

# Einführung in die Wahrscheinlichkeitstheorie

Prof. Dr. Andreas Eberle

20. April 2010

# Inhaltsverzeichnis

<b>Inhaltsverzeichnis</b>	<b>2</b>
<b>1 Diskrete Zufallsvariablen</b>	<b>8</b>
1.1 Ereignisse und ihre Wahrscheinlichkeit . . . . .	10
Ereignisse als Mengen . . . . .	10
Wahrscheinlichkeitsverteilungen . . . . .	12
Diskrete Wahrscheinlichkeitsverteilungen . . . . .	14
Spezielle Wahrscheinlichkeitsverteilungen . . . . .	17
Gleichverteilungen / Laplace-Modelle . . . . .	17
Empirische Verteilungen . . . . .	18
1.2 Diskrete Zufallsvariablen und ihre Verteilung . . . . .	22
Binomialverteilung . . . . .	24
Poissonverteilung . . . . .	26
Hypergeometrische Verteilung . . . . .	27
1.3 Simulation von Gleichverteilungen . . . . .	28
Lineare Kongruenzgeneratoren (LCG) . . . . .	29
Shift-Register-Generatoren . . . . .	34
Kombination von Zufallszahlengeneratoren . . . . .	35
Zufallszahlen aus $[0,1)$ . . . . .	35
Zufallspermutationen . . . . .	35
1.4 Erwartungswert . . . . .	36
Transformationssatz . . . . .	37
Linearität und Monotonie des Erwartungswertes . . . . .	39
<b>2 Bedingte Wahrscheinlichkeiten und Unabhängigkeit</b>	<b>43</b>
2.1 Bedingte Wahrscheinlichkeiten . . . . .	43
Berechnung von Wahrscheinlichkeiten durch Fallunterscheidung . . . . .	44

---

Bayessche Regel . . . . .	46
2.2 Mehrstufige diskrete Modelle . . . . .	47
Produktmodelle . . . . .	50
Markov-Ketten . . . . .	51
Berechnung von Wahrscheinlichkeiten . . . . .	54
2.3 Unabhängigkeit von Ereignissen . . . . .	55
Verteilungen für unabhängige Ereignisse . . . . .	57
Geometrische Verteilung . . . . .	57
Binomialverteilung . . . . .	58
2.4 Unabhängige Zufallsvariablen und Random Walk . . . . .	63
Unabhängigkeit von diskreten Zufallsvariablen . . . . .	63
Der Random Walk auf $\mathbb{Z}$ . . . . .	64
Symmetrischer Random Walk . . . . .	67
2.5 Simulationsverfahren . . . . .	71
Das direkte Verfahren . . . . .	71
Acceptance-Rejection-Verfahren . . . . .	72
<b>3 Konvergenzsätze und Monte Carlo Verfahren</b> . . . . .	<b>75</b>
3.1 Varianz und Kovarianz . . . . .	76
3.2 Das schwache Gesetz der großen Zahlen . . . . .	80
3.3 Monte Carlo-Verfahren . . . . .	82
Varianzreduktion durch Importance Sampling . . . . .	85
3.4 Gleichgewichte von Markov-Ketten . . . . .	88
Gleichgewichte und Stationarität . . . . .	88
Metropolis-Algorithmus und Gibbs-Sampler . . . . .	92
Die Metropolis-Kette . . . . .	92
Der Gibbs-Sampler . . . . .	93
3.5 Konvergenz ins Gleichgewicht . . . . .	96
<b>4 Stetige und Allgemeine Modelle</b> . . . . .	<b>101</b>
4.1 Unendliche Kombinationen von Ereignissen . . . . .	101
4.2 Allgemeine Wahrscheinlichkeitsräume . . . . .	109
Beispiele von Wahrscheinlichkeitsräumen . . . . .	109
Konstruktion von $\sigma$ -Algebren . . . . .	112
Existenz und Eindeutigkeit von Wahrscheinlichkeitsverteilungen . . . . .	114

4.3	Allgemeine Zufallsvariablen und ihre Verteilung . . . . .	118
	Allgemeine Zufallsvariablen . . . . .	119
	Verteilungen von Zufallsvariablen . . . . .	121
	Charakterisierung der Verteilung . . . . .	122
4.4	Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$ . . . . .	125
	Eigenschaften der Verteilungsfunktion . . . . .	125
	Diskrete Verteilungen . . . . .	126
	Stetige Verteilungen . . . . .	128
	Transformation von absolutstetigen Zufallsvariablen . . . . .	133
4.5	Quantile und Inversionsverfahren . . . . .	135
	Quantile . . . . .	135
	Konstruktion und Simulation reellwertiger Zufallsvariablen . . . . .	137
4.6	Normalapproximation der Binomialverteilung . . . . .	142
	Der Satz von De Moivre - Laplace . . . . .	143
	Approximative Konfidenzintervalle . . . . .	149
<b>5</b>	<b>Unabhängigkeit und Produktmodelle</b>	<b>152</b>
5.1	Unabhängigkeit in allgemeinen Modellen . . . . .	152
	Unabhängigkeit von Ereignissen . . . . .	152
	Unabhängigkeit von Zufallsvariablen . . . . .	155
	Konfidenzintervalle für Quantile . . . . .	159
5.2	Gemeinsame Verteilungen und endliche Produktmodelle . . . . .	161
	Wahrscheinlichkeitsverteilungen auf endlichen Produkträumen . . . . .	161
	Absolutstetigkeit von multivariaten Verteilungen . . . . .	164
	Absolutstetigkeit von endlichen Produktmodellen . . . . .	164
	Gemeinsame Verteilungen . . . . .	165
	Diskrete Zufallsvariablen . . . . .	167
	Reelle Zufallsvariablen . . . . .	169
5.3	Unendliche Produktmodelle . . . . .	173
	Konstruktion von unabhängigen Zufallsvariablen . . . . .	173
	Unendliche Produktmaße . . . . .	178
5.4	Asymptotische Ereignisse . . . . .	179
	Das 0-1-Gesetz von Kolmogorov . . . . .	181
	Anwendungen auf Random Walks und Perkulationsmodelle . . . . .	181

<b>6</b>	<b>Erwartungswert und Varianz</b>	<b>186</b>
6.1	Erwartungswert . . . . .	186
	Definition des Erwartungswerts . . . . .	186
	Elementare Zufallsvariablen . . . . .	186
	Nichtnegative Zufallsvariablen . . . . .	187
	Allgemeine Zufallsvariablen . . . . .	190
	Eigenschaften des Erwartungswerts . . . . .	190
	Konvergenzsätze . . . . .	192
6.2	Berechnung von Erwartungswerten; Dichten . . . . .	194
	Diskrete Zufallsvariablen . . . . .	195
	Allgemeine Zufallsvariablen . . . . .	195
	Zufallsvariablen mit Dichten . . . . .	198
	Existenz von Dichten . . . . .	202
6.3	Varianz, Kovarianz und lineare Regression . . . . .	203
	Varianz und Standardabweichung . . . . .	203
	Quadratintegrierbare Zufallsvariablen . . . . .	205
	Beste Prognosen . . . . .	206
	Kovarianz und Korrelation . . . . .	208
	Lineare Regression . . . . .	211
	Unabhängigkeit und Unkorreliertheit . . . . .	215
<b>7</b>	<b>Gesetze der großen Zahlen</b>	<b>217</b>
7.1	Ungleichungen und Konvergenz von ZVn . . . . .	217
	Konvergenzbegriffe für Zufallsvariablen . . . . .	217
	Die Markov-Čebyšev-Ungleichung . . . . .	220
	Wichtige Spezialfälle: . . . . .	220
	Die Jensensche Ungleichung . . . . .	222
7.2	Starke Gesetze der großen Zahlen . . . . .	224
	Das schwache Gesetz der großen Zahlen . . . . .	225
	Das starke Gesetz für quadratintegrierbare Zufallsvariablen . . . . .	226
	Von $\mathcal{L}^2$ nach $\mathcal{L}^1$ mit Unabhängigkeit . . . . .	230
7.3	Empirische Verteilungen . . . . .	234
	Schätzen von Kenngrößen einer unbekanntem Verteilung . . . . .	234
	Konvergenz der empirischen Verteilungsfunktionen . . . . .	236
	Histogramme und Multinomialverteilung . . . . .	238

7.4	Entropie . . . . .	240
	Definition und Eigenschaften . . . . .	241
	Extrema der Entropie: . . . . .	242
	Statistische Interpretation der Entropie . . . . .	244
	Entropie und Kodierung . . . . .	245
<b>8</b>	<b>Grenzwertsätze</b>	<b>248</b>
8.1	Charakteristische und Momentenerzeugende Funktionen . . . . .	249
	Definition und Eigenschaften . . . . .	249
	Inversion der Fouriertransformation . . . . .	253
8.2	Erste Anwendungen auf Grenzwertsätze . . . . .	255
	Zentraler Grenzwertsatz . . . . .	256
	Große Abweichungen vom Gesetz der großen Zahlen . . . . .	257
8.3	Verteilungskonvergenz . . . . .	262
	Schwache Konvergenz von Wahrscheinlichkeitsverteilungen . . . . .	263
	Konvergenz der Verteilungen von Zufallsvariablen . . . . .	268
	Existenz schwach konvergenter Teilfolgen . . . . .	271
	Schwache Konvergenz über charakteristische Funktionen . . . . .	273
8.4	Der Zentrale Grenzwertsatz . . . . .	275
	ZGS für Summen von i.i.d. Zufallsvariablen . . . . .	276
	Normalapproximationen . . . . .	278
	Heavy Tails, Konvergenz gegen $\alpha$ -stabile Verteilungen . . . . .	281
	Der Satz von Lindeberg-Feller . . . . .	282
<b>9</b>	<b>Multivariate Verteilungen und Statistik</b>	<b>287</b>
9.1	Mehrstufige Modelle . . . . .	287
	Stochastische Kerne und der Satz von Fubini . . . . .	287
	Wichtige Spezialfälle . . . . .	290
	Bedingte Dichten und Bayessche Formel . . . . .	291
9.2	Summen unabhängiger Zufallsvariablen, Faltung . . . . .	295
	Verteilungen von Summen unabhängiger Zufallsvariablen . . . . .	296
	Wartezeiten, Gamma-Verteilung . . . . .	298
9.3	Transformationen, Gaußmodelle und Parameterschätzung . . . . .	300
	Der Dichtetransformationssatz . . . . .	300
	Multivariate Normalverteilungen und multivariater ZGS . . . . .	301

---

Parameterschätzung im Gaußmodell . . . . .	305
Exkurs zu Hypothesentests . . . . .	309
<b>10 Bedingte Erwartungen</b>	<b>312</b>
10.1 Bedingen auf diskrete Zufallsvariablen . . . . .	312
Bedingte Erwartungen als Zufallsvariablen . . . . .	312
Formel von der totalen Wahrscheinlichkeit . . . . .	314
Bedingte Varianz . . . . .	315
Anwendung auf zufällige Summen . . . . .	316
Charakterisierende Eigenschaften der bedingten Erwartung . . . . .	317
10.2 Erzeugende Funktionen, Verzweigungsprozesse und Erneuerungen . . . . .	318
Erzeugende Funktionen von ganzzahligen Zufallsvariablen . . . . .	318
Erzeugende Funktionen zufälliger Summen . . . . .	320
Galton-Watson-Verzweigungsprozesse . . . . .	320
Rekurrente Ereignisse und Erneuerungsgleichung . . . . .	323
10.3 Bedingen auf allgemeine Zufallsvariablen . . . . .	326
Das Faktorisierungslemma . . . . .	327
Definition allgemeiner bedingter Erwartung . . . . .	328
Diskreter und absolutstetiger Fall . . . . .	331
Reguläre bedingte Verteilungen . . . . .	333
<b>Stichwortverzeichnis</b>	<b>337</b>

# Kapitel 1

## Diskrete Zufallsvariablen

Unser Ziel in diesem Kapitel ist die mathematische Modellierung von **Zufallsvorgängen**. Einfache Beispiele für Zufallsvorgänge sind das Werfen eines Würfels oder Münzwürfe. Anhand dieser Beispiele wollen wir zunächst einige grundlegende Begriffe der Wahrscheinlichkeitstheorie veranschaulichen.

NOTATIONEN:  $|A|$  bezeichnet die Anzahl der Elemente einer Menge  $A$ ,  $A^C$  bezeichnet das Komplement der Menge  $A$  innerhalb einer bestimmten Menge  $B$ , die  $A$  enthält.

**Beispiel** (Werfen eines Würfels).

- **Mögliche Fälle** sind 1, 2, 3, 4, 5, 6. Mit  $\Omega = \{1, 2, 3, 4, 5, 6\}$  wird die Menge aller möglichen Fälle bezeichnet. Ein **Elementarereignis** ist ein möglicher Fall, also ein Element  $\omega \in \Omega$ .
- **Ereignisse** sind die Objekte, denen man eine Wahrscheinlichkeit zuordnen kann, zum Beispiel:

»Augenzahl ist 3«	$\{3\}$
»Augenzahl ist gerade«	$\{2, 4, 6\}$
»Augenzahl ist <b>nicht</b> gerade«	$\{1, 3, 5\} = \{2, 4, 6\}^C$
»Augenzahl ist größer als 3«	$\{4, 5, 6\}$
»Augenzahl ist gerade <b>und</b> größer als 3«	$\{4, 6\} = \{2, 4, 6\} \cap \{4, 5, 6\}$
»Augenzahl gerade <b>oder</b> größer als 3«	$\{2, 4, 5, 6\} = \{2, 4, 6\} \cup \{4, 5, 6\}$

Jedes **Ereignis** kann durch eine **Teilmenge**  $A$  von  $\Omega$  dargestellt werden!



- **Wahrscheinlichkeiten** werden mit  $P$  (für »probability«) bezeichnet. Zum Beispiel sollte für einen »fairen« Würfel gelten:

$$P[\text{»3«}] = \frac{1}{6},$$

$$P[\text{»Augenzahl gerade«}] = \frac{\text{Anzahl günstige Fälle}}{\text{Anzahl mögliche Fälle}} = \frac{|\{2, 4, 6\}|}{|\{1, 2, 3, 4, 5, 6\}|} = \frac{3}{6} = \frac{1}{2},$$

$$P[\text{»Augenzahl gerade oder größer als 3«}] = \frac{4}{6} = \frac{2}{3}.$$

- **Zufallsvariablen** sind Abbildungen  $X : \Omega \rightarrow S$ , wobei  $S$  eine beliebige Menge ist, zum Beispiel:

$$X(\omega) = \omega, \quad \text{»Augenzahl des Wurfs«, oder}$$

$$X(\omega) = \begin{cases} 1 & \text{falls } \omega \in \{1, 2, 3, 4, 5\}, \\ -5 & \text{falls } \omega \in 6, \end{cases} \quad \text{»Gewinn bei einem fairen Spiel«.}$$

**Beispiel** (Münzwürfe). a) EIN MÜNZWURF:

Die Menge der möglichen Fälle ist  $\Omega = \{0, 1\}$ , wobei 0 für »Kopf« und 1 für »Zahl« steht.

Die Wahrscheinlichkeiten sind

$$P[\{1\}] = p \quad \text{und} \quad P[\{0\}] = 1 - p \quad \text{mit } 0 \leq p \leq 1.$$

Für  $p = \frac{1}{2}$  ist der Münzwurf fair.

b) ENDLICH VIELE FAIRE MÜNZWÜRFE:

Die Menge der möglichen Fälle lautet

$$\Omega = \{\omega = (x_1, \dots, x_n) \mid x_i \in \{0, 1\}\} =: \{0, 1\}^n.$$

Alle Ausgänge sind genau dann gleich wahrscheinlich, wenn  $P[\{\omega\}] = 2^{-n}$  für alle  $\omega \in \Omega$  gilt. Dies wird im folgenden angenommen. Zufallsvariablen von Interesse sind beispielsweise:

- $X_i(\omega) := x_i$ , das Ergebnis des  $i$ -ten Wurfs. Das Ereignis » $i$ -ter Wurf ist Kopf« wird durch die Menge  $A_i = \{\omega \in \Omega \mid X_i(\omega) = 0\} =: \{X_i = 0\}$  beschrieben, und hat die Wahrscheinlichkeit  $P[A_i] = \frac{1}{2}$ .
- $S_n(\omega) := \sum_{i=1}^n X_i(\omega)$ , die Anzahl der Einsen in  $n$  Münzwürfen. Das Ereignis »genau  $k$ -mal Zahl« wird durch die Menge  $A = \{\omega \in \Omega \mid S_n(\omega) = k\} =: \{S_n = k\}$  beschrieben und hat die Wahrscheinlichkeit  $P[A] = \binom{n}{k} 2^{-n}$ .

c) UNENDLICH VIELE MÜNZWÜRFE:

Die Menge der möglichen Fälle ist nun

$$\Omega = \{\omega = (x_1, x_2, \dots) \mid x_i \in \{0, 1\}\} = \{0, 1\}^{\mathbb{N}}.$$

Diese Menge ist überabzählbar, da die Abbildung

$$\begin{aligned} \Omega &\rightarrow [0, 1] \\ (x_1, x_2, \dots) &\mapsto 0.x_1x_2\dots \end{aligned}$$

surjektiv ist, (wobei das Einheitsintervall binär dargestellt wird). Die Definition von Ereignissen und Wahrscheinlichkeiten ist daher in diesem Fall aufwändiger. Wahrscheinlichkeitsverteilungen auf überabzählbaren Mengen werden systematisch in der Vorlesung »Einführung in die Wahrscheinlichkeitstheorie« betrachtet.

