$ gelte

$$\sum_{x\in S}\nu(x)\,p(x,y) = \nu(y). \tag{6.2.8}$$

Der folgende wichtige Satz zeigt, dass die relative Entropie $H(\mu p^n|\nu)$ der Verteilung zur Zeit n einer Markovkette mit Startverteilung μ und Übergangsmatrix p bezüglich des Gleichgewichts ν monoton fällt:

Satz 6.8 (Abfall der relativen Entropie). *Ist p eine stochastische Matrix auf S und ν ein Gleichgewicht von p, dann gilt für jede Wahrscheinlichkeitsverteilung μ auf S:*

$$H(\mu p|\nu) \geq H(\mu|\nu)\,. \tag{6.2.9}$$

Insbesondere ist $n \mapsto H(\mu p^n|\nu)$ stets monoton fallend.

Beweis. Ist μ nicht absolutstetig bezüglich ν, dann ist die Aussage (6.2.9) automatisch erfüllt. Andernfalls sei w eine Version der relativen Dichte $d\mu/d\nu$. Dann gilt

$$(\mu p)(y) \;=\; \sum_{x\in S} \mu(x)\, p(x,y) \;=\; \sum_{x\in S} w(x)\, \nu(x)\, p(x,y). \tag{6.2.10}$$

Aus der Gleichgewichtsbedingung (6.2.8) folgt $\nu(x)p(x,y) \le \nu(y)$ für alle $x, y \in S$. Also ist auch μp absolutstetig bzgl. ν, mit relativer Dichte

$$\frac{(\mu p)(y)}{\nu(y)} \;=\; \sum_{x\in S} w(x)\, \frac{\nu(x)\, p(x,y)}{\nu(y)}. \tag{6.2.11}$$

Aus der Gleichgewichtsbedingung folgt auch, dass $x \mapsto \nu(x)p(x,y)/\nu(y)$ die Massenfunktion einer Wahrscheinlichkeitsverteilung auf S ist. Darum können wir die Jensensche Ungleichung auf die konvexe Funktion $u(x) = x \log_+ x$ anwenden, und erhalten

$$u\left(\sum_{x\in S} w(x)\, \frac{\nu(x)\, p(x,y)}{\nu(y)}\right) \quad\le\quad \sum_{x\in S} u(w(x))\, \frac{\nu(x)\, p(x,y)}{\nu(y)}.$$

Zusammen mit (6.2.11) ergibt sich

$$\begin{aligned} H(\mu p|\nu) \;&=\; \sum_{y:\nu(y)\neq 0} u\left(\sum_{x\in S} w(x)\, \nu(x)\, p(x,y)/\nu(y)\right)\, \nu(y) \\ &\le\; \sum_{y\in S}\sum_{x\in S} u(w(x))\, \nu(x)\, p(x,y) \\ &=\; \sum_{x\in S} u(w(x))\, \nu(x) \quad=\quad H(\mu|\nu). \end{aligned}$$

□

Bemerkung (Zunahme der Entropie; thermodynamische Irreversibilität). Als Spezialfall ergibt sich wegen $H(\mu) = \log|S| - H(\mu|\nu)$ die Aussage, dass die Entropie $H(\mu p^n)$ der Verteilung einer Markovkette zur Zeit n monoton wächst, falls der Zustandsraum endlich und die Gleichverteilung ν ein Gleichgewicht ist. In der Interpretation der statistischen Physik geht die zeitliche Entwicklung auf makroskopischer Ebene (Thermodynamik) von einem geordneten hin zu einem ungeordneten Zustand mit (lokal) maximaler Entropie (»thermodynamische Irreversibilität«). Trotzdem ist auf mikroskopischer Ebene die Dynamik rekurrent, d.h. jeder Zustand $x \in S$ wird von der Markovkette mit Wahrscheinlichkeit 1 unendlich oft besucht – dies dauert nur eventuell astronomisch lange. Die Einführung eines Markov-Modells durch die österreichischen Physiker Tatjana und Paul Ehrenfest konnte eine entsprechende Kontroverse von Zermelo („Dynamik kehrt immer wieder zurück“) und Boltzmann („soll solange warten“) lösen.

Beispiel (**Irrfahrten**). Ist p die Übergangsmatrix eines symmetrischen Random Walks auf dem diskreten Kreis $\mathbb{Z}_k = \mathbb{Z}/(k\mathbb{Z})$, der symmetrischen Gruppe S_n („Mischen eines Kartenspiels“), oder dem diskreten Hyperwürfel $\{0,1\}^n$ („Ehrenfest-Modell“), dann ist die Gleichverteilung ein Gleichgewicht, und die Entropie wächst monoton.

6.3 Untere Schranken durch Maßwechsel

Wir zeigen nun, dass die relative Entropie eine untere Schranken für die exponentielle Abfallrate der Größe typischer Mengen unter einem Wahrscheinlichkeitsmaß bei Übergang zu einem anderen Wahrscheinlichkeitsmaß liefert. Als Konsequenzen hieraus ergeben sich neben der Aussage des Satzes von Cramér unter anderem auch eine untere Schranke für große Abweichungen von empirischen Verteilungen (Satz von Sanov) sowie der Quellenkodierungssatz von Shannon.

6.3.1 Eine allgemeine untere Schranke

Seien $X_1, X_2, \ldots$ unter P_ν bzw. P_μ unabhängige Zufallsvariablen mit Verteilung ν bzw. μ.

Definition (**Wesentliche Mengen in Produktmodellen**). *Eine Folge von Mengen $B_n \subseteq S^n$ ($n \in \mathbb{N}$) heißt **wesentlich** bzgl. μ, falls*

$$P_\mu[(X_1, \ldots, X_n) \in B_n] \quad = \quad \mu^n[B_n] \quad \longrightarrow \quad 1 \qquad \textit{für } n \to \infty.$$

Wie wahrscheinlich muss eine bzgl. μ wesentliche Folge von Mengen unter ν mindestens sein ? Der folgende Satz beantwortet diese Frage auf der exponentiellen Skala:

Satz 6.9 (**Shannon-Mc Millan II**). *(1). Für jedes $\varepsilon > 0$ ist die Folge*

$$B_{n,\varepsilon} := \{(x_1, \ldots, x_n) : e^{n(H(\mu \mid \nu)-\varepsilon)} \leq w_n(x_1, \ldots, x_n) \leq e^{n(H(\mu \mid \nu)+\varepsilon)}\} \subseteq S^n$$

wesentlich bzgl. μ, und

$$\nu^n[B_{n,\varepsilon}] \quad \leq \quad e^{-n(H(\mu \mid \nu)-\varepsilon)} \qquad \textit{für alle } n \in \mathbb{N}. \tag{6.3.1}$$

(2). Für beliebige messbare Mengen $A_n \subseteq S^n$ mit

$$\liminf \mu^n[A_n] > 0 \tag{6.3.2}$$

gilt

$$\liminf \frac{1}{n} \log \nu^n[A_n] \geq -H(\mu \mid \nu). \tag{6.3.3}$$

Beweis. (1). Die Mengen $B_{n,\varepsilon}, n \in \mathbb{N}$, sind wesentlich bzgl. μ nach Satz 6.7. Zudem gilt:

$$1 \geq \mu^n[B_{n,\varepsilon}] = \int_{B_{n,\varepsilon}} w_n \, d\nu^n \geq \nu^n[B_{n,\varepsilon}] \cdot e^{n(H(\mu \mid \nu)-\varepsilon)}.$$

(2). Aus

$$\nu^n[A_n] = \int_{A_n} \frac{1}{w_n} \, d\mu_n \geq e^{-n(H(\mu \mid \nu)+\varepsilon)} \mu^n[A_n \cap B_{n,\varepsilon}]$$

folgt

$$\begin{aligned} \liminf \frac{1}{n} \log \nu^n[A_n] &\geq -(H(\mu \mid \nu) + \varepsilon) + \liminf \frac{1}{n} \log \mu^n[A_n \cap B_{n,\varepsilon}] \\ &= -(H(\mu \mid \nu) + \varepsilon), \end{aligned}$$

da $\liminf \mu^n[A_n \cap B_{n,\varepsilon}] = \liminf \mu^n[A_n] > 0$ nach (1) gilt. Die Behauptung folgt für $\varepsilon \to 0$.

□

Die zweite Aussage der Satzes können wir als eine allgemeine untere Schranke für große Abweichungen interpretieren: Ist $A_n \subseteq S^n$ eine Folge von Ereignissen, deren Wahrscheinlichkeit bzgl. ν^n gegen 0 geht, dann liefert uns (6.3.3) für jede Wahrscheinlichkeitsverteilung μ mit (6.3.2) eine asymptotische untere Schranke für die Wahrscheinlichkeiten

$$P_\nu[(X_1, \ldots, X_n) \in A_n] = \nu^n[A_n]$$

auf der exponentiellen Skala. Wir werden dies im Folgenden verwenden, um verschiedene Sätze über große Abweichungen zu beweisen.

Beispiel (Cramér revisited). Als erste Anwendung betrachten wir nochmal die Situation aus dem Satz von Cramér (Satz 6.2): Sei ν ein Wahrscheinlichkeitsmaß auf einem meßbaren Raum $(S, \mathcal{S})$, $U : S \to \mathbb{R}$ eine meßbare Funktion mit endlicher momentenerzeugender Funktion $Z(t) = \int e^{tU} \, d\nu$, $t \in \mathbb{R}$, und sei

$$a > m = \int U \, d\nu.$$

Um aus (6.3.3) eine bestmögliche asymptotische untere Schranke für die Wahrscheinlichkeiten $\nu^n[A_n]$ der großen Abweichungen

$$A_n = \left\{ (x_1, \ldots, x_n) \in S^n \, : \, \frac{1}{n} \sum_{i=1}^n U(x_i) \geq a \right\}$$

zu erhalten, müssen wir eine Wahrscheinlichkeitsverteilung μ finden, die die relative Entropie $H(\mu \mid \nu)$ unter allen Wahrscheinlichkeitsverteilungen μ mit (6.3.2) minimiert. Die Bedingung

(6.3.2) ist aber genau dann erfüllt, wenn $\int U \, d\mu \geq a$ gilt, denn aus dem Gesetz der großen Zahlen und dem zentralen Grenzwertsatz folgt:

$$\lim_{n\to\infty} \mu^n \left[\frac{1}{n} \sum_{i=1}^{n} U(x_i) \geq a \right] = \begin{cases} 1 & \text{für } a < \int U \, d\mu, \\ 1/2 & \text{für } a = \int U \, d\mu, \\ 0 & \text{für } a > \int U \, d\mu. \end{cases} \tag{6.3.4}$$

Das sich ergebende Variationsproblem

$$H(\mu \mid \nu) \quad = \quad \stackrel{!}{=} \quad \min \qquad \text{unter der Nebenbedingung} \qquad \int U \, d\mu \quad \geq \quad a$$

wird nach Satz 6.6 durch das Wahrscheinlichkeitsmaß μ_{t^*} aus der exponentiellen Familie zu μ und U zum eindeutigen Schwellenwert t^* mit $\int U \, d\mu_{t^*} = a$ gelöst. Wählt man $\mu = \mu_{t^*}$, dann gilt $\lim \mu_{t^*}^n[A_n] = 1/2$ nach (6.3.4). Damit erhalten wir nach Satz 6.9 (2) die asymptotische untere Schranke

$$\liminf \frac{1}{n} \log \nu^n[A_n] \quad \geq \quad -H(\mu_{t^*} \mid \nu) \quad = \quad t^* \cdot a - \log Z(t^*) \quad \geq \quad -I(a),$$

wobei $I(a) = \sup_{t\in\mathbb{R}}(ta - \log Z(t))$ die Ratenfunktion aus Satz dem Chernoff ist. Da nach dem Satz von Chernoff auch $-I(a) \geq \frac{1}{n} \log \nu^n[A_n]$ gilt, folgt

$$\nu^n[A_n] \quad \simeq \quad \exp(-nI(a)), \qquad \text{und} \qquad I(a) = H(\mu_{t^*}|\nu).$$

Das beschriebene Vorgehen liefert nicht nur die untere Schranke im Satz von Cramér. Es demonstriert auch, dass der Maßwechsel über die exponentielle Familie asymptotisch optimal ist, weil μ_{t^*} die relative Entropie unter allen Wahrscheinlichkeitsmaßen minimiert, bezüglich derer die Ereignisse A_n asymptotisch relevant sind.