In dieser Vorlesung ist die Menge der möglichen Fälle  $\Omega$  abzählbar. Solche Zufallsvorgänge werden **diskret** genannt.

## 1.1 Ereignisse und ihre Wahrscheinlichkeit

### Ereignisse als Mengen

Sei  $\Omega$  die Menge der möglichen Fälle und  $A \subseteq \Omega$  ein Ereignis. Als Notationen für die Menge  $A$  werden wir auch verwenden:

$$A = \{\omega \in \Omega \mid \omega \in A\} = \{\omega \in A\} = \{\text{»}A \text{ tritt ein«}\}.$$

Wir wollen nun **Kombinationen von Ereignissen** betrachten.

Seien  $A, B, A_i, i \in I$ , Ereignisse. Was bedeuten Ereignisse wie  $A^C, A \cup B, \bigcap_{i \in I} A_i$  anschaulich?

Um dies herauszufinden, betrachten wir einen möglichen Fall  $\omega$  und untersuchen, wann dieser eintritt:

- $A \cup B$ :

$$\begin{aligned} \omega \in A \cup B &\Leftrightarrow \omega \in A \text{ oder } \omega \in B, \\ \text{»}A \cup B \text{ tritt ein«} &\Leftrightarrow \text{»}A \text{ tritt ein oder } B \text{ tritt ein«}. \end{aligned}$$

- $\bigcup_{i \in I} A_i$ :

$$\begin{aligned}\omega \in \bigcup_{i \in I} A_i &\Leftrightarrow \text{es gibt ein } i \in I \text{ mit } \omega \in A_i. \\ \text{»}\bigcup_{i \in I} A_i \text{ tritt ein«} &\Leftrightarrow \text{»mindestens eines der Ereignisse } A_i \text{ tritt ein«.}\end{aligned}$$

• WEITERE BEISPIELE:

$$\begin{aligned}A \cap B &\Leftrightarrow \text{»}A \text{ und } B \text{ treten ein«}, \\ \bigcap_{i \in I} A_i &\Leftrightarrow \text{»jedes der } A_i \text{ tritt ein«}, \\ A^C = \Omega &\Leftrightarrow \text{»}A \text{ tritt nicht ein«}, \\ A = \emptyset &\Leftrightarrow \text{»unmögliches Ereignis« (tritt nie ein),} \\ A = \Omega &\Leftrightarrow \text{»sicheres Ereignis« (tritt immer ein),} \\ A = \{\omega\} &\Leftrightarrow \text{»Elementarereignis« (tritt nur im Fall } \omega \text{ ein).}\end{aligned}$$

Sei  $\mathcal{A}$  die Kollektion aller im Modell zugelassenen bzw. in Betracht gezogenen Ereignisse.  $\mathcal{A}$  besteht aus Teilmengen von  $\Omega$ , d.h.

$$\begin{aligned}\mathcal{A} &\subseteq \mathcal{P}(\Omega), \quad \text{wobei} \\ \mathcal{P}(\Omega) &:= \{A \mid A \subseteq \Omega\}\end{aligned}$$

die Potenzmenge von  $\Omega$ , d.h. die Menge aller Teilmengen von  $\Omega$  bezeichnet. Die Kollektion  $\mathcal{A}$  sollte unter den obigen Mengenoperationen, also abzählbaren Vereinigungen, Durchschnitten und Komplementbildung abgeschlossen sein. Wir fordern daher:

**Axiom.**  $\mathcal{A} \subseteq \mathcal{P}(\Omega)$  ist eine  $\sigma$ -Algebra, d.h.

- (i)  $\Omega \in \mathcal{A}$ ,
- (ii) Für alle  $A \in \mathcal{A}$  gilt:  $A^C \in \mathcal{A}$ ,
- (iii) Für  $A_1, A_2, \dots \in \mathcal{A}$  gilt:  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ .

**Bemerkung.** Für  $\sigma$ -Algebren gilt auch:

- a) Nach (i) und (ii) ist  $\emptyset = \Omega^C \in \mathcal{A}$ .
- b) Sind  $A, B \in \mathcal{A}$ , so gilt nach (iii) und a):  $A \cup B = A \cup B \cup \emptyset \cup \emptyset \cup \dots \in \mathcal{A}$ .
- c) Sind  $A_1, A_2, \dots \in \mathcal{A}$ , so ist nach (ii) und (iii):  $\bigcap_{i=1}^{\infty} A_i = (\bigcup_{i=1}^{\infty} A_i^C)^C \in \mathcal{A}$ .

**Beispiel.** Die Potenzmenge  $\mathcal{A} = \mathcal{P}(\Omega)$  ist eine  $\sigma$ -Algebra.

Üblicherweise verwendet man  $\mathcal{A} = \mathcal{P}(\Omega)$  bei **diskreten** Modellen, d.h. für abzählbare  $\Omega$ . Bei nichtdiskreten Modellen kann man **nicht** jede Wahrscheinlichkeitsverteilung  $P$  auf einer  $\sigma$ -Algebra  $\mathcal{A} \subset \mathcal{P}(\Omega)$  zu einer Wahrscheinlichkeitsverteilung auf  $\mathcal{P}(\Omega)$  erweitern (siehe »Einführung in die Wahrscheinlichkeitstheorie«).

## Wahrscheinlichkeitsverteilungen

Sei  $\Omega$  eine nichtleere Menge und  $\mathcal{A} \subseteq \mathcal{P}(\Omega)$  eine  $\sigma$ -Algebra. Wir wollen nun Ereignissen  $A \in \mathcal{A}$  eine Wahrscheinlichkeit  $P[A]$  zuordnen. Für Ereignisse  $A, B \in \mathcal{A}$  gilt:

$$A \cup B \text{ tritt ein} \Leftrightarrow A \text{ oder } B \text{ tritt ein.}$$

Angenommen,  $A$  und  $B$  **treten nicht gleichzeitig ein**, d.h.

$$A \cap B = \emptyset, \quad (A \text{ und } B \text{ sind »disjunkt«}).$$

Dann sollte »endliche Additivität« gelten:

$$P[A \cup B] = P[A] + P[B].$$

**Axiom.** Eine Abbildung

$$P: \mathcal{A} \rightarrow [0, \infty]$$

$$A \mapsto P[A]$$

ist eine **Wahrscheinlichkeitsverteilung** auf  $(\Omega, \mathcal{A})$ , wenn gilt:

(i)  $P$  ist » **$\sigma$ -additiv**«, d.h. für Ereignisse  $A_1, A_2, \dots \in \mathcal{A}$  mit  $A_i \cap A_j = \emptyset$  für  $i \neq j$  gilt:

$$P\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} P[A_i].$$

(ii)  $P$  ist »**normiert**«, d.h.

$$P[\Omega] = 1.$$

Ein **Wahrscheinlichkeitsraum**  $(\Omega, \mathcal{A}, P)$  besteht aus einer Menge  $\Omega$ , einer  $\sigma$ -Algebra  $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ , und einer Wahrscheinlichkeitsverteilung  $P$  auf  $(\Omega, \mathcal{A})$ .

**Satz 1.1** (Elementare Rechenregeln). Sei  $(\Omega, \mathcal{A}, P)$  ein Wahrscheinlichkeitsraum.

i) Es gilt  $P[\emptyset] = 0$ ,

ii) Für  $A, B \in \mathcal{A}$  mit  $A \cap B = \emptyset$  gilt **endliche Additivität**:

$$P[A \cup B] = P[A] + P[B].$$

iii) Für  $A, B \in \mathcal{A}$  mit  $A \subseteq B$  gilt:

$$P[B] = P[A] + P[B \setminus A].$$

Insbesondere gilt:

$$\begin{aligned} P[A] &\leq P[B], && \text{»Monotonie«,} \\ P[A^C] &= 1 - P[A], && \text{»Gegenereignis«,} \\ P[A] &\leq 1. \end{aligned}$$

iv) Für  $A, B \in \mathcal{A}$  gilt:

$$P[A \cup B] = P[A] + P[B] - P[A \cap B] \leq P[A] + P[B].$$

*Beweis.* i) Wegen der  $\sigma$ -Additivität von  $P$  gilt

$$1 = P[\Omega] = P[\Omega \cup \emptyset \cup \emptyset \cup \dots] = \underbrace{P[\Omega]}_{=1} + \underbrace{P[\emptyset]}_{\geq 0} + \underbrace{P[\emptyset]}_{\geq 0} + \dots,$$

und damit

$$P[\emptyset] = 0.$$

ii) Für disjunkte Ereignisse  $A, B$  folgt aus der  $\sigma$ -Additivität und mit i):

$$\begin{aligned} P[A \cup B] &= P[A \cup B \cup \emptyset \cup \emptyset \cup \dots] \\ &= P[A] + P[B] + P[\emptyset] + \dots \\ &= P[A] + P[B]. \end{aligned}$$

iii) Falls  $A \subseteq B$ , ist  $B = A \cup (B \setminus A)$ . Da diese Vereinigung disjunkt ist, folgt mit ii):

$$P[B] = P[A] + P[B \setminus A] \geq P[A].$$

Insbesondere ist  $1 = P[\Omega] = P[A] + P[A^C]$  und somit  $P[A] \leq 1$ .

iv) Nach iii) gilt:

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

□

Aussage iv) des Satzes lässt sich für endlich viele Ereignisse verallgemeinern. Nach iv) gilt für die Vereinigung von drei Ereignissen:

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

Mit vollständiger Induktion folgt:

**Korollar** (Einschluss-/Ausschlussprinzip). Für  $n \in \mathbb{N}$  mit Ereignissen  $A_1, \dots, A_n \in \mathcal{A}$  gilt:

$$P[\underbrace{A_1 \cup A_2 \cup \dots \cup A_n}_{\text{»eines der } A_i \text{ tritt ein«}}] = \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} P[\underbrace{A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}}_{\text{»}A_{i_1}, A_{i_2}, \dots \text{ und } A_{i_k} \text{ treten ein«}}].$$

Das Einschluss-/Ausschlussprinzip werden wir auf eine elegantere Weise am Ende dieses Kapitels beweisen (siehe Satz 1.9).

## Diskrete Wahrscheinlichkeitsverteilungen

Als Beispiel für eine diskrete Wahrscheinlichkeitsverteilung haben wir den Münzwurf betrachtet:

$$\begin{aligned} \Omega &= \{0, 1\}, & \mathcal{A} &= \{\{\emptyset\}, \{0\}, \{1\}, \{0, 1\}\}, \\ P[\{1\}] &= p, & P[\emptyset] &= 0, \\ P[\{0\}] &= 1 - p, & P[\Omega] &= 1. \end{aligned}$$

ALLGEMEIN: Ist die Menge der möglichen Fälle  $\Omega$  endlich oder abzählbar unendlich, dann setzen wir als zugehörige  $\sigma$ -Algebra  $\mathcal{A} = \mathcal{P}[\Omega]$ .

**Satz 1.2.** i) Sei  $0 \leq p(\omega) \leq 1$ ,  $\sum_{\omega \in \Omega} p(\omega) = 1$  eine **Gewichtung der möglichen Fälle**. Dann ist durch

$$P[A] := \sum_{\omega \in A} p(\omega), \quad (A \subseteq \Omega),$$

eine **Wahrscheinlichkeitsverteilung** auf  $(\Omega, \mathcal{A})$  definiert.

ii) Umgekehrt ist jede Wahrscheinlichkeitsverteilung  $P$  auf  $(\Omega, \mathcal{A})$  von dieser Form mit

$$p(\omega) = P[\{\omega\}] \quad (\omega \in \Omega).$$

$p: \Omega \rightarrow [0, 1]$  heißt **Massenfunktion** (»probability mass function«) der diskreten Wahrscheinlichkeitsverteilung  $P$ .

Für den Beweis des Satzes brauchen wir einige Vorbereitungen.

**Bemerkung** (Vorbemerkung zu Summen mit positiven Summanden). Sei  $A$  eine abzählbare Menge,  $p(\omega) \geq 0$  für alle  $\omega \in A$ . Dann definieren wir

$$\sum_{\omega \in A} p(\omega) := \sum_{i=1}^{\infty} p(\omega_i),$$

wobei  $\omega_1, \omega_2, \dots$  eine beliebige Abzählung von  $A$  ist.

**Lemma 1.3.** i)  $\sum_{\omega \in A} p(\omega) \in [0, \infty]$  und ist wohldefiniert (d.h. unabhängig von der Abzählung). Es gilt:

$$\sum_{\omega \in A} p(\omega) = \sup_{\substack{F \subseteq A \\ |F| < \infty}} \sum_{\omega \in F} p(\omega). \quad (1.1.1)$$

Insbesondere gilt **Monotonie**:

$$\sum_{\omega \in A} p(\omega) \leq \sum_{\omega \in B} p(\omega), \quad (A \subseteq B). \quad (1.1.2)$$

ii) Ist  $A = \bigcup_{i=1}^{\infty} A_i$  eine disjunkte Zerlegung, dann gilt:

$$\sum_{\omega \in A} p(\omega) = \sum_{i=1}^{\infty} \sum_{\omega \in A_i} p(\omega).$$

*Beweis.* i) Sei  $\omega_1, \omega_2, \dots$  eine beliebige Abzählung von  $A$ . Aus  $p(\omega_i) \geq 0$  für alle  $i \in \mathbb{N}$  folgt, dass die Partialsummen  $\sum_{i=1}^n p(\omega_i)$  monoton wachsend sind. Daraus folgt:

$$\sum_{i=1}^{\infty} p(\omega_i) = \sup_{n \in \mathbb{N}} \sum_{i=1}^n p(\omega_i).$$

Falls die Menge der Partialsummen von oben beschränkt ist, existiert dieses Supremum in  $[0, \infty)$ . Andernfalls divergiert die Folge der Partialsummen bestimmt gegen  $+\infty$ . Zu zeigen bleibt:

$$\sup_{n \in \mathbb{N}} \sum_{i=1}^n p(\omega_i) = \sup_{\substack{F \subseteq A \\ |F| < \infty}} \sum_{\omega \in F} p(\omega) \quad \text{ist unabhängig von der Abzählung von } A.$$

»≤«: Für alle  $n \in \mathbb{N}$  gilt:

$$\sum_{i=1}^n p(\omega_i) \leq \sup_{\substack{F \subseteq A \\ |F| < \infty}} \sum_{\omega \in F} p(\omega),$$

da das Supremum auch über  $F = \{\omega_1, \dots, \omega_n\}$  gebildet wird. Damit folgt »≤«.

»≥«: Da  $F \subseteq A$  endlich ist, gibt es ein  $n \in \mathbb{N}$ , so dass  $F \subseteq \{\omega_1, \dots, \omega_n\}$ . Also gilt:

$$\sum_{\omega \in F} p(\omega) \leq \sum_{i=1}^n p(\omega_i) \leq \sum_{i=1}^{\infty} p(\omega_i).$$

Damit folgt »≥«.

- ii) • Falls  $A$  endlich ist, gilt  $A_i \neq \emptyset$  nur für endlich viele  $i \in \mathbb{N}$  und alle  $A_i$  sind endlich. Die Behauptung folgt dann aus dem Kommutativ- und dem Assoziativgesetz.
- Sei andernfalls  $A$  abzählbar unendlich.

»≤«: Da  $F \subseteq A$  endlich, ist  $F = \bigcup_{i \in \mathbb{N}} F \cap A_i$ . Da diese Vereinigung disjunkt ist, folgt mit  $\sigma$ -Additivität und Gleichung (1.1.2):

$$P[F] = \sum_{i \in \mathbb{N}} P[F \cap A_i] \leq \sum_{i \in \mathbb{N}} P[A_i].$$

Mit (i) gilt auch:

$$P[A] = \sup_{\substack{F \subseteq A \\ |F| < \infty}} P[F] \leq \sum_{i \in \mathbb{N}} P[A_i].$$

Damit folgt »≤«.