6.3.2 Große Abweichungen für empirische Verteilungen

Mithilfe von Satz 6.9 können wir noch eine stärkere Form der unteren Schranke für große Abweichungen vom Gesetz der großen Zahlen herleiten. Sei dazu ν ein Wahrscheinlichkeitsmaß auf einem metrischen Raum S mit Borelscher σ-Algebra $\mathcal{S}$. Wir nehmen an, dass S separabel ist, also eine abzählbare dichte Teilmenge besitzt. Seien

$$\hat{\nu}_n(\omega) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}, \qquad n \in \mathbb{N},$$

die empirischen Verteilungen einer Folge $(X_i)_{i\in\mathbb{N}}$ unabhängiger Zufallsvariablen mit Verteilung ν bzgl. P_ν. Aus dem Gesetz der großen Zahlen folgt, dass die Erwartungswerte einer integrierbaren

Funktion $U \in \mathcal{L}^1(\nu)$ bezüglich der zufälligen Maße $\hat{\nu}_n(\omega)$ für fast alle ω gegen die Erwartungswerte bzgl. ν konvergieren:

$$\int U \, d\hat{\nu}_n(\omega) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} U(X_i(\omega)) \quad \longrightarrow \quad \int U \, d\nu \qquad \text{für } P_\nu\text{-fast alle } \omega. \tag{6.3.5}$$

Die Ausnahmemenge kann dabei aber von der Funktion U abhängen. Da der Raum der stetigen beschränkten Funktionen $U : S \to \mathbb{R}$ im Allgemeinen nicht mehr separabel ist, ist die folgende Erweiterung des Gesetzes der großen Zahlen nicht offensichtlich:

Satz 6.10 (GGZ für empirische Verteilungen; Varadarajan). *Ist S separabel, dann konvergieren die empirischen Verteilungen fast sicher schwach gegen ν:*

$$\hat{\nu}_n(\omega) \quad \xrightarrow{w} \quad \nu \qquad \textit{für } P_\nu\textit{-fast alle } \omega. \tag{6.3.6}$$

Beweisskizze. Wir bezeichnen mit $C_{b,L}(S)$ den Raum aller beschränkten, Lipschitz-stetigen Funktionen $U : S \to \mathbb{R}$. Nach Satz 5.1 genügt es zu zeigen, dass mit Wahrscheinlichkeit 1

$$\int U \, d\hat{\nu}_n(\omega) \; \longrightarrow \; \int U \, d\nu \qquad \text{für alle } U \in C_{b,L}(S) \tag{6.3.7}$$

gilt. Man kann beweisen, dass im Raum $C_{b,L}(S)$ eine *bezüglich der Supremums-Norm* dichte, abzählbare Teilmenge $\{U_k : k \in \mathbb{N}\}$ existiert, wenn S separabel ist, siehe z.B. [DUDLEY: REAL ANALYSIS AND PROBABILITY]. Aus (6.3.5) können wir schließen, dass (6.3.7) außerhalb einer P_ν-Nullmenge N für die Funktionen U_k, $k \in \mathbb{N}$, simultan gilt. Hieraus folgt aber, dass (6.3.7) für $\omega \notin N$ sogar für alle beschränkten Lipschitz-stetigen Funktionen wahr ist: Ist $U \in C_{b,L}(S)$, dann existiert zu jedem $\varepsilon > 0$ ein $k \in \mathbb{N}$ mit $\sup |U - U_k| \leq \varepsilon$, und damit gilt

$$\limsup_{n\to\infty} \left| \int U \, d\hat{\nu}_n(\omega) - \int U \, d\nu \right| \;\leq\; \limsup_{n\to\infty} \left| \int U_k \, d\hat{\nu}_n(\omega) - \int U_k \, d\nu \right| + 2\varepsilon \;\leq\; 2\varepsilon$$

für alle $\omega \notin N$ und $\varepsilon > 0$. □

Nach dem Satz konvergiert die Wahrscheinlichkeit $P_\nu[\hat{\nu}_n \notin \mathcal{U}]$ für jede Umgebung $\mathcal{U}$ des Wahrscheinlichkeitsmaßes ν bzgl. der Topologie der schwachen Konvergenz gegen 0. Hierbei ist eine Menge $\mathcal{U}$ von Wahrscheinlichkeitsmaßen *offen*, wenn ihr Komplement $\mathcal{A} = WV(S) \setminus \mathcal{U}$ abgeschlossen ist, also alle schwachen Limiten von Folgen in $\mathcal{A}$ enthält.

Die Konvergenzgeschwindigkeit auf der exponentiellen Skala lässt sich durch ein Prinzip der großen Abweichungen auf dem Raum WV(S) der Wahrscheinlichkeitsverteilungen auf $(S, \mathcal{S})$ mit der Topologie der schwachen Konvergenz beschreiben:

Satz 6.11 (Sanov). *Die empirischen Verteilungen $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ erfüllen das folgende Prinzip der großen Abweichungen:*

(1). Obere Schranke: Für jede abgeschlossene Menge $\mathcal{A} \subseteq$WV(S) gilt:

$$\limsup_{n\to\infty} \frac{1}{n} \log P_\nu[\hat{\nu}_n \in \mathcal{A}] \quad \leq \quad -\inf_{\mu\in\mathcal{A}} H(\mu \mid \nu).$$

(2). Untere Schranke: Für jede offene Menge $\mathcal{O} \subseteq$WV(S) gilt:

$$\liminf_{n\to\infty} \frac{1}{n} \log P_\nu[\hat{\nu}_n \in \mathcal{O}] \quad \geq \quad -\inf_{\mu\in\mathcal{O}} H(\mu \mid \nu).$$

Beweis. (2). Zum Beweis der unteren Schranke wechseln wir wieder das zugrundeliegende Maß, und wenden Satz 6.9 an. Sei $\mathcal{O} \subseteq$ WV(S) offen und $\mu \in \mathcal{O}$. Nach (6.3.6) ist dann die Folge

$$A_n \quad = \quad \left\{(x_1,\ldots,x_n) \in S^n \,:\, \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \in \mathcal{O}\right\}$$

wesentlich bzgl. μ, denn

$$\mu^n[A_n] \quad = \quad P_\mu[\hat{\nu}_n \in \mathcal{O}] \quad \longrightarrow \quad 1$$

für $n \to \infty$. Daher folgt nach Korollar 6.9(2):

$$\liminf_{n\to\infty} \frac{1}{n} \log P_\nu[\hat{\nu}_n \in \mathcal{O}] \quad = \quad \liminf_{n\to\infty} \frac{1}{n} \log \nu^n[A_n] \quad \geq \quad -H(\mu \mid \nu).$$

Die Behauptung ergibt sich, da dies für alle $\mu \in \mathcal{O}$ gilt.

(1). Die obere Schranke beweisen wir hier nur für endliche Zustandsräume S, s. z.B. [DEMBO, ZEITOUNI: LARGE DEVIATIONS] für den Beweis im allgemeinen Fall. Ist S endlich, und μ eine bzgl. ν absolutstetige Wahrscheinlichkeitsverteilung mit Dichte $w = d\mu/d\nu$, dann gilt für alle $(x_1,\ldots,x_n) \in S^n$ mit empirischer Verteilung $\frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} = \mu$:

$$\begin{aligned}\frac{d\mu^n}{d\nu^n}(x_1,\ldots,x_n) \quad &= \quad \prod_{i=1}^{n} \frac{d\mu}{d\nu}(x_i) \quad = \quad \exp\left(\sum_{i=1}^{n} \log\left(\frac{d\mu}{d\nu}(x_i)\right)\right) \\ &= \quad \exp\left(n \int \log\left(\frac{d\mu}{d\nu}\right)\, d\mu\right) \quad = \quad \exp(nH(\mu \mid \nu)).\end{aligned}$$

Damit folgt

$$\begin{aligned}P_\nu[\hat{\nu}_n = \mu] \quad &= \quad \nu^n\left[\left\{(x_1,\ldots,x_n) \,:\, \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} = \mu\right\}\right] \\ &= \quad e^{-nH(\mu\mid\nu)} \cdot \mu^n\left[\left\{(x_1,\ldots,x_n) \,:\, \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} = \mu\right\}\right] \qquad (6.3.8) \\ &\leq \quad e^{-nH(\mu\mid\nu)}.\end{aligned}$$

Jeder empirischen Verteilung von n Elementen $x_1, \ldots, x_n \in S$ entspricht ein Histogramm $\vec{h} = (h_a)_{a \in S} \in \{0, 1, \ldots, n\}^S$. Für die Anzahl der möglichen empirischen Verteilungen gilt daher

$$\left| \left\{ \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \,:\, (x_1, \ldots, x_n) \in S^n \right\} \right| \quad \leq \quad (n+1)^{|S|}.$$

Nach (6.3.8) erhalten wir nun für eine beliebige Menge $\mathcal{A} \subseteq \mathrm{WV}(S)$ die (nicht-asymptotische) Abschätzung

$$P_\nu[\hat{\nu}_n \in \mathcal{A}] \quad = \quad \sum_{\mu \in \mathcal{A}} P_\nu[\hat{\nu}_n = \mu] \quad \leq \quad (n+1)^{|S|} \cdot \exp\left(-n \inf_{\mu \in \mathcal{A}} H(\mu \mid \nu) \right),$$

aus der die asymptotische obere Schranke wegen $|S| < \infty$ folgt.

□

Bemerkung. Wie der Beweis schon andeutet, gilt auch die obere Schranke in diesem Fall nur noch asymptotisch und modulo subexponentiell wachsender Faktoren. Der Übergang von endlichen zu allgemeinen Zustandsräumen ist bei der oberen Schranke nicht trivial, s. [DEMBO, ZEITOUNI: LARGE DEVIATIONS].

Den Satz von Sanov bezeichnet man gelegentlich auch als ein „Prinzip der großen Abweichungen auf Level II“, d.h. für die empirischen Verteilungen. Wir bemerken abschließend, dass sich eine Version des Satzes von Cramér, d.h. ein „Prinzip der großen Abweichungen auf Level I“ als Spezialfall ergibt:

Für eine stetige beschränkte Funktion $U : S \to \mathbb{R}$ und eine offene Menge $B \subseteq \mathbb{R}$ gilt nach dem Satz von Sanov:

$$P_\nu \left[\frac{1}{n} \sum_{i=1}^{n} U(X_i) \in B \right] \quad = \quad P_\nu[\hat{\nu}_n \in \mathcal{O}] \quad \succeq \quad \exp\left(- \inf_{\mu \in \mathcal{O}} H(\mu \mid \nu) \right)$$

mit $\mathcal{O} = \{\mu \in \mathrm{WV}(S) \,:\, \int U \, d\mu \in B\}$. Entsprechend ergibt sich eine analoge obere Schranke, falls B abgeschlossen ist.

6.3.3 Entropie und Kodierung

Als Spezialfall der Aussagen (1) und (2) in Satz 6.9 erhalten wir, wenn S endlich und ν die Gleichverteilung ist, zwei bekannte Aussagen aus der Informationstheorie: die „asymptotische Gleichverteilungseigenschaft” und den Quellenkodierungssatz von Shannon:

Wir betrachten die *möglichst effiziente Beschreibung/Kodierung einer Zufallsfolge.* Eine unbekannte Signalfolge mit Werten in einer endlichen Menge S (dem zugrundeliegenden „Alphabet“)

beschreibt man im einfachsten A-Priori-Modell durch unabhängige Zufallsvariablen $X_1, X_2, ...$ mit Verteilung μ, wobei $\mu(x)$ die relative Häufigkeit des Buchstabens x in der verwendeten Sprache ist. Eine „perfekte“ Kodierung ordnet jedem Wort mit einer vorgegebenen Anzahl n von Buchstaben, also jedem Element des Produktraums S^n, eine Binärfolge zu. Will man alle Wörter mit n Buchstaben perfekt kodieren, werden $n \cdot \log_2 |S|$ Bits benötigt. Wir betrachten stattdessen „effiziente“ Kodierungen, die nur den „meisten“ Wörtern mit n Buchstaben eindeutig eine Binärfolge zuordnen.

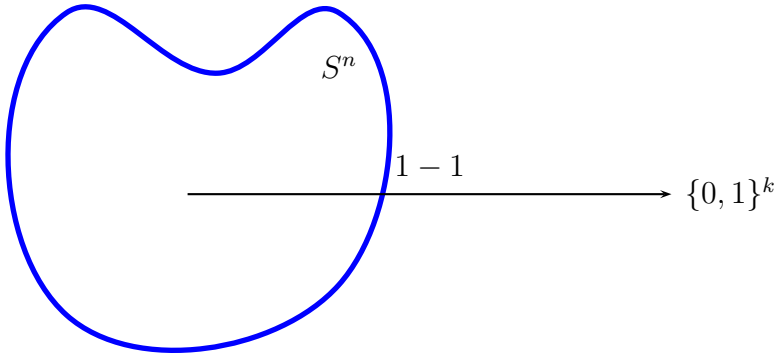


Abbildung 6.2: Perfekte Kodierung

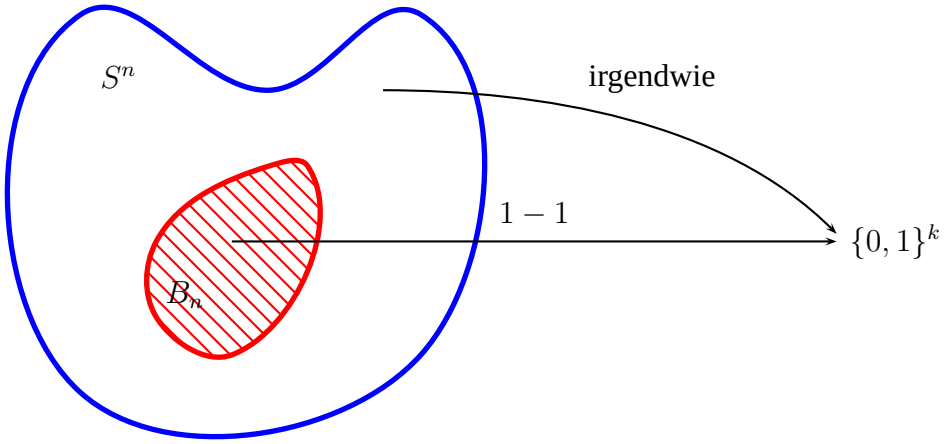


Abbildung 6.3: Effiziente Kodierung bzgl. einer Folge von wesentlichen Mengen B_n.