»≥«: Seien  $F_i \subseteq A_i$  endlich. Da die  $F_i$  disjunkt sind, folgt mit  $\sigma$ -Additivität und Gleichung (1.1.2) für alle  $n \in \mathbb{N}$ :

$$\sum_{i=1}^n P[F_i] = P\left[\bigcup_{i=1}^n F_i\right] \leq P\left[\bigcup_{i=1}^{\infty} A_i\right] = P[A].$$

Mit (1.1.1) folgt

$$\sum_{i=1}^n P[A_i] \leq P[A],$$

und für  $n \rightarrow \infty$  schließlich

$$\sum_{i=1}^{\infty} P[A_i] \leq P[A].$$

Damit folgt »≥«.

□

*Beweis von Satz 1.2.* i) Es ist  $P[\Omega] = \sum_{\omega \in \Omega} p(\omega) = 1$  nach Voraussetzung.

Seien  $A_i$ , ( $i \in \mathbb{N}$ ) disjunkt und  $A := \bigcup_{i=1}^{\infty} A_i$ . Die  $\sigma$ -Additivität von  $P$  folgt aus Lemma 1.3.ii):

$$P\left[\bigcup_{i=1}^{\infty} A_i\right] = P[A] = \sum_{\omega \in A} p(\omega) = \sum_{i=1}^{\infty} \sum_{\omega \in A_i} p(\omega) = \sum_{i=1}^{\infty} P[A_i]$$



ii) Aus der  $\sigma$ -Additivität von  $P$  folgt:

$$P[A] = P\left[\underbrace{\bigcup_{\omega \in A} \{\omega\}}_{\text{disjunkt}}\right] = \sum_{\omega \in A} P[\{\omega\}].$$

□

## Spezielle Wahrscheinlichkeitsverteilungen

### Gleichverteilungen / Laplace-Modelle

Sei  $\Omega$  endlich und nichtleer,  $\mathcal{A} = \mathcal{P}(\Omega)$  und  $p(\omega) = \frac{1}{|\Omega|}$  für alle  $\omega \in \Omega$ . Dann ist

$$P[A] = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl »günstiger« Fälle}}{\text{Anzahl aller Fälle}}, \quad (A \subseteq \Omega),$$

die Wahrscheinlichkeitsverteilung zu  $p$  und wird **Gleichverteilung** genannt.

**Beispiele.** a)  $n$  FAIRE MÜNZWÜRFE:

Sei  $\Omega = \{0, 1\}^n$  und  $P$  die Gleichverteilung. Dann ist

$$p(\omega) = \frac{1}{2^n}.$$

b) ZUFÄLLIGE PERMUTATIONEN:

Sei  $\Omega = \mathcal{S}_n = \{\omega : \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, n\} \text{ bijektive Abbildungen}\}$  und  $P$  die Gleichverteilung. Dann ist

$$P[A] = \frac{|A|}{n!}.$$

Beispiele für zufällige Permutationen sind das Mischen eines Kartenspiels, Vertauschen von Hüten oder Umzug in die LWK, wobei  $n$  Schlüssel zufällig vertauscht werden. Es gilt:

$$P[\text{»der } k\text{-te Schlüssel passt auf Schloss } i\text{«}] = P[\{\omega \in \mathcal{S}_n \mid \omega(i) = k\}] = \frac{(n-1)!}{n!} = \frac{1}{n}.$$

Wie groß ist die Wahrscheinlichkeit, dass einer der Schlüssel sofort passt?

Das Ereignis »Schlüssel  $i$  passt« ist  $A_i = \{\omega \mid \omega(i) = i\} = \{\text{»}i \text{ ist Fixpunkt«}\}$ . Die Wahrscheinlichkeit für das Ereignis »ein Schlüssel passt« ist nach dem Einschluss-/Ausschlussprinzip (Satz 1.9):

$$\begin{aligned} P[\text{»es gibt mindestens einen Fixpunkt«}] &= P[A_1 \cup A_2 \cup \dots \cup A_n] \\ &= \sum_{k=1}^n (-1)^{k+1} \sum_{1 \leq i_1 < i_2 < \dots < i_k \leq n} P[A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}] \\ &= \sum_{k=1}^n (-1)^{k+1} \sum_{1 \leq i_1 < i_2 < \dots < i_k \leq n} \frac{(n-k)!}{n!}, \end{aligned}$$

wobei die innere Summe über alle  $k$ -elementigen Teilmengen läuft. Es folgt:

$$\begin{aligned} &= \sum_{k=1}^n (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} \\ &= - \sum_{k=1}^n \frac{(-1)^k}{k!} \end{aligned}$$

Für das Gegenereignis erhalten wir:

$$\begin{aligned} P[\text{»kein Schlüssel passt«}] &= P[\text{»kein Fixpunkt«}] - P[\text{»mindestens ein Fixpunkt«}] \\ &= 1 + \sum_{k=1}^n \frac{(-1)^k}{k!} \\ &= \sum_{k=0}^n \frac{(-1)^k}{k!}. \end{aligned}$$

Die letzte Summe konvergiert für  $n \rightarrow \infty$  gegen  $e^{-1}$ . Der Grenzwert existiert also und ist weder 0 noch 1! Die Wahrscheinlichkeit hängt für große  $n$  nur wenig von  $n$  ab.

## Empirische Verteilungen

Seien  $x_1, x_2, \dots, x_n \in \Omega$  Beobachtungsdaten oder Merkmalsausprägungen, zum Beispiel das Alter aller Einwohner von Bonn. Sei

$$\begin{aligned} N[A] &:= |\{i \in \{1, \dots, n\} \mid x_i \in A\}|, && \text{die Anzahl bzw. Häufigkeit der Werte in } A, \text{ und} \\ P[A] &:= \frac{N[A]}{n}, && \text{die relative Häufigkeit der Werte in } A. \end{aligned}$$

Dann ist  $P$  eine Wahrscheinlichkeitsverteilung auf  $(\Omega, \mathcal{P}(\Omega))$  mit Massenfunktion

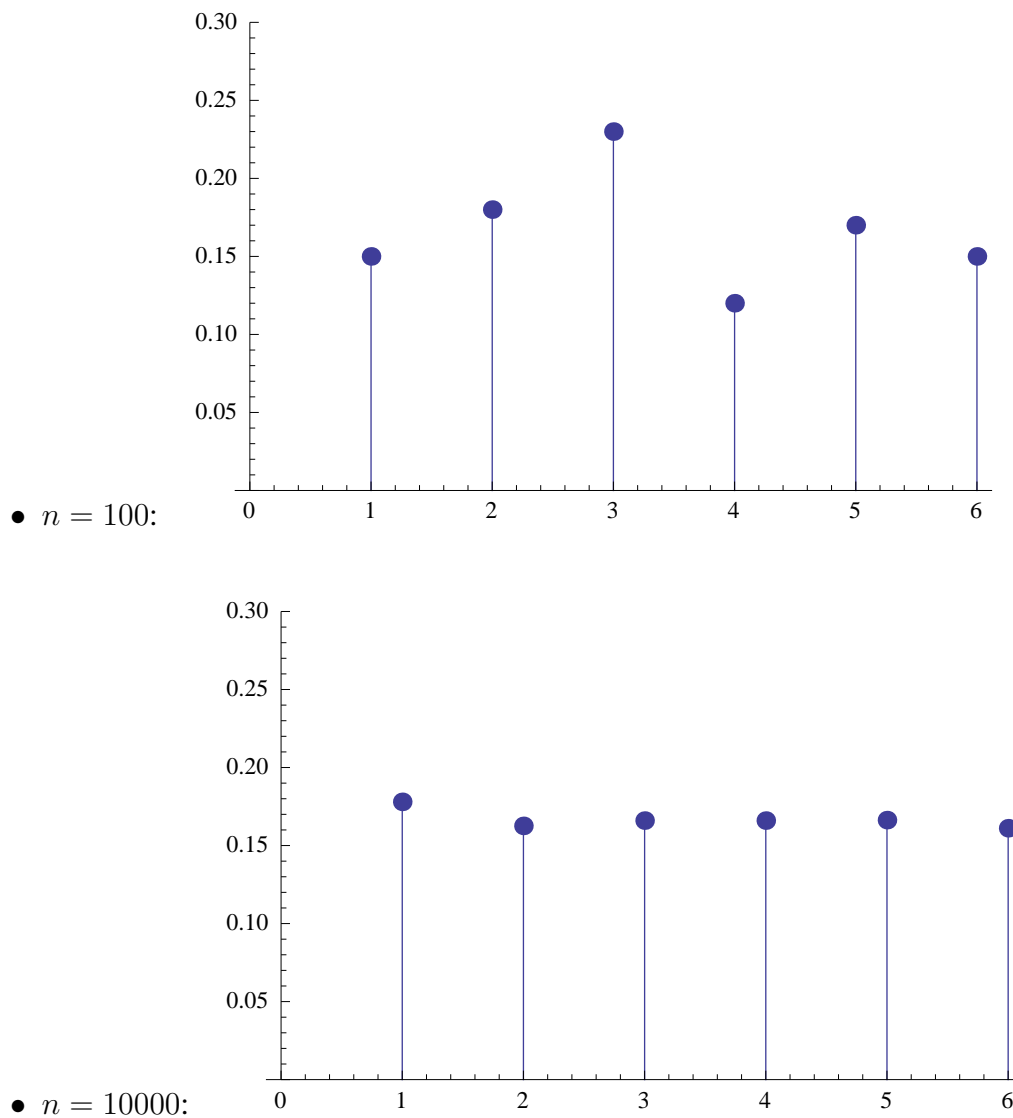
$$p(\omega) = \frac{N[\{\omega\}]}{n}, \quad \text{der relativen Häufigkeit der Merkmalsausprägungen.}$$

**Beispiele.** a) ABZÄHLUNG ALLER MÖGLICHEN FÄLLE:

Sei  $x_1, \dots, x_n$  eine Abzählung der Elemente in  $\Omega$ . Dann stimmt die empirische Verteilung mit der Gleichverteilung überein.

b) EMPIRISCHE VERTEILUNG VON  $n$  ZUFALLSZAHLEN AUS  $\{1, 2, 3, 4, 5, 6\}$ :

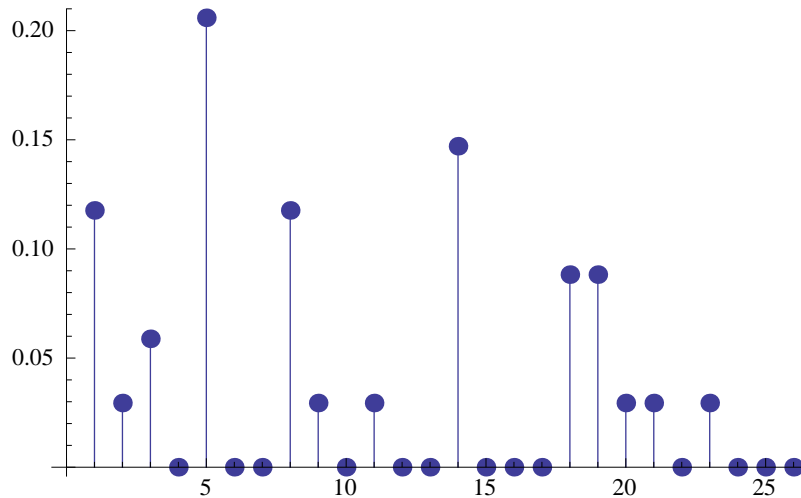
```
x=RandomChoice[{1,2,3,4,5,6},n];
ListPlot[BinCounts[x[[1 ;; n]],{1,7,1}]/n,
  Filling -> Axis, PlotRange -> {0,0.3},
  PlotStyle -> PointSize[Large]],{n,1,100,1}
```



## c) EMPIRISCHE VERTEILUNG DER BUCHSTABEN »A« BIS »Z«:

- in dem Wort »Eisenbahnschrankenwaerterhaeuschen«:

```
freq = StringCount["eisenbahnschrankenwaerterhaeuschen", #] & /@
  CharacterRange["a", "z"];
relfreq = freq/Total[freq];
ListPlot[relfreq, Filling -> Axis, PlotStyle -> PointSize[Large]]
```

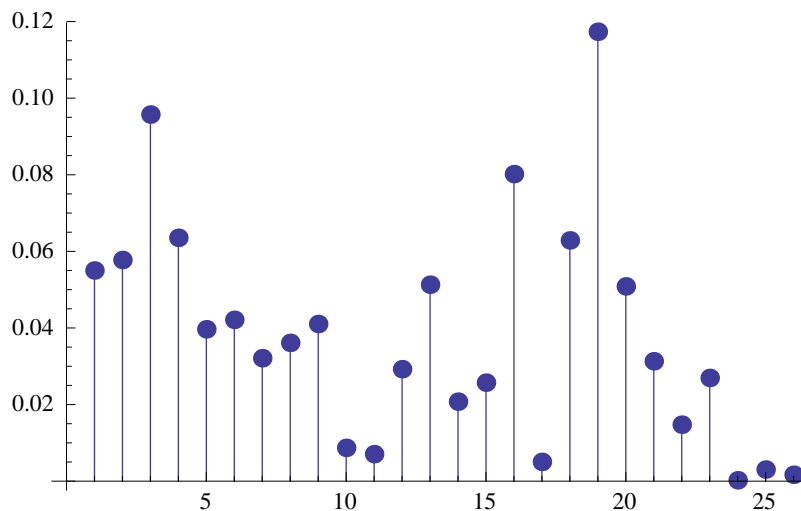


- in einem englischen Wörterbuch:

```

freq = Length[DictionaryLookup[# ~~ ___]] & /@
      CharacterRange["a", "z"];
relfreq = freq/Total[freq];
ListPlot[relfreq, Filling -> Axis, PlotStyle -> PointSize[Large]]

```



d) BENFORDSCHES GESETZ:

Das Benfordsche Gesetz, auch Newcomb-Benford's Law (NBL) beschreibt eine Gesetzmäßigkeit in der Verteilung der Ziffernstrukturen von Zahlen in empirischen Datensätzen, zum Beispiel ihrer ersten Ziffern. Es lässt sich etwa in Datensätzen über Einwohnerzahlen von Städten, Geldbeträge in der Buchhaltung, Naturkonstanten etc. beobachten. Kurz gefasst besagt es:

**»Je niedriger der zahlenmäßige Wert einer Ziffernsequenz bestimmter Länge an einer bestimmten Stelle einer Zahl ist, umso wahrscheinlicher ist ihr Auftreten. Für die Anfangsziffern in Zahlen des Zehnersystems gilt zum Beispiel: Zahlen mit der Anfangsziffer 1 treten etwa 6,5-mal so häufig auf wie solche mit der Anfangsziffer 9.«**

1881 wurde diese Gesetzmäßigkeit von dem Mathematiker Simon Newcomb entdeckt und im „American Journal of Mathematics“ publiziert. Er soll bemerkt haben, dass in den benutzten Büchern mit Logarithmentafeln, die Seiten mit Tabellen mit Eins als erster Ziffer deutlich schmutziger waren als die anderen Seiten, weil sie offenbar öfter benutzt worden seien. Die Abhandlung Newcombs blieb unbeachtet und war schon in Vergessenheit geraten, als der Physiker Frank Benford (1883–1948) diese Gesetzmäßigkeit wiederentdeckte und darüber 1938 neu publizierte. Seither war diese Gesetzmäßigkeit nach ihm benannt, in neuerer Zeit wird aber durch die Bezeichnung »Newcomb-Benford’s Law« (NBL) dem eigentlichen Urheber wieder Rechnung getragen. Bis vor wenigen Jahren war diese Gesetzmäßigkeit nicht einmal allen Statistikern bekannt. Erst seit der US-amerikanische Mathematiker Theodore Hill versucht hat, die Benford-Verteilung zur Lösung praktischer Probleme nutzbar zu machen, ist ihr Bekanntheitsgrad gewachsen. (Quelle: »Wikipedia«)

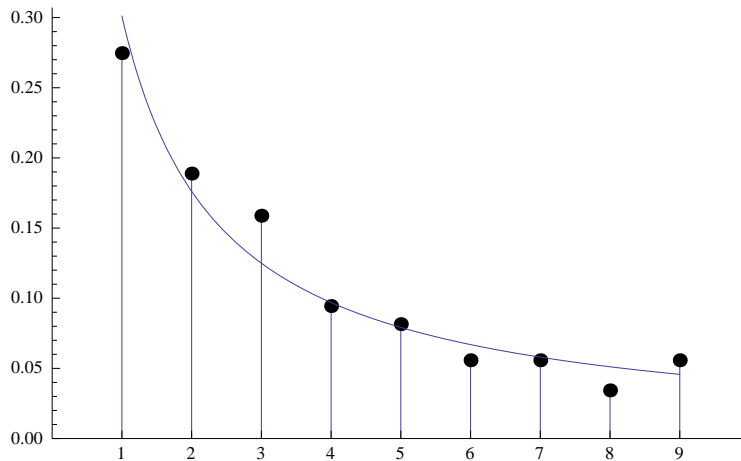
HÄUFIGKEITSVERTEILUNG DER ANFANGSZIFFERN VON ZAHLEN:

Ist  $d$  die erste Ziffer einer Dezimalzahl, so tritt sie nach dem Benfordschen Gesetz in empirischen Datensätzen näherungsweise mit folgenden relativen Häufigkeiten  $p(d)$  auf:

$$p(d) = \log_{10} 1 + \frac{1}{d} = \log_{10} d + 1 - \log_{10} d.$$

In der Grafik unten (Quelle: »Wolfram Demonstrations Project«) werden die relativen Häufigkeiten der Anfangsziffern 1 bis 9 in den Anzahlen der Telefonanschlüsse in allen Ländern der Erde mit den nach dem Benfordschen Gesetz prognostizierten relativen Häu-

figkeiten verglichen.



## 1.2 Diskrete Zufallsvariablen und ihre Verteilung

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

**Definition.** i) Eine *diskrete Zufallsvariable (ZV)* ist eine Abbildung

$$X: \Omega \rightarrow S, \quad S \text{ abzählbar,}$$

so dass für alle  $a \in S$  gilt:

$$X^{-1}(a) := \{\omega \in \Omega \mid X(\omega) = a\} \in \mathcal{A}. \quad (1.2.1)$$

Für  $X^{-1}(a)$  (das Urbild von  $a$  unter  $X$ ) schreiben wir im folgenden  $\{X = a\}$ .

ii) Die **Verteilung** von  $X$  ist die Wahrscheinlichkeitsverteilung  $\mu_X$  auf  $S$  mit Gewichten

$$p_X(a) := P[\{X = a\}], \quad (a \in S).$$

Für  $P[\{X = a\}]$  schreiben wir im folgenden  $P[X = a]$ .

**Bemerkung.** a) In der Tat ist  $p_X$  Massenfunktion einer Wahrscheinlichkeitsverteilung (siehe Satz 1.2):

i) Für alle  $a \in S$  gilt:  $p_X(a) \geq 0$

ii) Da die Ereignisse  $\{X = a\}$  disjunkt sind, folgt:

$$\sum_{a \in S} p_X(a) = \sum_{a \in S} P[X = a] = P\left[\bigcup_{a \in S} \{X = a\}\right] = P[\Omega] = 1.$$

b) Für  $B \subseteq S$  gilt:

$$\{X \in B\} := \underbrace{\{\omega \in \Omega \mid X(\omega) \in B\}}_{X^{-1}(B)} = \bigcup_{a \in B} \underbrace{\{X = a\}}_{\in \mathcal{A}}, \quad \text{sowie}$$

$$P[X \in B] = \sum_{a \in B} P[X = a] = \sum_{a \in B} p_X(a) = \mu_X(B).$$

Die Verteilung  $\mu_X$  gibt also an, mit welchen Wahrscheinlichkeiten die Zufallsvariable  $X$  Werte in bestimmten Mengen annimmt.

**Beispiele** (Zweimal würfeln). Sei  $\Omega = \{\omega = (\omega_1, \omega_2) \mid \omega_i \in \{1, \dots, 6\}\}$  und sei  $P$  die Gleichverteilung.

a)

Sei  $X_i: \Omega \rightarrow S := \{1, 2, 3, 4, 5, 6\}$ ,

$X(\omega) := \omega_i$ , die Augenzahl des  $i$ -ten Wurfs.

$X_i$  ist eine diskrete Zufallsvariable mit Verteilung  $\mu_{X_i}$ . Die Gewichte von  $\mu_{X_i}$  sind

$$p_{X_i}(a) = P[X_i = a] = \frac{6}{36} = \frac{1}{6} \quad \text{für alle } a \in S,$$

d.h.  $X_i$  ist gleichverteilt.

b)

Sei  $Y: \Omega \rightarrow \tilde{S} := \{2, 3, \dots, 12\}$

$Y(\omega) := X_1(\omega) + X_2(\omega)$ , die Summe der Augenzahlen.

Die Gewichte der Verteilung von  $Y$  sind

$$p_Y(a) = P[Y = a] = \begin{cases} \frac{1}{36} & \text{falls } a \in \{2, 12\}, \\ \frac{2}{36} & \text{falls } a \in \{3, 11\}, \\ \dots & \end{cases}$$

d.h.  $Y$  ist nicht mehr gleichverteilt!

Allgemeiner:

**Beispiel.** Sei  $\Omega = \{\omega_1, \dots, \omega_n\}$  endlich,  $P$  die Gleichverteilung,  $X: \Omega \rightarrow S$  eine Zufallsvariable und  $x_i := X(\omega_i)$ . Dann ist

$$P[X = a] = \frac{|\{\omega \in \Omega \mid X(\omega) = a\}|}{|\Omega|} = \frac{|\{1 \leq i \leq n \mid x_i = a\}|}{n},$$

also ist  $\mu_x$  die empirische Verteilung von  $x_1, \dots, x_n$ .

### Binomialverteilung

**Beispiel** (»Ziehen mit Zurücklegen«). Wir betrachten eine endliche Grundgesamtheit (Population, Zustandsraum)  $S$ , zum Beispiel Kugeln in einer Urne, Vögel im Wald, Einwohner in NRW. Wir wollen nun die zufällige Entnahme von  $n$  Einzelstichproben mit Zurücklegen aus  $S$  beschreiben und setzen daher

$$\Omega = S^n = \{\omega = (x_1, \dots, x_n) \mid x_i \in S\}.$$

Wir nehmen an, daß alle kombinierten Stichproben gleich wahrscheinlich sind, d.h.  $P$  sei die Gleichverteilung auf  $\Omega$ .

RELEVANTE ZUFALLSVARIABLEN UND EREIGNISSE:

- $i$ -ter Stichprobenwert:

$$X_i(\omega) = x_i,$$

$$P[X_i = a] = \frac{|S|^{n-1}}{|\Omega|} = \frac{|S|^{n-1}}{|S|^n} = \frac{1}{|S|}, \quad \text{für alle } a \in S,$$

d.h.  $X_i$  ist gleichverteilt auf  $S$ .

Sei  $E \subseteq S$  eine bestimmte Merkmalsausprägung der Stichprobe, die wir im folgenden als »Erfolg« bezeichnen (zum Beispiel schwarze Kugel, Beobachtung einer Amsel). Dann können wir die Ereignisse

$$\{X_i \in E\}, \text{ »Erfolg bei } i\text{-ter Stichprobe«},$$

betrachten. Es gilt:

$$P[X_i \in E] = \mu_{X_i}(E) = \frac{|E|}{|S|}.$$

Wir setzen

$$q := \frac{|E|}{|S|}, \quad \text{»Erfolgswahrscheinlichkeit«}$$

- Häufigkeit von  $E$  / »Anzahl der Erfolge«:

Sei nun

$$N: \Omega \rightarrow \{0, 1, 2, \dots, n\},$$

$$N(\omega) := |\{1 \leq i \leq n \mid X_i(\omega) \in E\}|$$

die Anzahl der Einzelstichproben mit Merkmalsausprägung  $E$ .



**Lemma 1.4.** Für  $k \in \{0, 1, \dots, n\}$  gilt:

$$P[N = k] = \binom{n}{k} p^k (1 - p)^{n-k}.$$

*Beweis.* Es gilt

$$|\{\omega \in \Omega \mid N(\omega) = k\}| = \binom{n}{k} |E|^k |S \setminus E|^{n-k},$$

wobei

$$\binom{n}{k} = \text{Anzahl der Möglichkeiten } k \text{ Indizes aus } \{1, \dots, n\} \text{ auszuwählen,}$$

für die ein Erfolg eintritt,

$$|E|^k = \text{Anzahl der Möglichkeiten für jeden Erfolg,}$$

$$|S \setminus E|^{n-k} = \text{Anzahl der Möglichkeiten für jeden Mißerfolg.}$$

Also gilt:

$$\begin{aligned} P[N = k] &= \frac{\binom{n}{k} |E|^k |S \setminus E|^{n-k}}{|S|^n} = \binom{n}{k} \left(\frac{|E|}{|S|}\right)^k \left(\frac{|S \setminus E|}{|S|}\right)^{n-k} \\ &= \binom{n}{k} p^k (1 - p)^{n-k}. \end{aligned}$$

□

**Definition.** Sei  $n \in \mathbb{N}$  und  $p \in [0, 1]$ . Die Wahrscheinlichkeitsverteilung auf  $\{0, 1, \dots, n\}$  mit Massenfunktion

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

heißt **Binomialverteilung mit Parametern  $n$  und  $p$**  (kurz:  $\text{Bin}(n, p)$ ).

**Bemerkung.** Daß  $p_{n,p}$  eine Massenfunktion einer Wahrscheinlichkeitsverteilung ist, folgt aus Lemma 1.3!

**Bemerkung.** Ereignisse  $E_1, \dots, E_n$  heißen **unabhängig**, falls

$$P[E_{i_1} \cap E_{i_2} \cap \dots \cap E_{i_k}] = P[E_{i_1}] \cdot P[E_{i_2}] \cdot \dots \cdot P[E_{i_k}]$$

für alle  $k \leq n$  und  $1 \leq i_1 < i_2 < \dots < i_k \leq n$  gilt.

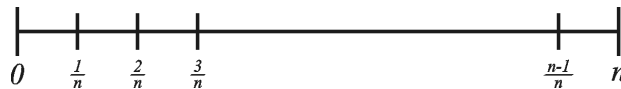
Sind  $E_1, \dots, E_n$  unabhängig und  $P[E_i] = p$ , dann ist

$$P[\text{»genau } k \text{ der } E_i \text{ treten ein«}] = \binom{n}{k} p^k (1 - p)^{n-k},$$

d.h. die Anzahl der Ereignisse, die eintreten, ist binomialverteilt. Der Beweis folgt weiter unten.

### Poissonverteilung

**Beispiel** (Warteschlange). Angenommen, die Kunden in einer Warteschlange kommen **unabhängig voneinander** zu **zufälligen** (gleichverteilten) Zeitpunkten. Wie viele Kunden kommen in einer Zeitspanne der Länge  $t_0$  an? Sei  $N$  die Anzahl dieser Kunden und  $t_0 = 1$ . Wir unterteilen das Intervall  $[0, 1]$ :



Wir machen die folgende Annahme (die natürlich in zu modellierenden Anwendungsproblemen zu überprüfen ist):

»Wenn  $n$  sehr groß ist, dann kommt in einer Zeitspanne der Länge  $\frac{1}{n}$  fast immer höchstens ein Kunde«.

$E_i$  stehe für das Ereignis, daß ein Kunde im Zeitintervall  $[\frac{i-1}{n}, \frac{i}{n}]$  ankommt ( $1 \leq i \leq n$ ). Wir nehmen außerdem an, daß die Wahrscheinlichkeit unabhängig von  $i$  und näherungsweise proportional zu  $\frac{1}{n}$  ist, also:

$$P[E_i] \approx \frac{\lambda}{n}, \quad \lambda \in (0, \infty).$$

Für das Ereignis, daß genau  $k$  Kunden im Zeitintervall  $[0, 1]$  ankommen, sollte dann gelten, dass

$$P[N = k] \approx P[\text{»genau } k \text{ der } E_i \text{ treten ein«}] \approx p_{n, \frac{\lambda}{n}}(k),$$

wobei  $p_{n, \frac{\lambda}{n}}(k)$  das Gewicht von  $k$  unter der Binomialverteilung mit Parametern  $n$  und  $\frac{\lambda}{n}$  ist. Diese Näherung sollte **»für große  $n$  immer genauer werden«**.

**Satz 1.5** (Poissonapproximation der Binomialverteilung). Sei  $\lambda \in (0, \infty)$ . Dann gilt:

$$\lim_{n \rightarrow \infty} p_{n, \frac{\lambda}{n}}(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots$$

*Beweis.* Es gilt:

$$\begin{aligned} p_{n, \frac{\lambda}{n}}(k) &= \frac{n!}{k!(n-k)!} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \cdot \underbrace{\frac{n \cdot (n-1) \cdot \dots \cdot (n-k+1)}{n^k}}_{\rightarrow 1} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\rightarrow e^{-\lambda}} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\rightarrow 1} \\ &\rightarrow \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{für } n \rightarrow \infty. \end{aligned}$$

□

**Definition.** Die Wahrscheinlichkeitsverteilung auf  $\{0, 1, 2, \dots\}$  mit Massenfunktion

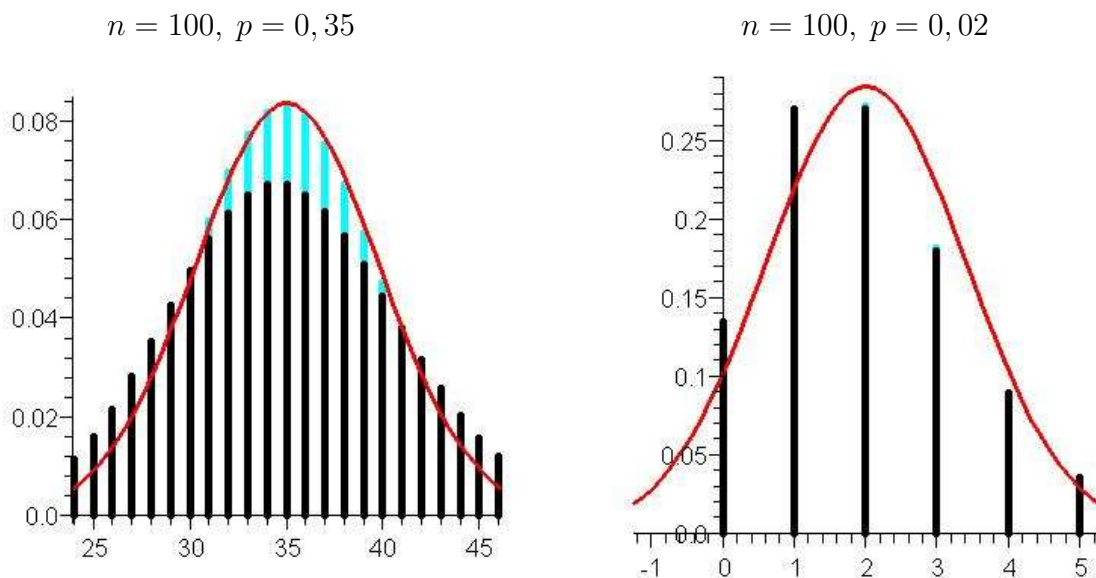
$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots,$$

heißt **Poissonverteilung mit Parameter  $\lambda$** .

Aufgrund des Satzes verwendet man die Poissonverteilung zur näherungsweisen Modellierung der Häufigkeit seltener Ereignisse (zum Beispiel Tippfehler in einem Buch, Schadensfälle bei Versicherung, Zusammenbrüche des T-Mobile-Netzes, ...) und damit zur »Approximation« von Binomialverteilungen mit kleiner Erfolgswahrscheinlichkeit  $p$ .

Für häufigere Ereignisse (zum Beispiel wenn Erfolgswahrscheinlichkeit  $p$  unabhängig von  $n$  ist) verwendet man hingegen besser eine Normalverteilung zur näherungsweisen Modellierung der (geeignet reskalierten) relativen Häufigkeit  $\frac{k}{n}$  des Ereignisses für große  $n$ . Definition und Eigenschaften von Normalverteilungen werden wir später kennenlernen.