Sei $p_n(x_1, \ldots, x_n) = \prod_{i=1}^n \mu(x_i)$ die A-Priori-Wahrscheinlichkeit von $(x_1, \ldots, x_n)$ unter dem Produktmaß μ^n. Wählen wir für ν die Gleichverteilung auf S, dann gilt

$$|A_n| \;=\; \nu^n[A_n] \cdot |S|^n \qquad \text{für alle Mengen } A_n \subseteq S^n,\ n \in \mathbb{N}.$$

Damit können wir die Aussage von Satz 6.9 (1) folgendermaßen umformulieren:

Korollar 6.12 (Asymptotische Gleichverteilungseigenschaft). *Für jedes $\varepsilon > 0$ ist die Folge*

$$B_{n,\varepsilon} \;:=\; \left\{(x_1, ...x_n) \in S^n \;:\; e^{-n(H(\mu)+\varepsilon)} \le p_n(x_1, ..., x_n) \le e^{-n(H(\mu)-\varepsilon)}\right\}, \qquad n \in \mathbb{N},$$

wesentlich bzgl. μ, und es gilt

$$|B_{n,\varepsilon}| \;\leq\; e^{n(H(\mu)+\varepsilon)} \qquad \textit{für alle } n \in \mathbb{N}.$$

Beweis. Die Aussage folgt aus Satz 6.9 (1) wegen $H(\mu|\nu) = \log|S| - H(\mu)$ und $d\mu^n/d\nu^n = |S|^n \cdot p_n$. □

Die asymptotische Gleichverteilungseigenschaft zeigt, dass Folgen von wesentlichen Mengen existieren, deren Mächtigkeit auf der exponentiellen Skala nicht viel schneller als $\exp(n \cdot H(\mu))$ wächst.

Wieviele Elemente enthalten wesentliche Mengen mindestens ? Für $p \in (0,1)$ sei

$$K(n,p) \;=\; \inf\{|A_n| \,:\, A_n \subseteq S^n \text{ mit } P[(X_1,...,X_n) \in A_n] \geq p\}$$

die mindestens benötigte Anzahl von Wörtern, um den Text $(X_1,...,X_n)$ mit Wahrscheinlichkeit $\geq p$ korrekt zu erfassen. Dann ist $\log_2 K(n,p)$ die für eine korrekte binäre Kodierung von $(X_1,...,X_n)$ mit Wahrscheinlichkeit $\geq p$ mindestens benötigte Anzahl von Bits. Aus dem zweiten Teil von Satz 6.9 ergibt sich:

Korollar 6.13 (Quellenkodierungssatz von Shannon). *Für alle $p \in (0,1)$ gilt:*

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \log K(n,p) &= H(\mu), \quad \textit{bzw.} \\
\lim_{n\to\infty} \frac{1}{n} \log_2 K(n,p) &= H_2(\mu) := -\sum_{x:\mu(x)\neq 0} \mu(x) \log_2 \mu(x).
\end{aligned}$$

Insbesondere gilt: Ist A_n $(n \in \mathbb{N})$ wesentlich bzgl. μ, so ist $|A_n| \succeq \exp(nH(\mu))$.

Die Größe $\frac{1}{n}\log_2 K(n,p)$ kann als die für eine mit Wahrscheinlichkeit $\geq p$ korrekte Kodierung benötigte Zahl von Bits pro gesendetem Buchstaben interpretiert werden.

Bemerkung. Der Quellenkodierungssatz zeigt, dass es keine Folge von wesentlichen Mengen gibt, die auf der exponentiellen Skala deutlich langsamer wächst als die in Korollar 6.12 angegebenen Folgen $B_{n,\varepsilon}$ $(n \in \mathbb{N})$.

Beweis. Die Aussage ergibt sich wieder als Spezialfall von Satz 6.9 wenn wir $\nu = \mathcal{U}_S$ setzen:

Obere Schranke: $\limsup \frac{1}{n} \log K(n,p) \leq H(\mu)$:

Sei $\varepsilon > 0$. Nach Korollar 6.12 erfüllt die dort konstruierte Folge $B_{n,\varepsilon}$ $(n \in \mathbb{N})$ die Bedingung

$$\lim_{n\to\infty} P[(X_1,...,X_n) \in B_{n,\varepsilon}] \quad = \quad 1 \quad > \quad p,$$

und es gilt $\frac{1}{n}\log|B_{n,\varepsilon}| \le H(\mu)+\varepsilon$. Damit folgt

$$\limsup_{n\to\infty}\frac{1}{n}\log K(n,p) \quad\le\quad \limsup_{n\to\infty}\frac{1}{n}\log|B_{n,\varepsilon}| \quad\le\quad H(\mu)+\varepsilon.$$

Die Behauptung ergibt sich für $\varepsilon \to 0$.

Untere Schranke: $\liminf \frac{1}{n}\log K(n,p) \ge H(\mu)$:

Für Mengen $A_n \subseteq S^n$ mit $P[(X_1,...,X_n)\in A_n] \ge p$ gilt nach Satz 6.9 (2):

$$\liminf_{n\to\infty}\frac{1}{n}\log|A_n| \;=\; \liminf_{n\to\infty}\frac{1}{n}\log\left(\nu^n(A_n)\cdot S^n\right) \;\ge\; \log|S| - H(\mu|\nu) \;=\; H(\mu).$$

□

6.4 Likelihood

Praktisch unterscheidet man Wahrscheinlichkeitsverteilungen in der Schätz- und Testtheorie durch Likelihood-basierte statistische Verfahren. Der Zusammenhang von relativer Entropie und statistischer Unterscheidbarkeit kann genutzt werden, um die Qualität dieser Verfahren asymptotisch zu beurteilen.