Die folgenden (mit »Maple« erstellten) Graphiken zeigen die Poisson- und Normalapproximation (Poisson schwarz, Normalverteilung rot) der Binomialverteilung (blau) für unterschiedliche Parameterwerte:



### Hypergeometrische Verteilung

**Beispiel** (Ziehen ohne Zurücklegen). Wir betrachten  $m$  Kugeln in einer Urne (Wähler, Fische im See, ...), davon  $r$  rote und  $m - r$  schwarze. Gezogen wird eine zufällige Stichprobe von  $n$

Kugeln,  $n \leq \min(r, m - r)$ . Sind alle Stichproben gleich wahrscheinlich, dann ist ein geeignetes Modell gegeben durch:

$$\begin{aligned}\Omega &= \mathcal{P}(\{1, \dots, m\}) = \text{alle Teilmengen von } \{1, \dots, m\} \text{ der Kardinalität } n, \\ P &= \text{Gleichverteilung auf } \Omega.\end{aligned}$$

Wir definieren eine Zufallsvariable  $N: \Omega \rightarrow \{1, \dots, m\}$  durch

$$N(\omega) := \text{Anzahl der roten Kugeln in } \omega.$$

Für das Ereignis, daß genau  $k$  rote Kugeln in der Stichprobe sind, gilt:

$$P[N = k] = \frac{|\{\omega \in \Omega \mid N(\omega) = k\}|}{|\Omega|} = \frac{\binom{r}{k} \cdot \binom{m-r}{n-k}}{\binom{m}{n}}, \quad (k = 0, 1, \dots, n).$$

Diese Wahrscheinlichkeitsverteilung wird **hypergeometrische Verteilung mit Parametern  $m$ ,  $r$  und  $n$**  genannt.

**Bemerkung.** Untersucht man die Asymptotik der hypergeometrischen Verteilung für  $m \rightarrow \infty$ ,  $r \rightarrow \infty$ ,  $p = \frac{r}{m}$  fest und  $n$  fest, so gilt:

$$P[N = k] \longrightarrow \binom{n}{k} p^k (1 - p)^{n-k},$$

d.h. die hypergeometrische Verteilung nähert sich der Binomialverteilung an. Eine anschauliche Erklärung dafür ist:

Befinden sich sehr viele Kugeln in der Urne, dann ist der Unterschied zwischen Ziehen mit und ohne Zurücklegen vernachlässigbar, da nur sehr selten dieselbe Kugel zweimal gezogen wird.

### 1.3 Simulation von Gleichverteilungen

Ein **(Pseudo-) Zufallszahlengenerator** ist ein Algorithmus, der eine deterministische Folge von ganzen Zahlen  $x_1, x_2, x_3, \dots$  mit Werten zwischen 0 und einem Maximalwert  $m - 1$  erzeugt, welche durch eine vorgegebene Klasse statistischer Tests nicht von einer Folge von Stichproben unabhängiger, auf  $\{0, 1, 2, \dots, m - 1\}$  gleichverteilter Zufallsgrößen unterscheidbar ist. Ein Zufallszahlengenerator erzeugt also nicht wirklich zufällige Zahlen. Die von »guten« Zufallszahlengeneratoren erzeugten Zahlen haben aber statistische Eigenschaften, die denen von echten Zufallszahlen in vielerlei (aber nicht in jeder) Hinsicht sehr ähnlich sind.

Konkret werden die Pseudozufallszahlen üblicherweise über eine deterministische Rekurrenzrelation vom Typ

$$x_{n+1} = f(x_{n-k+1}, x_{n-k+2}, \dots, x_n), \quad n = k, k+1, k+2, \dots,$$

aus **Saatwerten**  $x_1, x_2, \dots, x_k$  erzeugt. In vielen Fällen hängt die Funktion  $f$  nur von der letzten erzeugten Zufallszahl  $x_n$  ab. Wir betrachten einige Beispiele:

### Lineare Kongruenzgeneratoren (LCG)

Bei linearen Kongruenzgeneratoren ist die Rekurrenzrelation vom Typ

$$x_{n+1} = (ax_n + c) \bmod m, \quad n = 0, 1, 2, \dots$$

Hierbei sind  $a$ ,  $c$  und  $m$  geeignet zu wählende positive ganze Zahlen, zum Beispiel:

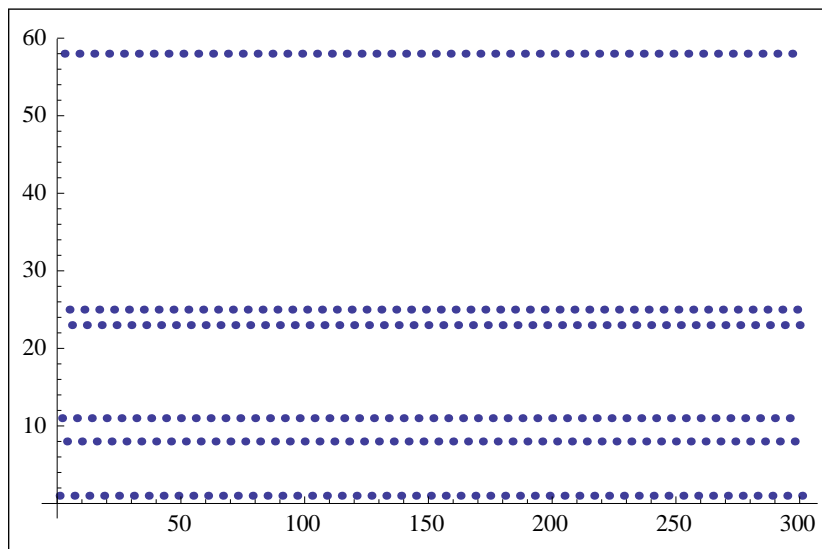
ZX81-Generator :	$m = 2^{16} + 1,$	$a = 75,$	$c = 0.$
RANDU, IBM 360/370 :	$m = 2^{31},$	$a = 65539,$	$c = 0.$
Marsaglia-Generator :	$m = 2^{32},$	$a = 69069,$	$c = 1.$
Langlands-Generator :	$m = 2^{48},$	$a = 142412240584757,$	$c = 11.$

Um einen ersten Eindruck zu erhalten, wie die Qualität der erzeugten Pseudozufallszahlen von  $a$ ,  $c$  und  $m$  abhängt, implementieren wir die Generatoren mit »Mathematica«:

```
f[x_] := Mod[a x + c, m]
```

**Beispiel.** Wir beginnen zur Demonstration mit dem Beispiel eines ganz schlechten LCG:

```
a = 11; c = 0; m = 63;
pseudorandomdata = NestList[f, 1, 300];
ListPlot[pseudorandomdata]
```



Die Folge von Zufallszahlen ist in diesem Fall periodisch mit einer Periode, die viel kleiner ist als die maximal mögliche (63). Dies rechnet man auch leicht nach.

Periodizität mit Periode kleiner als  $m$  kann man leicht ausschließen. Es gilt nämlich:

**Satz (Knuth).** *Die Periode eines LCG ist gleich  $m$  genau dann, wenn*

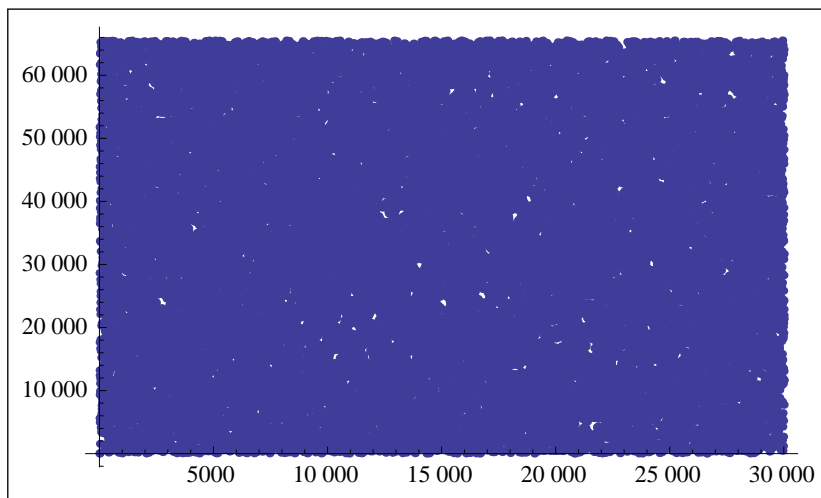
- i)  $c$  und  $m$  teilerfremd sind,*
- ii) jeder Primfaktor von  $m$  ein Teiler von  $a - 1$  ist, und*
- iii) falls 4 ein Teiler von  $m$  ist, dann auch von  $a - 1$ .*

*Beweis.* siehe D. Knuth: »The art of computer programming, Vol. 2.«

□

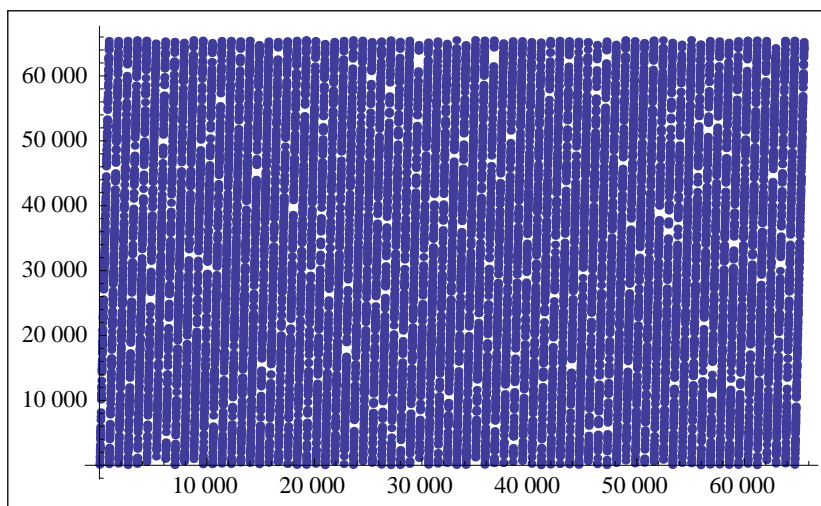
**Beispiel (ZX 81-Generator).** Hier ergibt sich ein besseres Bild, solange wir nur die Verteilung der einzelnen Zufallszahlen betrachten:

```
a = 75; c = 0; m = 2^16 + 1;
pseudorandomdata = NestList[f, 1, 30000];
ListPlot[pseudorandomdata]
```



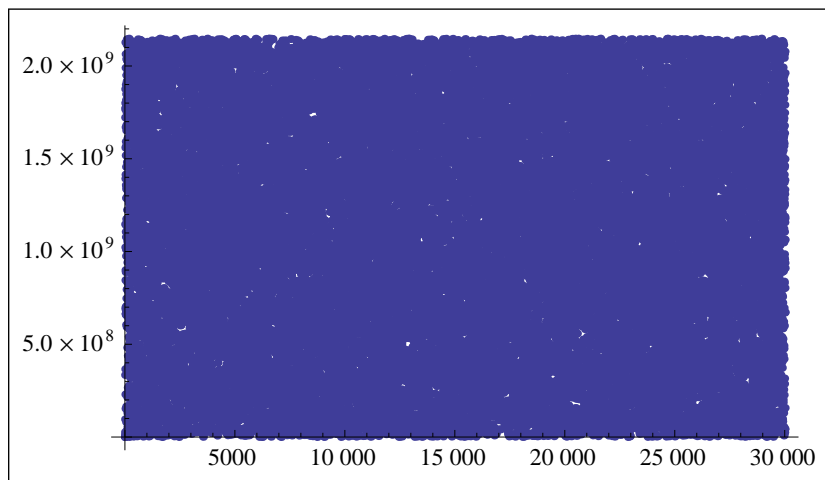
Fassen wir jedoch Paare  $(x_i, x_{i+1})$  von aufeinanderfolgenden Pseudozufallszahlen als Koordinaten eines zweidimensionalen Pseudozufallsvektors auf, und betrachten die empirische Verteilung dieser Vektoren, so ergibt sich keine besonders gute Approximation einer zweidimensionalen Gleichverteilung:

```
blocks = Partition[pseudorandomdata, 2]; ListPlot[blocks]
```

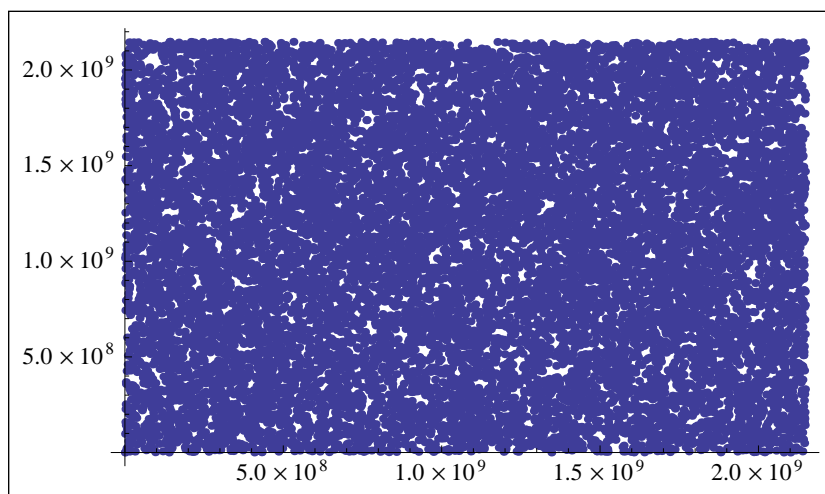


**Beispiel (RANDU).** Hier scheinen sowohl die einzelnen Pseudozufallszahlen  $x_i$  als auch die Vektoren  $(x_i, x_{i+1})$  näherungsweise gleichverteilt zu sein:

```
a = 65539; c = 0; m = 2^31;
pseudorandomdata = NestList[f, 1, 30000];
ListPlot[pseudorandomdata]
```



```
blocks = Partition[pseudorandomdata , 2]; ListPlot[blocks]
```

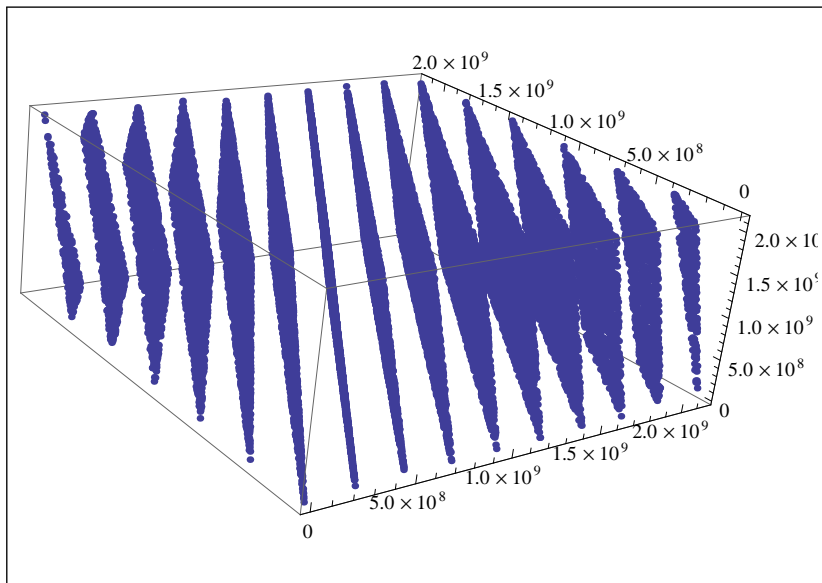


Fassen wir aber jeweils drei aufeinanderfolgende Pseudozufallszahlen als Koordinaten eines Vektors

$(x_i, x_{i+1}, x_{i+2})$  im  $\mathbb{Z}^3$  auf, dann ist die empirische Verteilung dieser Pseudozufallsvektoren keine Gleichverteilung mehr, sondern konzentriert sich auf nur 15 zweidimensionalen Hyperebenen:

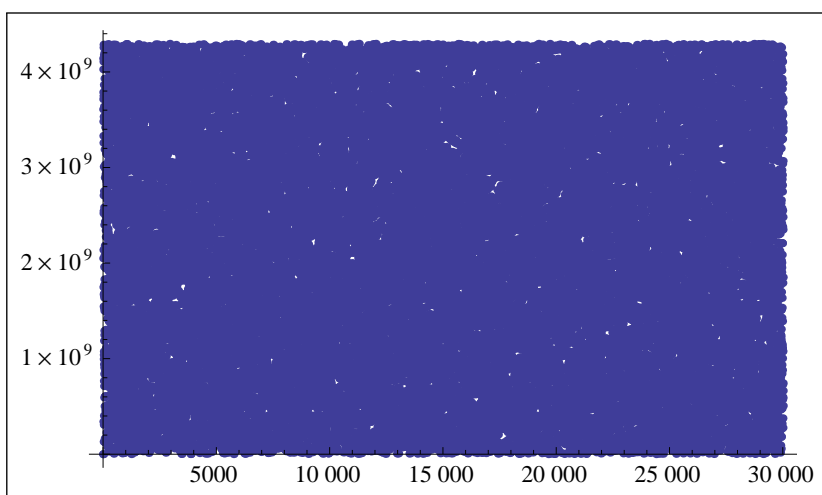
```
blocks3 = Partition[pseudorandomdata , 3]; ListPointPlot3D[blocks3]
```