6.4.1 Konsistenz von Maximum-Likelihood-Schätzern

Sei $(\nu_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsverteilungen auf $S=\mathbb{R}^d$ (oder einem diskreten Raum) mit Dichten (bzw. Massenfunktionen) f_θ wobei θ ein unbekannter Parameter ist. Ferner sei

$$L_n(\theta; x_1,...,x_n) \;=\; \prod_{i=1}^n f_\theta(x_i), \qquad \theta\in\Theta,$$

die Likelihoodfunktion zu n unabhängigen Stichproben $x_1,...,x_n$ von ν_θ. Ein wichtiges Ad-hoc-Verfahren zur Konstruktion eines Schätzers für θ ist das

Maximum-Likelihood-Prinzip: Wähle $\hat\theta(x_1,...,x_n)$ als den Parameterwert θ, für den die Likelihood der beobachteten Werte $x_1,\dots,x_n$ maximal ist.

Definition. *(1). Eine Zufallsvariable vom Typ $\hat\theta(X_1,\dots,X_n)$, $\hat\theta : S^n \to \Theta$ messbar, heißt* **Statistik** *der Daten $X_1,\dots,X_n$.*

*(2). Die Statistik heißt **Maximum-Likelihood-Schätzer (MLE)** für den Parameter θ, falls*

$$L_n(\hat\theta(x_1,...,x_n);\, x_1,...,x_n) \quad=\quad \max_{\theta\in\Theta} L_n(\theta;\, x_1,...,x_n) \qquad \textit{für alle } x_1,\dots,x_n\in S \textit{ gilt.}$$

Um einen Maximum-Likelihood-Schätzer zu berechnen, ist es oft günstig, die **log-Likelihood**

$$\theta \mapsto \log L_n(\theta;\, x_1, ..., x_n) = \sum_{i=1}^{n} \log f_\theta(x_i) \qquad \text{zu maximieren.}$$

Beispiel. (1). **Gaußmodell:** $\Theta = \{(m, v) \mid m \in \mathbb{R}, v > 0\},\ \nu_{m,v} = N(m, v)$.

$$L_n(m, v; X_1, ..., X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v}}\, e^{-\frac{(X_i - m)^2}{2v}}$$

ist maximal für $\hat{m}(X) = \overline{X}_n$, $\hat{v}(X) = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}_n)^2$. Dieser Maximum-Likelihood-Schätzer ist **nicht** erwartungstreu, da die Stichprobenvarianz mit dem Faktor $\frac{1}{n}$ statt $\frac{1}{n-1}$ gebildet wird.

(2). **Doppelexponentialverteilung:** $\Theta = \mathbb{R}$, $f_\theta(X_i) = \frac{1}{2}\, e^{-|X_i - \theta|}$.

$$\log L_n(\theta; X_1, ..., X_n) = -n \log 2 - \sum_{i=1}^{n} |X_i - \theta|$$

ist maximal, falls $\hat{\theta}$ ein Median von $X_1, ..., X_n$ ist.

(3). **Zufallszahlen** aus $[0, \theta]$, $\theta > 0$ unbekannt.

$$\begin{aligned} f_\theta(X_i) &= \frac{1}{\theta} I_{[0,\theta]}(X_i), \\ L_n(\theta; X_1, ..., X_n) &= \frac{1}{\theta^n} I_{[0,\theta]}(\max_{1 \le i \le n} X_i). \end{aligned}$$

Der Maximum-Likelihood-Schätzer ist $\hat{\theta}(X_1, ..., X_n) = \max_{1 \le i \le n} X_i$. Dieser Schätzer ist sicher nicht optimal, da mit Wahrscheinlichkeit 1 $\theta > \hat{\theta}(X_1, \ldots, X_n)$ gilt !

Wie das letzte Beispiel zeigt, sind Maximum-Likelihood-Schätzer für ein festes n nicht immer optimal. Unter bestimmten Voraussetzungen haben sie aber gute asymptotische Eigenschaften für $n \to \infty$. Sei etwa ν_θ ($\theta \in \Theta$) eine einparametrige (d.h. $\Theta \subseteq \mathbb{R}$) Familie von Wahrscheinlichkeitsverteilungen mit Dichten bzw. Massenfunktionen f_θ. Es gelte:

Annahme (Unimodalität): Für alle $n \in \mathbb{N}$ und $x \in S^n$ existiert ein $\hat{\theta}_n(x_1, ..., x_n)$, sodass

$$\theta \mapsto L_n(\theta; x_1, ..., x_n) \begin{cases} \text{ist monoton wachsend für } \theta \le \hat{\theta}_n(x_1, ..., x_n). \\ \text{ist monoton fallend für } \theta \ge \hat{\theta}_n(x_1, ..., x_n). \end{cases}$$

Bemerkung. (1). Die Annahme ist z.B. erfüllt, falls $\theta \mapsto \log f_\theta(x)$ für jedes x konkav ist - denn dann ist auch $\log L_n(\theta, x_1, ..., x_n) = \sum_{i=1}^{n} \log f_\theta(x_i)$ konkav in θ.

(2). $\hat{\theta}_n(X_1, ..., X_n)$ ist im unimodalen Fall eindeutiger Maximum-Likelihood-Schätzer für θ.

Satz 6.14. *Es gelte die Annahme, sowie $\nu_\theta \neq \nu_{\widetilde{\theta}}$ für $\theta \neq \widetilde{\theta}$. Dann ist $\hat{\theta}_n(X_1, \ldots, X_n)$ $(n \in \mathbb{N})$ eine **konsistente** Folge von Schätzern für θ, d.h. für jedes $\varepsilon > 0$ gilt:*

$$P_\theta[|\hat{\theta}_n(X_1, ..., X_n) - \theta| < \varepsilon] \to 1 \quad \textit{für } n \to \infty.$$

Beweis. Wegen der Unimodalität gilt $\hat{\theta}_n(x_1, ..., x_n) \in (\theta - \varepsilon, \theta + \varepsilon)$ falls

$$L_n(\theta; x_1, ..., x_n) \; > \; L_n(\theta \pm \varepsilon; x_1, ..., x_n).$$

Also:

$$P_\theta[|\hat{\theta}_n(X_1, ..., X_n) - \theta| < \varepsilon] \geq P_\theta \left[\frac{L_n(\theta; X_1, ..., X_n)}{L_n(\theta \pm \varepsilon; X_1, ..., X_n)} > 1 \right].$$

Die rechte Seite konvergiert aber für $n \to \infty$ nach Satz 6.7 für jedes θ gegen 1. □

Bemerkung (Asymptotische Normalität von Maximum-Likelihood-Schätzern). Unter geeigneten Regularitätsvoraussetzungen an die Dichten f_θ gilt für die Maximum-Likelihood-Schätzer neben der Konsistenz (also dem Gesetz der großen Zahlen) auch ein zentraler Grenzwertsatz:

Satz (Fisher, Wilkes, Wold). Unter geeigneten Voraussetzungen gilt:

$$\sqrt{n}(\hat{\theta}_n(X_1, ..., X_n) - \theta) \; \xrightarrow{\mathcal{D}} \; N\left(0, \frac{1}{I(\theta)}\right),$$

wobei

$$I(\theta) \; = \; \int \left| \frac{\partial}{\partial \theta} \, \log f_\theta(x) \right|^2 \nu_\theta(dx) \; = \; \lim_{\varepsilon \to 0} \, \frac{2}{\varepsilon^2} \, H(\nu_{\theta+\varepsilon} | \, \nu_\theta)$$

die **Fisher-Information** des statistischen Modells ist.

Da man andererseits unter geeigneten Regularitätsbedingungen zeigen kann, daß die Varianz eines erwartungstreuen Schätzers für θ basierend auf n unabhängigen Stichproben stets größer als $\frac{1}{nI(\theta)}$ ist (*Informationsungleichung von Cramér-Rao*), folgt, daß Maximum-Likelihood-Schätzer in gewisser Hinsicht asymptotisch optimal sind.

6.4.2 Asymptotische Macht von Likelihoodquotiententests

Angenommen, wir haben n unabhängige Stichproben $X_1, ..., X_n$ von einer unbekannten Verteilung vorliegen und wir gehen davon aus, daß die zugrundeliegende Verteilung aus einer Familie μ_θ $(\theta \in \Theta)$ von Wahrscheinlichkeitsverteilungen kommt. Sei Θ_0 eine Teilmenge des Parameterbereichs. Wir wollen entscheiden zwischen der

$$\textit{Nullhypothese } H_0\text{:} \qquad \text{»}\theta \in \Theta_0\text{«}$$

und der

$$\textit{Alternative } H_1\text{:} \qquad \text{»}\theta \notin \Theta_0\text{«}$$

Ein **Hypothesentest** für ein solches Problem ist bestimmt durch eine messbare Teilmenge $C \subseteq S^n$ (den **Verwerfungsbereich**) mit zugehöriger Entscheidungsregel:

$$\text{akzeptiere } H_0 \iff (X_1, ..., X_n) \notin C.$$

Beispiel (t-Test). Seien $X_1, X_2, \ldots, X_n$ unabhängige Stichproben von einer Normalverteilung mit unbekanntem Parameter $(m, v) \in \Theta = \mathbb{R} \times \mathbb{R}^+$. Wir wollen testen, ob der Mittelwert der Verteilung einen bestimmten Wert m_0 hat:

$$\textit{Nullhypothese } H_0\text{:} \qquad \text{»}m = m_0\text{«}\,, \qquad \Theta_0 = \{m_0\} \times \mathbb{R}^+\,.$$

Ein solches Problem tritt z.B.in der Qualitätskontrolle auf, wenn man überprüfen möchte, ob ein Sollwert m_0 angenommen wird. Eine andere Anwendung ist der Vergleich zweier Verfahren, wobei X_i die Differenz der mit beiden Verfahren erhaltenen Messwerte ist. Die Nullhypothese mit $m_0 = 0$ besagt hier, daß kein signifikanter Unterschied zwischen den Verfahren besteht.

Im *t–Test* für obiges Testproblem wird die Nullhypothese akzeptiert, falls der Betrag der *Student-schen t-Statistik* unterhalb einer angemessen zu wählenden Konstanten c liegt, bzw. verworfen, falls

$$|T_{n-1}| \;=\; \left|\frac{\sqrt{n}\cdot(\overline{X}_n - m_0)}{\sqrt{V_n}}\right| \;>\; c$$

gilt.

Seien nun allgemein $X_1, X_2, \ldots$ unter P_θ unabhängige Zufallsvariablen mit Verteilung ν_θ. Bei einem Hypothesentest können zwei Arten von Fehlern auftreten:

Fehler 1. Art: H_0 wird verworfen, obwohl wahr. Wahrscheinlichkeit:

$$P_\theta[(X_1, ..., X_n) \in C] = \mu_\theta^n(C)\,, \qquad \theta \in \Theta_0.$$

Fehler 2. Art: H_0 wird akzeptiert, obwohl falsch. Wahrscheinlichkeit:

$$P_\theta[(X_1, ..., X_n) \notin C] = \mu_\theta^n(C^C) \,, \qquad \theta \in \Theta \setminus \Theta_0.$$

Obwohl das allgemeine Testproblem im Prinzip symmetrisch in H_0 und H_1 ist, interpretiert man beide Fehler i.a. unterschiedlich. Die Nullhypothese beschreibt in der Regel den Normalfall, die Alternative eine Abweichung oder einen zu beobachtenden Effekt. Da ein Test Kritiker überzeugen soll, sollte die Wahrscheinlichkeit für den Fehler 1. Art (Effekt prognostiziert, obgleich nicht vorhanden) unterhalb einer vorgegebenen (kleinen) Schranke α liegen. Die Wahrscheinlichkeit

$$\mu_\theta^n(C) \,, \qquad \theta \in \Theta \setminus \Theta_0 \,,$$

daß kein Fehler 2. Art auftritt, sollte unter dieser Voraussetzung möglichst groß sein.

Definition. *Die Funktion*

$$G(\theta) \;=\; P_\theta[(X_1, ..., X_n) \in C] \;=\; \mu_\theta^n(C)$$

*heißt **Gütefunktion** des Tests. Der Test hat **Niveau** α, falls*

$$G(\theta) \;\leq\; \alpha \quad \textit{für alle } \theta \in \Theta_0$$

*gilt. Die Funktion $G(\theta)$ mit $\theta \in \Theta_1$ heißt **Macht** des Tests.*

Beispiel. Der Studentsche t-Test hat Niveau α falls c ein $(1 - \frac{\alpha}{2})$-Quantil der Studentschen t-Verteilung mit $n - 1$ Freiheitsgraden ist.

Ein Ziel bei der Konstruktion eines Testverfahrens sollte es sein, die Machtfunktion bei vorgegebenem Niveau zu maximieren. Dies ist im Allgemeinen nicht simultan für alle Parameter $\theta \in \Theta \setminus \Theta_0$ möglich. Eine Ausnahme bildet der Fall einer einfachen Hypothese und Alternative, in dem ein optimaler Test existiert:

a) Einfache Hypothese und Alternative

Angenommen, wir wissen, daß die Stichproben von einer der beiden Verteilungen $\mu_0 := \nu$ und $\mu_1 := \mu$ stammen und wir wollen entscheiden zwischen der

$$\textit{Nullhypothese } H_0\text{:} \qquad \text{»}X_i \sim \nu\text{«}$$

und der

$$\textit{Alternative } H_1\text{:} \qquad \text{»}X_i \sim \mu\text{«}.$$

Ein solches Problem tritt in Anwendungen zwar selten auf, bildet aber einen ersten Schritt zum Verständnis allgemeinerer Testprobleme. Sei

$$\varrho_n(x_1, \ldots, x_n) \;=\; \frac{L_n(\mu; x_1, ..., x_n)}{L_n(\nu; x_1, ..., x_n)} \;=\; \prod_{i=1}^{n} \frac{f(x_i)}{g(x_i)}$$

der Quotient der Likelihoods der Stichproben $x_1, \ldots, x_n$ im Produktmodell. Hierbei sind f und g die Dichte bzw. Massenfunktion der Verteilungen μ und ν.