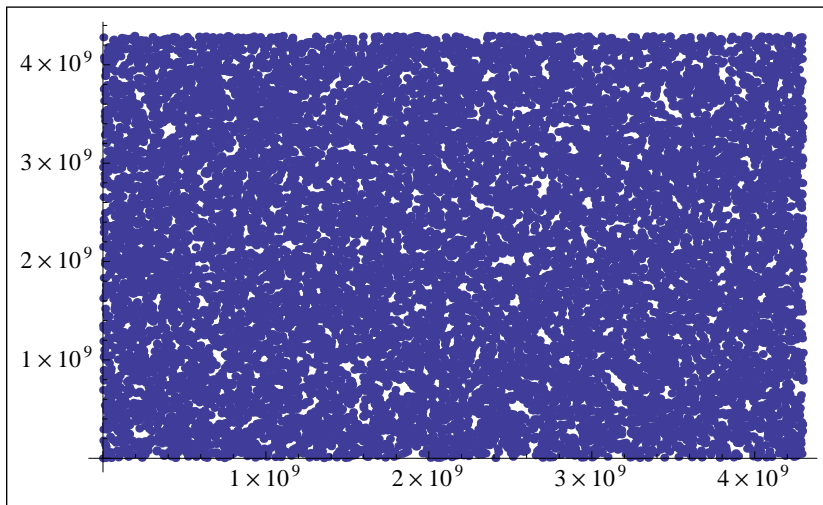


**Beispiel** (Marsaglia-Generator). Der von Marsaglia 1972 vorgeschlagene LCG besteht dagegen alle obigen Tests (und einige weitere):

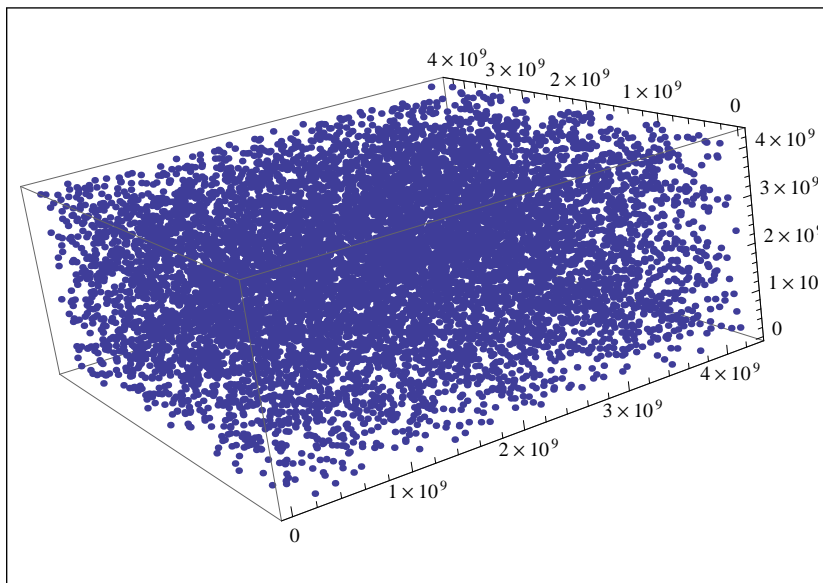
```
a = 60069; c = 1; m = 2^32;
pseudorandomdata = NestList[f, 1, 30000];
ListPlot[pseudorandomdata]
```



```
blocks = Partition[pseudorandomdata, 2];
ListPlot[blocks]
```



```
blocks3 = Partition[pseudorandomdata, 3];
ListPointPlot3D[blocks3]
```



Dies bedeutet natürlich nicht, daß die vom Marsaglia-Generator erzeugte Folge eine für **alle** Zwecke akzeptable Approximation einer Folge von unabhängigen Stichproben von der Gleichverteilung ist. Da die Folge in Wirklichkeit deterministisch ist, kann man einen Test konstruieren, der sie von einer echten Zufallsfolge unterscheidet.

### Shift-Register-Generatoren

Bei Shift-Register-Generatoren interpretiert man eine Zahl  $x_n \in \{0, 1, \dots, 2^k - 1\}$  zunächst als Binärzahl bzw. als Vektor aus  $\{0, 1\}^k$ , und wendet dann eine gegebene Matrix  $T$  darauf an, um  $x_{n+1}$  zu erhalten:

$$x_{n+1} = Tx_n, \quad n = 0, 1, 2, \dots$$

### Kombination von Zufallszahlengeneratoren

Zufallszahlengeneratoren lassen sich kombinieren, zum Beispiel indem man die von mehreren Zufallszahlengeneratoren erzeugten Folgen von Pseudozufallszahlen aus  $\{0, 1, \dots, m - 1\}$  modulo  $m$  addiert. Auf diese Weise erhält man sehr leistungsfähige Zufallszahlengeneratoren, zum Beispiel den Kiss-Generator von Marsaglia, der einen LCG und zwei Shift-Register-Generatoren kombiniert, Periode  $2^{95}$  hat, und umfangreiche statistische Tests besteht.

### Zufallszahlen aus $[0,1)$

Ein Zufallszahlengenerator kann natürlich nicht wirklich reelle Pseudozufallszahlen erzeugen, die die Gleichverteilung auf dem Intervall  $[0, 1)$  simulieren, denn dazu würden unendlich viele »zufällige« Nachkommastellen benötigt. Stattdessen werden üblicherweise (pseudo-)zufällige Dezimalzahlen vom Typ

$$u_n = \frac{x_n}{m}, \quad x_n \in \{0, 1, \dots, m - 1\},$$

erzeugt, wobei  $m$  vorgegeben ist (zum Beispiel Darstellungsgenauigkeit des Computers), und  $x_n$  eine Folge ganzzahliger Pseudozufallszahlen aus  $\{0, 1, \dots, m - 1\}$  ist. In »Mathematica« kann man mit

```
RandomReal [spec, WorkingPrecision → k]
```

pseudozufällige Dezimalzahlen mit einer beliebigen vorgegebenen Anzahl  $k$  von Nachkommastellen erzeugen.

### Zufallspermutationen

Der folgende Algorithmus erzeugt eine (pseudo-)zufällige Permutation aus  $S_n$  :

#### Algorithmus 1.6 (RPERM).

```
rperm [n_] :=
Module[{x = Range[n], k, a}, Beginn mit Liste 1,2,...,n
Do[
k = RandomInteger[{i, n}];
a = x[[i]]; x[[i]] = x[[k]]; x[[k]] = a; (Vertausche x[[i]] und x[[k]])
, {i, n - 1}]; (Schleife, i läuft von 1 bis n - 1)
x (Ausgabe von x) ]
```

```
rperm [17]
```

```
{12, 5, 13, 8, 17, 9, 10, 6, 1, 7, 16, 15, 14, 4, 2, 3, 11}
```

ÜBUNG:

Sei  $\Omega_n = \{1, 2, \dots, n\} \times \{2, 3, \dots, n\} \times \dots \times \{n-1, n\}$ .

- Zeigen Sie, daß die Abbildung  $X(\omega) = \tau_{n-1, \omega_{n-1}} \circ \dots \circ \tau_{2, \omega_2} \circ \tau_{1, \omega_1}$  eine Bijektion von  $\Omega_n$  nach  $S_n$  ist ( $\tau_{i,j}$  bezeichnet die Transposition von  $i$  und  $j$ ).
- Folgern Sie, daß der Algorithmus oben tatsächlich eine Stichprobe einer gleichverteilten Zufallspermutation aus  $S_n$  simuliert.

## 1.4 Erwartungswert

Sei  $(\Omega, \mathcal{A}, P)$  ein Wahrscheinlichkeitsraum,  $S \subseteq \mathbb{R}$  abzählbar und  $X: \Omega \rightarrow S$  eine Zufallsvariable auf  $(\Omega, \mathcal{A}, P)$ .

**Definition.** Der *Erwartungswert* von  $X$  bzgl.  $P$  ist definiert als

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

sofern die Summe auf der rechten Seite wohldefiniert ist (d.h. unabhängig von der Abzählung von  $S$ ).

**Bemerkung.** a) Falls  $X(\omega) \geq 0$  für alle  $\omega \in \Omega$  gilt, sind alle Summanden der Reihe nichtnegativ und der Erwartungswert  $E[X] \in [0, \infty]$  wohldefiniert.

b) Falls die Reihe absolut konvergiert, d.h. falls  $\sum_{a \in S} |a| \cdot P[X = a]$  endlich ist, ist der Erwartungswert  $E[X] \in \mathbb{R}$  wohldefiniert.

$E[X]$  kann als der **Prognosewert** oder (**gewichteter**) **Mittelwert** für  $X(\omega)$  interpretiert werden.

**Beispiel** (Indikatorfunktion eines Ereignisses  $A \in \mathcal{A}$ ). Sei

$$X(\omega) = I_A(\omega) := \begin{cases} 1 & \text{falls } \omega \in A, \\ 0 & \text{falls } \omega \in A^C. \end{cases}$$

Dann ist der Erwartungswert

$$E[X] = 1 \cdot P[X = 1] + 0 \cdot P[X = 0] = P[A].$$

Ein Beispiel dafür ist ein elementarer Versicherungsvertrag mit Leistung

$$Y = \begin{cases} c & \text{falls } \omega \in A, \quad \text{»Schadensfall«,} \\ 0 & \text{sonst.} \end{cases}$$

Dann gilt:

$$Y = c \cdot I_A \quad \text{und} \quad E[Y] = c \cdot P[A].$$

**Beispiel** (Poissonverteilung). Sei  $X$  Poisson-verteilt mit Parameter  $\lambda$ . Dann ist der Erwartungswert

$$E[X] = \sum_{k=0}^{\infty} k \cdot P[X = k] = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{(k)!} e^{-\lambda} = \lambda.$$

Wir können daher den Parameter  $\lambda$  als Erwartungswert oder die mittlere Häufigkeit des Ereignisses interpretieren.

### Transformationsatz

Sei nun  $S$  eine beliebige abzählbare Menge,  $g : S \rightarrow \mathbb{R}$  eine Funktion und  $X : \Omega \rightarrow S$  eine Zufallsvariable. Wir definieren

$$g(X) : \Omega \rightarrow \mathbb{R}, \\ \omega \mapsto g(X(\omega)).$$

$g(X)$  ist eine **reellwertige Zufallsvariable**.

**Satz 1.7** (Transformationsatz). *Es gilt*

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

*falls die Summe wohldefiniert ist (zum Beispiel falls  $g$  nichtnegativ ist oder die Summe absolut konvergiert).*

*Beweis.* Es gilt mit Verwendung der  $\sigma$ -Additivität

$$\begin{aligned} E[g(X)] &= \sum_{b \in g(S)} b \cdot P[g(X) = b] = \sum_{b \in g(S)} b \cdot P\left[\bigcup_{a \in g^{-1}(b)} \{X = a\}\right] \\ &= \sum_{b \in g(S)} b \cdot \sum_{a \in g^{-1}(b)} P[X = a] \\ &= \sum_{b \in g(S)} \sum_{a \in g^{-1}(b)} g(a) \cdot P[X = a] \\ &= \sum_{a \in S} g(a) \cdot P[X = a]. \end{aligned}$$

□

**Bemerkung.** a) Insbesondere gilt:

$$E[|X|] = \sum_{a \in S} |a| \cdot P[X = a].$$

Ist  $E[|X|]$  endlich, dann konvergiert  $E[X] = \sum a \cdot P[X = a]$  absolut.

b) Ist  $\Omega$  abzählbar, dann folgt für  $X : \Omega \rightarrow \mathbb{R}$ :

$$E[X] = E[X \circ id_{\Omega}] = \sum_{\omega \in \Omega} X(\omega) \cdot P[\{\omega\}] = \sum_{\omega \in \Omega} X(\omega) p(\omega),$$

wobei  $id_{\Omega}$  die identische Abbildung auf  $\Omega$  bezeichnet. Der Erwartungswert ist das **gewichtete Mittel**. Ist  $P$  die Gleichverteilung auf  $\Omega$ , folgt weiter:

$$E[X] = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} X(\omega).$$

Der Erwartungswert ist in diesem Spezialfall das **arithmetische Mittel**.

**Beispiel** (Sankt-Petersburg-Paradoxon). Wir betrachten ein Glücksspiel mit fairen Münzwürfen  $X_1, X_2, \dots$ , wobei sich der Gewinn in jeder Runde verdoppelt bis zum ersten Mal »Kopf« fällt, dann ist das Spiel beendet. Wie hoch wäre eine faire Teilnahmegebühr für dieses Spiel?

Der Gewinn ist

$$G(\omega) = 2^{T(\omega)}, \quad \text{mit} \\ T(\omega) := \min\{n \in \mathbb{N} \mid X_n(\omega) = 1\}, \quad \text{der Wartezeit auf »Kopf«.}$$

Für den erwarteten Gewinn ergibt sich

$$E[G] = \sum_{k=1}^{\infty} 2^k \cdot P[T = k] = \sum_{k=1}^{\infty} 2^k \cdot P[X_1 = \dots = X_{k-1} = 1, X_k = 0] = \sum_{k=1}^{\infty} 2^k 2^{-k} \\ = \infty.$$

Das Spiel sollte also auf den ersten Blick bei beliebig hoher Teilnahmegebühr attraktiv sein – dennoch wäre wohl kaum jemand bereit, einen sehr hohen Einsatz zu zahlen.

Eine angemessenere Beschreibung – vom Blickwinkel des Spielers aus betrachtet – erhält man, wenn man eine (üblicherweise als monoton wachsend und konkav vorausgesetzte) Nutzenfunktion  $u(x)$  einführt, die den Nutzen beschreibt, den der Spieler vom Kapital  $x$  hat. Für kleine  $x$  könnte etwa  $u(x) = x$  gelten, aber für große  $x$  wäre plausibler  $u(x) < x$ . Dann ist  $c$  ein fairer Einsatz aus Sicht des Spielers, wenn  $u(c) = E[u(G)]$  gilt.

## Linearität und Monotonie des Erwartungswertes

**Satz 1.8** (Linearität des Erwartungswerts). *Seien  $X : \Omega \rightarrow S_X \subseteq \mathbb{R}$  und  $Y : \Omega \rightarrow S_Y \subseteq \mathbb{R}$  diskrete reellwertige Zufallsvariablen auf  $(\Omega, \mathcal{A}, P)$ , für die  $E[|X|]$  und  $E[|Y|]$  endlich sind, dann gilt:*

$$E[\lambda X + \mu Y] = \lambda E[X] + \mu E[Y] \quad \text{für alle } \lambda, \mu \in \mathbb{R}.$$

*Beweis.* Wir definieren  $g : S_X \times S_Y \rightarrow \mathbb{R}$ ,  $(x, y) \mapsto \lambda x + \mu y$ . Dann ist  $g(X, Y) = \lambda X + \mu Y$  eine Zufallsvariable mit Werten in  $S_X \times S_Y$ . Mit dem Transformationssatz folgt:

$$\begin{aligned} E[\lambda X + \mu Y] &= E[g(X, Y)] \\ &= \sum_{a \in S_X} \sum_{b \in S_Y} g(a, b) P[X = a, Y = b] \\ &= \sum_{a \in S_X} \sum_{b \in S_Y} (\lambda a + \mu b) P[X = a, Y = b] \\ &= \lambda \sum_{a \in S_X} a \sum_{b \in S_Y} P[X = a, Y = b] + \mu \sum_{b \in S_Y} b \sum_{a \in S_X} P[X = a, Y = b] \\ &= \lambda \sum_{a \in S_X} a P[X = a] + \mu \sum_{b \in S_Y} b P[Y = b] \\ &= \lambda E[X] + \mu E[Y]. \end{aligned} \tag{1.4.1}$$

Hierbei konvergiert die Reihe in (1.4.1) absolut, da

$$\begin{aligned} \sum_{a \in S_X} \sum_{b \in S_Y} |\lambda a + \mu b| P[X = a, Y = b] &\leq |\lambda| \sum_{a \in S_X} |a| P[X = a] + |\mu| \sum_{b \in S_Y} |b| P[Y = b] \\ &= |\lambda| E[|X|] + |\mu| E[|Y|] \end{aligned}$$

nach Voraussetzung endlich ist. □

**Korollar** (Monotonie des Erwartungswerts). *Seien die Voraussetzungen von Satz 1.8 erfüllt. Sei zusätzlich  $X(\omega) \leq Y(\omega)$  für alle  $\omega \in \Omega$ , dann gilt:*

$$E[X] \leq E[Y].$$

*Beweis.* Nach Voraussetzung gilt  $(Y - X)(\omega) \geq 0$  für alle  $\omega \in \Omega$ , weshalb der Erwartungswert  $E[Y - X]$  nichtnegativ ist. Aufgrund der Linearität des Erwartungswerts folgt:

$$0 \leq E[Y - X] = E[Y] - E[X].$$

□

**Beispiele** (Unabhängige 0-1-Experimente). Seien  $A_1, A_2, \dots, A_n \in \mathcal{A}$  unabhängige Ereignisse mit Wahrscheinlichkeit  $p$ , und sei

$$X_i = I_{A_i}, \quad \text{die Indikatorfunktion des Ereignisses } A_i, i = 0, \dots, n.$$

a) Die Zufallsvariablen  $X_i$  sind **Bernoulli-verteilt mit Parameter  $p$** , d.h.