Definition. *Ein Test mit Entscheidungsregel*

$$\textit{Akzeptiere } H_0 \iff \varrho_n(X_1, ..., X_n) \le c,$$

$c \in (0, \infty)$*, heißt* ***Likelihoodquotiententest.***

Der Verwerfungsbereich eines Likelihoodquotiententests ist also $C = \{\varrho_n > c\}$, die Wahrscheinlichkeit für den Fehler 1. Art beträgt

$$\alpha := \nu^n(\varrho_n > c).$$

Satz 6.15 (Neyman-Pearson-Lemma). *Der Likelihoodquotiententest mit Parameter c ist der* ***beste Test zum Niveau*** *α, d.h. jeder Test mit*

$$\textit{Wahrscheinlichkeit (Fehler 1.Art)} \le \alpha$$

hat eine kleinere Macht (d.h. eine höhere Wahrscheinlichkeit für den Fehler 2. Art).

Beweis. Sei $A \subseteq S^n$ der Verwerfungsbereich eines Tests mit $\nu^n(A) \le \alpha$, und sei

$$\chi = I_C - I_A = I_{A^C} - I_{C^C}.$$

Zu zeigen ist:

$$0 \le \mu^n(A^C) - \mu^n(C^C) = \int \chi \, d\mu^n.$$

Offensichtlich gilt $\chi \ge 0$ auf $C = \{\varrho_n > c\}$ und $\chi \le 0$ auf $C^C = \{\varrho_n \le c\}$, also $\chi \cdot (\varrho_n - c) \ge 0$. Durch Integration erhalten wir:

$$0 \le \int \chi \cdot (\varrho_n - c) \, d\nu^n = \int \chi \, d\mu^n - c \cdot \int \chi \, d\nu^n \le \int \chi \, d\mu^n,$$

da $\int \chi \, d\nu^n = \nu^n(C) - \nu^n(A) \ge 0$. □

Wie gut ist der Likelihoodquotiententest (also der beste Test zur Unterscheidung von ν und μ) asymptotisch für große n? Wir betrachten ein festes Niveau $\alpha \in (0,1)$, und wählen $c_n \in (0,\infty)$ $(n \in \mathbb{N})$ mit

$$\nu^n(\varrho_n > c_n) \leq \alpha \leq \nu^n(\varrho_n \geq c_n) \tag{6.4.1}$$

Satz 6.16 (Asymptotische Macht des Likelihoodquotiententests). *Es gilt:*

(i)
$$\frac{1}{n} \log c_n \longrightarrow -H(\nu|\mu) \qquad \textit{für } n \to \infty.$$

(ii)
$$\frac{1}{n} \log \mu^n(\varrho_n \leq c_n) \longrightarrow -H(\nu|\mu) \qquad \textit{für } n \to \infty,$$

d.h. die Wahrscheinlichkeit für den Fehler 2. Art fällt exponentiell mit Rate $H(\nu|\mu)$.

Beweis. (i) Sei $\varepsilon > 0$. Für große n gilt nach dem Satz von Shannon-McMillan:

$$\nu^n\left(\varrho_n > e^{-n(H(\nu|\mu)+\varepsilon)}\right) > \alpha \overset{6.4.1}{\geq} \nu^n(\varrho_n > c_n).$$

Es folgt $e^{-n(H(\nu|\mu)+\varepsilon)} < c_n$. Analog zeigt man $e^{-n(H(\nu|\mu)-\varepsilon)} > c_n$. Die Behauptung folgt dann für $\varepsilon \to 0$.

(ii) a) *Untere Schranke:* Wegen

$$\nu^n(\varrho_n \leq c_n) \geq 1-\alpha > 0 \qquad \forall\, n \in \mathbb{N}$$

folgt nach Korollar 6.9:

$$\underline{\lim}\, \frac{1}{n} \log \mu^n(\varrho_n \leq c_n) \geq -H(\nu|\mu).$$

Obere Schranke: Wegen

$$\mu^n(\varrho_n \leq c_n) = \int_{\varrho_n \leq c_n} \varrho_n \, d\nu^n \leq c_n$$

folgt nach (i)

$$\overline{\lim}\, \frac{1}{n} \log \mu^n(\varrho_n \leq c_n) \leq \overline{\lim}\, \frac{1}{n} \log c_n = -H(\nu|\mu).$$

□

Der Satz demonstriert erneut, daß die relative Entropie ein gutes Maß für die Unterscheidbarkeit zweier Wahrscheinlichkeitsverteilungen ist.

b) Zusammengesetzte Hypothesen und/oder Alternativen

Wenn Θ_0 und/oder Θ_1 aus mehr als einem Element bestehen, kann man den **verallgemeinerten Likelihoodquotienten**

$$\bar{\varrho}_n(x_1, \ldots, x_n) \;=\; \frac{\sup_{\theta\in\Theta_1} L_n(\theta; x_1, ..., x_n)}{\sup_{\theta\in\Theta_0} L_n(\theta; x_1, ..., x_n)} \;=\; \frac{\text{max. Lik. von } x\text{, falls } H_1 \text{ wahr}}{\text{max. Lik. von } x\text{, falls } H_0 \text{ wahr}}$$

betrachten. Der entsprechende Likelihoodquotiententest ist ähnlich wie der Maximum-Likelihood-Schätzer ein häufig verwendetes ad hoc Verfahren. Im Gegensatz zum Fall einer einfachen Hypothese und Alternative ist der verallgemeinerte Likelihoodquotiententest allerdings nicht immer optimal.

Beispiel. Im Beispiel von oben ist der t-Test der Likelihoodquotiententest. Mit einem Neyman-Pearson-Argument kann man zeigen, daß er im Gaußschen Produktmodell der beste unverfälschte Test zu einem vorgegebenen Niveau α ist, d.h. der mächtigste Test mit

$$G(\theta) \leq \alpha \quad \forall\, \theta \in \Theta_0 \quad \text{und} \quad G(\theta) \geq \alpha \quad \forall\, \theta \in \Theta_1.$$

Auch in nicht-Gaußschen Modellen wird häufig der t-Test eingesetzt – eine partielle Rechtfertigung dafür liefert der zentrale Grenzwertsatz.