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

Also gilt:

$$E[X_i] = E[I_{A_i}] = P[A_i] = p,$$

analog zu Beispiel 1.4.

b) Die Anzahl

$$S_n = X_1 + X_2 + \dots + X_n$$

der Ereignisse, die eintreten, ist binomialverteilt mit Parametern  $n$  und  $p$  (siehe Übung), d.h.

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

Den Erwartungswert kann man daher wie folgt berechnen:

$$\begin{aligned} E[S_n] &= \sum_{k=0}^n k \cdot P[S_n = k] = \sum_{k=0}^n k \binom{n}{k} p^k (1 - p)^{n-k} \\ &= \dots = n p. \end{aligned}$$

Einfacher benutzt man aber die Linearität des Erwartungswerts, und erhält

$$E[S_n] = E \left[ \sum_{i=1}^n X_i \right] = \sum_{i=1}^n E[X_i] = n p,$$

sogar **ohne Verwendung der Unabhängigkeit!**

**Beispiel** (Abhängige 0-1-Experimente). Wir betrachten eine Population aus  $m$  Objekten, davon  $r$  rote, aus der eine Zufallsstichprobe aus  $n$  Objekten ohne Zurücklegen entnommen wird,  $n \leq \min(r, m - r)$ . Sei  $A_i$  das Ereignis, daß das  $i$ -te Objekt in der Stichprobe rot ist, und  $X_i = I_{A_i}$ . Die Anzahl

$$S_n = X_1 + \dots + X_n$$



der roten Objekte in der Zufallsstichprobe ist dann hypergeometrisch verteilt mit Parametern  $m$ ,  $r$  und  $n$ . Als Erwartungswert dieser Verteilung erhalten wir analog zum letzten Beispiel:

$$E[S_n] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n P[A_i] = n \frac{r}{m}.$$

**Beispiel** (Inversionen von Zufallspermutationen). Seien  $\Omega = S_n$  die Menge aller Permutationen  $\omega: \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ ,  $P$  die Gleichverteilung auf  $\Omega$ , und

$$N(\omega) = |\{(i, j) \mid i < j \text{ und } \omega(i) > \omega(j)\}|,$$

die Anzahl der Inversionen einer Permutation  $\omega \in \Omega$ . Dann gilt

$$N = \sum_{1 \leq i < j \leq n} I_{A_{i,j}}, \quad \text{wobei}$$

$$A_{i,j} = \{\omega \in S_n \mid \omega(i) > \omega(j)\}$$

das Ereignis ist, daß eine Inversion von  $i$  und  $j$  auftritt. Es folgt:

$$E[N] = \sum_{i < j} E[I_{A_{i,j}}] = \sum_{i < j} P[\{\omega \in S_n \mid \omega(i) > \omega(j)\}] = \sum_{i < j} \frac{1}{2} = \frac{1}{2} \binom{n}{2} = \frac{n(n-1)}{4}.$$

**ANWENDUNG:** Beim Sortieralgorithmus »Insertion Sort« wird der Wert  $\omega(i)$  einer Liste  $\{\omega(1), \omega(2), \dots, \omega(n)\}$  beim Einfügen von  $\omega(j)$  genau dann verschoben, wenn  $\omega(j) < \omega(i)$  gilt. Ist die Anfangsanordnung eine Zufallspermutation der korrekten Anordnung, dann ist die mittlere Anzahl der Verschiebungen, die der Algorithmus vornimmt, also gleich  $\frac{n(n-1)}{4}$ .

**Satz 1.9** (Einschluss-/Ausschlussprinzip). Für  $n \in \mathbb{N}$  und Ereignisse  $A_1, \dots, A_n \in \mathcal{A}$  gilt:

$$P[A_1 \cup A_2 \cup \dots \cup A_n] = \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} P[A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}].$$

*Beweis.* Wir betrachten zunächst das Gegenereignis, und drücken die Wahrscheinlichkeiten als Erwartungswerte von Indikatorfunktionen aus. Unter Ausnutzung der Linearität des Erwartungswerts erhalten wir:

$$\begin{aligned}
 P[(A_1 \cup \dots \cup A_n)^C] &= P[A_1^C \cap \dots \cap A_n^C] = E[I_{A_1^C \cap \dots \cap A_n^C}] \\
 &= E\left[\prod_{i=1}^n I_{A_i^C}\right] = E\left[\prod_{i=1}^n (1 - I_{A_i})\right] \\
 &= \sum_{k=0}^n (-1)^k \sum_{1 \leq i_1 < \dots < i_k \leq n} E[I_{A_{i_1}} \cdot \dots \cdot I_{A_{i_k}}] \\
 &= \sum_{k=0}^n (-1)^k \sum_{1 \leq i_1 < \dots < i_k \leq n} E[I_{A_{i_1} \cap \dots \cap A_{i_k}}] \\
 &= \sum_{k=0}^n (-1)^k \sum_{1 \leq i_1 < \dots < i_k \leq n} P[A_{i_1} \cap \dots \cap A_{i_k}].
 \end{aligned}$$

Es folgt:

$$\begin{aligned}
 P[A_1 \cup \dots \cup A_n] &= 1 - P[(A_1 \cup \dots \cup A_n)^C] \\
 &= \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} P[A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}].
 \end{aligned}$$

□

# Kapitel 2

## Bedingte Wahrscheinlichkeiten und Unabhängigkeit

### 2.1 Bedingte Wahrscheinlichkeiten

Sei  $(\Omega, \mathcal{A}, P)$  ein Wahrscheinlichkeitsraum und  $A, B \in \mathcal{A}$  Ereignisse. Was ist die Wahrscheinlichkeit dafür, daß  $A$  eintritt, wenn wir schon wissen, daß  $B$  eintritt?

Relevante Fälle:  $\omega \in B$

Davon günstige Fälle:  $\omega \in A \cap B$

**Definition.** Sei  $P[B] \neq 0$ . Dann heißt

$$P[A|B] := \frac{P[A \cap B]}{P[B]}$$

die **bedingte Wahrscheinlichkeit von  $A$  gegeben  $B$** .

**Bemerkung.** a)  $P[\bullet|B]: \mathcal{A} \rightarrow [0, 1]$  ist eine Wahrscheinlichkeitsverteilung auf  $(\Omega, \mathcal{A})$ , die **bedingte Verteilung gegeben  $B$** . Der Erwartungswert

$$E[X|B] = \sum_{a \in S} a \cdot P[X = a|B]$$

einer diskreten Zufallsvariable  $X: \Omega \rightarrow S$  bzgl. der bedingten Verteilung heißt **bedingte Erwartung von  $X$  gegeben  $B$** .

b) Ist  $P$  die Gleichverteilung auf einer endlichen Menge  $\Omega$ , dann gilt:

$$P[A|B] = \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} = \frac{|A \cap B|}{|B|} \quad \text{für alle } A, B \subseteq \Omega.$$

**Beispiele.** a) Wir betrachten eine Familie mit 2 Kindern, und stellen die Frage nach dem Geschlecht der Kinder. Sei daher

$$\Omega = \{JJ, JM, MJ, MM\}.$$

Angenommen, alle Fälle wären gleich wahrscheinlich. Dann gilt:

$$P[\text{»beide Mädchen«} \mid \text{»eines Mädchen«}] = \frac{|\{MM\}|}{|\{MM, JM, MJ\}|} = \frac{1}{3},$$

$$P[\text{»beide Mädchen«} \mid \text{»das erste ist Mädchen«}] = \frac{|\{MM\}|}{|\{MM, MJ\}|} = \frac{1}{2}.$$

In Wirklichkeit sind die Kombinationen  $JJ$  und  $MM$  wahrscheinlicher.

b) Bei 20 fairen Münzwürfen fällt 15-mal »Zahl«. Wie groß ist die Wahrscheinlichkeit, daß die ersten 5 Würfe »Zahl« ergeben haben? Sei

$$\Omega = \{\omega = (x_1, \dots, x_{20}) \mid x_i \in \{0, 1\}\}, \quad \text{und} \\ X_i(\omega) = x_i, \quad \text{der Ausgang des } i\text{-ten Wurfs.}$$

Es gilt:

$$P[X_1 = \dots = X_5 = 1 \mid \sum_{i=1}^{20} X_i = 15] = \frac{P[X_1 = \dots = X_5 = 1 \text{ und } \sum_{i=6}^{20} X_i = 10]}{P[\sum_{i=1}^{20} X_i = 15]} \\ = \frac{2^{-5} \cdot 2^{-15} \binom{15}{10}}{2^{-20} \binom{20}{15}} = \frac{15 \cdot 14 \cdot \dots \cdot 11}{20 \cdot 19 \cdot \dots \cdot 16} \approx \frac{1}{5}.$$

Dagegen ist  $P[X_1 = \dots = X_5 = 1] = \frac{1}{32}$ .

## Berechnung von Wahrscheinlichkeiten durch Fallunterscheidung

Sei  $\Omega = \bigcup H_i$  eine disjunkte Zerlegung von  $\Omega$  in abzählbar viele Fälle (»Hypothesen«)  $H_i$ ,  $i \in I$ .

**Satz 2.1 (Formel von der totalen Wahrscheinlichkeit).** Für alle  $A \in \mathcal{A}$  gilt:

$$P[A] = \sum_{\substack{i \in I \\ P[H_i] \neq 0}} P[A|H_i] \cdot P[H_i]$$

*Beweis.* Es ist  $A = A \cap \left(\bigcup_{i \in I} H_i\right) = \bigcup_{i \in I} (A \cap H_i)$  eine disjunkte Vereinigung, also gilt nach  $\sigma$ -Additivität:

$$P[A] = \sum_{i \in I} P[A \cap H_i] = \sum_{i \in I} \underbrace{P[A \cap H_i]}_{=0, \text{ falls } P[H_i]=0} = \sum_{\substack{i \in I, \\ P[H_i] \neq 0}} P[A|H_i] \cdot P[H_i].$$

□

**Beispiel.** Urne 1 enthalte 2 rote und 3 schwarze Kugeln, Urne 2 enthalte 3 rote und 4 schwarze Kugeln. Wir legen eine Kugel  $K_1$  von Urne 1 in Urne 2 und ziehen eine Kugel  $K_2$  aus Urne 2. Mit welcher Wahrscheinlichkeit ist  $K_2$  rot?

$$\begin{aligned} P[K_2 \text{ rot}] &= P[K_2 \text{ rot} | K_1 \text{ rot}] \cdot P[K_1 \text{ rot}] + P[K_2 \text{ rot} | K_1 \text{ schwarz}] \cdot P[K_1 \text{ schwarz}] \\ &= \frac{4}{8} \cdot \frac{2}{5} + \frac{3}{8} \cdot \frac{3}{5} = \frac{17}{40}. \end{aligned}$$

**Beispiel (Simpson-Paradoxon).** Bewerbungen in Berkeley:

BEWERBUNGEN IN BERKELEY				
Statistik 1:	Männer	angenommen ( $A$ )	Frauen	angenommen ( $A$ )
	2083	996	1067	349
Empirische Verteilung:	$P[A M] \approx 0,48$		$P[A F] \approx 0,33$	

GENAUERE ANALYSE DURCH UNTERTEILUNG IN 4 FACHBEREICHE

Statistik 2:	Männer	angenommen ( $A$ )	Frauen	angenommen ( $A$ )
Bereich 1	825	511	108	89
Bereich 2	560	353	25	17
Bereich 3	325	110	593	219
Bereich 4	373	22	341	24

Sei  $P_M[A] := P[A|M]$  die empirische Verteilung unter Männern und  $P_F[A] := P[A|F]$  die empirische Verteilung unter Frauen, angenommen zu werden. Die Aufgliederung nach Fachbereichen ergibt folgende Zerlegung in Hypothesen:

$$P_M[A] = \sum_{i=1}^4 P_M[A|H_i] P_M[H_i], \quad P_F[A] = \sum_{i=1}^4 P_F[A|H_i] P_F[H_i].$$

Im Beispiel ist  $P_F[A|H_i] > P_M[A|H_i]$  für **alle**  $i$ , aber **dennoch**  $P_F[A] < P_M[A]$ . Die erste Statistik vermischt verschiedene Populationen und legt deshalb eventuell eine falsche Schlussfolgerung nahe.

## Bayessche Regel

Wie wahrscheinlich sind die Hypothesen  $H_i$ ? Ohne zusätzliche Information ist  $P[H_i]$  die Wahrscheinlichkeit von  $H_i$ . In der Bayesschen Statistik interpretiert man  $P[H_i]$  als unsere subjektive Einschätzung (aufgrund von vorhandenem oder nicht vorhandenem Vorwissen) über die vorliegende Situation (»a priori degree of belief«).

Angenommen, wir wissen nun zusätzlich, daß ein Ereignis  $A \in \mathcal{A}$  mit  $P[A] \neq 0$  eintritt, und wir kennen die bedingte Wahrscheinlichkeit (»likelihood«)  $P[A|H_i]$  für das Eintreten von  $A$  unter der Hypothese  $H_i$  für jedes  $i \in I$  mit  $P[H_i] \neq 0$ . Wie sieht dann unsere neue Einschätzung der Wahrscheinlichkeiten der  $H_i$  (»a posteriori degree of belief«) aus?

**Korollar** (Bayessche Regel). Für  $A \in \mathcal{A}$  mit  $P[A] \neq 0$  gilt:

$$P[H_i|A] = \frac{P[A|H_i] \cdot P[H_i]}{\sum_{\substack{j \in I \\ P[H_j] \neq 0}} P[A|H_j] \cdot P[H_j]} \quad \text{für alle } i \in I \text{ mit } P[H_i] \neq 0, \text{ d.h.}$$

$$P[H_i|A] = c \cdot P[H_i] \cdot P[A|H_i],$$

wobei  $c$  eine von  $i$  unabhängige Konstante ist.

*Beweis.* Es gilt:

$$P[H_i|A] = \frac{P[A \cap H_i]}{P[A]} = \frac{P[A|H_i] \cdot P[H_i]}{\sum_{\substack{j \in I \\ P[H_j] \neq 0}} P[A|H_j] \cdot P[H_j]}.$$

□

**Beispiel.** Von 10.000 Personen eines Alters habe einer die Krankheit  $K$ . Ein Test sei positiv (+) bei 96% der Kranken und 0,1% der Gesunden.

$$\text{A priori: } P[K] = \frac{1}{10000}, \quad P[K^C] = \frac{9999}{10000}.$$

$$\text{Likelihood: } P[+|K] = 0,96, \quad P[+|K^C] = 0,001.$$

A posteriori:

$$\begin{aligned} P[K|+] &= \frac{P[+|K] \cdot P[K]}{P[+|K] \cdot P[K] + P[+|K^C] \cdot P[K^C]} \\ &= \frac{0,96 \cdot 10^{-4}}{0,96 \cdot 10^{-4} + 10^{-3} \cdot 0,9999} \approx \frac{1}{11}. \end{aligned}$$

Daraus folgt insbesondere:  $P[K^C|+] \approx \frac{10}{11}$ , d.h. ohne zusätzliche Informationen muss man davon ausgehen, daß  $\frac{10}{11}$  der positiv getesteten Personen in Wirklichkeit gesund sind!

## 2.2 Mehrstufige diskrete Modelle

Wir betrachten ein  $n$ -stufiges Zufallsexperiment. Sind  $\Omega_1, \dots, \Omega_n$  abzählbare Stichprobenräume der Telexperimente, dann können wir

$$\Omega = \Omega_1 \times \dots \times \Omega_n = \{(\omega_1, \dots, \omega_n) \mid \omega_i \in \Omega_i\}$$

als Stichprobenraum des Gesamtexperiments auffassen und setzen  $\mathcal{A} = \mathcal{P}(\Omega)$ . Für  $\omega \in \Omega$  und  $k = 1, \dots, n$  sei

$$X_k(\omega) = \omega_k, \quad \text{der Ausgang des } k\text{-ten Telexperiments.}$$

Angenommen, wir kennen

$$P[X_1 = x_1] = p_1(x_1), \quad \text{für alle } x_1 \in \Omega_1, \quad (2.2.1)$$

die Verteilung (Massenfunktion) von  $X_1$ , sowie

$$P[X_k = x_k \mid X_1 = x_1, \dots, X_{k-1} = x_{k-1}] = p_k(x_k \mid x_1, \dots, x_{k-1}), \quad (2.2.2)$$

die bedingte Verteilung von  $X_k$  gegeben  $X_1, \dots, X_{k-1}$  für  $k = 2, \dots, n$ ,  $x_i \in \Omega_i$  mit  $P[X_1 = x_1, \dots, X_{k-1} = x_{k-1}] \neq 0$ .

Wie sieht die gesamte Wahrscheinlichkeitsverteilung  $P$  auf  $\Omega$  aus?

**Satz 2.2.** Seien  $p_1$  und  $p_k(\bullet \mid x_1, \dots, x_{k-1})$  für jedes  $k = 2, \dots, n$  und  $x_1 \in \Omega_1, \dots, x_{k-1} \in \Omega_{k-1}$  die Massenfunktion einer Wahrscheinlichkeitsverteilung auf  $\Omega_k$ . Dann existiert genau eine Wahrscheinlichkeitsverteilung  $P$  auf  $(\Omega, \mathcal{A})$  mit (2.2.1) und (2.2.2). Diese ist bestimmt durch die Massenfunktion

$$p(x_1, \dots, x_n) = p_1(x_1) p_2(x_2 \mid x_1) p_3(x_3 \mid x_1, x_2) \cdots p_n(x_n \mid x_1, \dots, x_{n-1}).$$

*Beweis.*

- EINDEUTIGKEIT:

Wir behaupten, dass für eine Verteilung  $P$  mit (2.2.1) und (2.2.2) gilt:

$$P[X_1 = x_1, \dots, X_k = x_k] = p_1(x_1) \cdot p_2(x_2 \mid x_1) \cdots p_k(x_k \mid x_1, \dots, x_{k-1}), \quad k = 1, \dots, n.$$

Der Induktionsanfang folgt aus Bedingung (2.2.1). Sei die Induktionsbehauptung für  $k - 1$  wahr, dann folgt nach Induktionsannahme und (2.2.2):

$$\begin{aligned} P[X_1 = x_1, \dots, X_k = x_k] &= P[X_1 = x_1, \dots, X_{k-1} = x_{k-1}] \\ &\quad \cdot P[X_1 = x_1, \dots, X_k = x_k \mid X_1 = x_1, \dots, X_{k-1} = x_{k-1}] \\ &= p_1(x_1) \cdot p_2(x_2 \mid x_1) \cdots p_{k-1}(x_{k-1} \mid x_1, \dots, x_{k-2}) \\ &\quad \cdot p_k(x_k \mid x_1, \dots, x_{k-1}), \end{aligned}$$

falls  $P[X_1 = x_1, \dots, X_{k-1} = x_{k-1}] \neq 0$ . Andernfalls verschwinden beide Seiten und die Behauptung folgt. Für  $k = n$  erhalten wir als Massenfunktion von  $P$ :

$$p(x_1, \dots, x_n) = P[X_1 = x_1, \dots, X_n = x_n] = p_1(x_1) \cdots p_n(x_n \mid x_1, \dots, x_{n-1}).$$

- EXISTENZ:

$p$  ist Massenfunktion einer Wahrscheinlichkeitsverteilung  $P$  auf  $\Omega_1 \times \cdots \times \Omega_n$ , denn:

$$\begin{aligned} \sum_{x_1 \in \Omega_1} \cdots \sum_{x_n \in \Omega_n} p(x_1, \dots, x_n) &= \sum_{x_1 \in \Omega_1} p_1(x_1) \sum_{x_2 \in \Omega_2} p_2(x_2 \mid x_1) \cdots \underbrace{\sum_{x_n \in \Omega_n} p_n(x_n \mid x_1, \dots, x_{n-1})}_{=1} \\ &= 1. \end{aligned}$$

Für  $P$  gilt:

$$\begin{aligned} P[X_1 = x_1, \dots, X_k = x_k] &= \sum_{x_{k+1} \in \Omega_{k+1}} \cdots \sum_{x_n \in \Omega_n} p(x_1, \dots, x_n) \\ &= p_1(x_1) p_2(x_2 \mid x_1) \cdots p_k(x_k \mid x_1, \dots, x_{k-1}), \quad k = 1, \dots, n. \end{aligned}$$

Damit folgen (2.2.1) und (2.2.2). □

**Beispiel.** Wie groß ist die Wahrscheinlichkeit, daß beim Skat jeder Spieler genau einen der vier Buben erhält? Sei

$$\begin{aligned} \Omega &= \{(\omega_1, \omega_2, \omega_3) \mid \omega_i \in \{0, 1, 2, 3, 4\}\}, \\ X_i(\omega) &= \omega_i = \text{Anzahl der Buben von Spieler } i. \end{aligned}$$



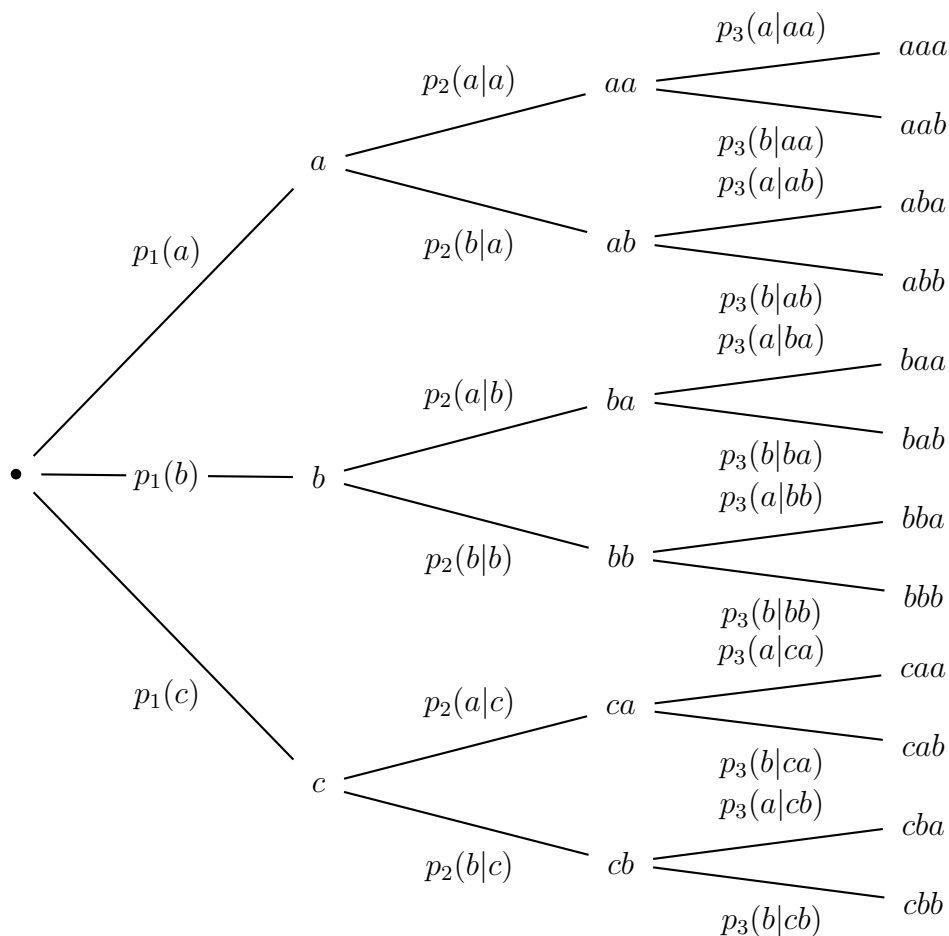


Abbildung 2.1: Baumdarstellung der Fallunterscheidungen

Es gilt:

$$p_1(x_1) = \frac{\binom{4}{x_1} \binom{28}{10-x_1}}{\binom{32}{10}}, \quad \text{hypergeometrische Verteilung,}$$

$$p_2(x_2 | x_1) = \frac{\binom{4-x_1}{x_2} \binom{18+x_1}{10-x_2}}{\binom{22}{10}}$$

$$p_3(x_3 | x_1, x_2) = \begin{cases} \frac{\binom{4-x_1-x_2}{x_3} \binom{18+x_1+x_2}{10-x_3}}{\binom{12}{10}} & \text{falls } 2 \leq x_1 + x_2 + x_3 \leq 4, \\ 0 & \text{sonst.} \end{cases}$$

Damit folgt:

$$p(1, 1, 1) = p_1(1) p_2(1 | 1) p_3(1 | 1, 1) \approx 5,56\%.$$

Im folgenden betrachten wir zwei fundamentale Klassen von mehrstufigen Modellen, Produktmodelle und Markov-Ketten.

## Produktmodelle

Angenommen, der Ausgang des  $i$ -ten Experiments hängt nicht von  $x_1, \dots, x_{i-1}$  ab. Dann sollte gelten:

$$p_i(x_i \mid x_1, \dots, x_{i-1}) = p_i(x_i)$$

mit einer von  $x_1, \dots, x_{i-1}$  unabhängigen Massenfunktion  $p_i$  einer Wahrscheinlichkeitsverteilung  $P_i$  auf  $\Omega_i$ . Die Wahrscheinlichkeitsverteilung  $P$  auf  $\Omega$  hat dann die Massenfunktion

$$p(x_1, \dots, x_n) = \prod_{i=1}^n p_i(x_i), \quad x \in \Omega. \quad (2.2.3)$$

**Definition.** Die Wahrscheinlichkeitsverteilung  $P$  auf  $\Omega = \Omega_1 \times \dots \times \Omega_n$  mit Massenfunktion (2.2.3) heißt **Produkt** von  $P_1, \dots, P_n$  und wird mit  $P_1 \otimes \dots \otimes P_n$  notiert.

**Beispiel** ( $n$ -dimensionale Bernoulli-Verteilung). Wir betrachten  $n$  unabhängige 0-1-Experimente mit Erfolgswahrscheinlichkeit  $p$ :

$$\Omega_1 = \dots = \Omega_n = \{0, 1\}, \quad p_i(1) = p, \quad p_i(0) = 1 - p, \quad i = 1, \dots, n.$$

Sei  $k = \sum_{i=1}^n x_i$  die Anzahl der Einsen. Dann ist

$$p(x_1, \dots, x_n) = \prod_{i=1}^n p_i(x_i) = p^k (1 - p)^{n-k}$$

die  **$n$ -dimensionale Bernoulli-Verteilung**.

**Bemerkung.** Sind die Mengen  $\Omega_i$ ,  $i = 1, \dots, n$  endlich, und ist  $P_i$  die Gleichverteilung auf  $\Omega_i$ , dann ist  $P_1 \otimes \dots \otimes P_n$  die Gleichverteilung auf  $\Omega_1 \times \dots \times \Omega_n$ .

Die Multiplikatивität im Produktmodell gilt nicht nur für die Massenfunktion, sondern allgemeiner für die Wahrscheinlichkeiten, daß in den Telexperimenten bestimmte Ereignisse  $A_1, \dots, A_n$  eintreten:

**Satz 2.3.** Im Produktmodell gilt für beliebige Ereignisse  $A_i \subseteq \Omega_i$ ,  $i = 1, \dots, n$ :

$$\begin{aligned} P[X_1 \in A_1, \dots, X_n \in A_n] &= \prod_{i=1}^n P[X_i \in A_i] & (2.2.4) \\ &\parallel & \\ &\parallel & \\ P[A_1 \times \dots \times A_n] &= \prod_{i=1}^n P_i[A_i] \end{aligned}$$

(d.h.  $X_1, \dots, X_n$  sind **unabhängige** Zufallsvariablen, siehe nächsten Abschnitt).

*Beweis.* Es gilt:

$$\begin{aligned} P[X_1 \in A_1, \dots, X_n \in A_n] &= P[(X_1, \dots, X_n) \in A_1 \times \dots \times A_n] = P[A_1 \times \dots \times A_n] \\ &= \sum_{x \in A_1 \times \dots \times A_n} p(x) = \sum_{x_1 \in A_1} \dots \sum_{x_n \in A_n} \prod_{i=1}^n p_i(x_i) \\ &= \prod_{i=1}^n \sum_{x_i \in A_i} p_i(x_i) = \prod_{i=1}^n P_i[A_i]. \end{aligned}$$

Insbesondere gilt:

$$P[X_i \in A_i] = P[X_1 \in \Omega, \dots, X_{i-1} \in \Omega, X_i \in A_i, X_{i+1} \in \Omega, \dots, X_n \in \Omega] = P_i[A_i].$$

□

## Markov-Ketten

Zur Modellierung einer zufälligen zeitlichen Entwicklung mit abzählbarem Zustandsraum  $S$  betrachten wir den Stichprobenraum

$$\Omega = S^{n+1} = \{(x_0, x_1, \dots, x_n) \mid x_i \in S\}.$$

Oft ist es naheliegend anzunehmen, daß die Weiterentwicklung des Systems nur vom gegenwärtigen Zustand, aber nicht vom vorherigen Verlauf abhängt (»kein Gedächtnis«), d.h. es sollte gelten:

$$p_k(x_k \mid x_0, \dots, x_{k-1}) = \underbrace{p_k(x_{k-1}, x_k)}_{\text{»Bewegungsgesetz«}}, \quad (2.2.5)$$

wobei  $p_k : S \times S \rightarrow [0, 1]$  folgende Bedingungen erfüllt:

- i)  $p_k(x, y) \geq 0$  für alle  $x, y \in S$
- ii)  $\sum_{y \in S} p_k(x, y) = 1$  für alle  $x \in S$

d.h.  $p_k(x, \bullet)$  ist für jedes  $x \in S$  die Massenfunktion einer Wahrscheinlichkeitsverteilung auf  $S$ .

**Definition.** Eine Matrix  $p_k(x, y)$  ( $x, y \in S$ ) mit i) und ii) heißt **stochastische Matrix** (oder **stochastischer Kern**) auf  $S$ .

Im **Mehrstufenmodell** folgt aus Gleichung (2.2.5):

$$p(x_0, x_1, \dots, x_n) = \underbrace{p_0(x_0)}_{\text{»Startverteilung«}} p_1(x_0, x_1) p_2(x_1, x_2) \cdots p_n(x_{n-1}, x_n), \quad \text{für } x_0, \dots, x_n \in S.$$

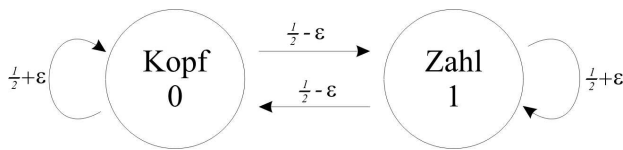
Den Fall, in dem der Übergangsmechanismus  $p_k(x, y) = p(x, y)$  unabhängig von  $k$  ist, nennt man **zeitlich homogen**.

**Beispiele.** a) PRODUKTMODELL (siehe oben):

$$p_k(x, y) = p_k(y) \text{ ist unabhängig von } x.$$

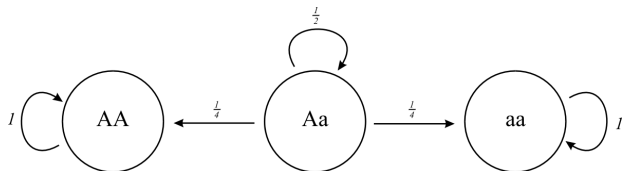
b) ABHÄNGIGE MÜNZWÜRFE:

$$S = \{0, 1\}, \quad \varepsilon \in \left[-\frac{1}{2}, \frac{1}{2}\right].$$



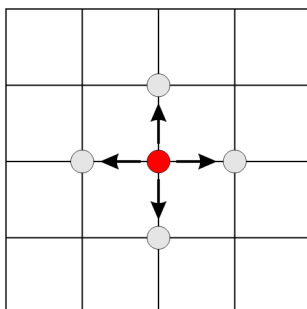
$$p = \begin{pmatrix} \frac{1}{2} + \varepsilon & \frac{1}{2} - \varepsilon \\ \frac{1}{2} - \varepsilon & \frac{1}{2} + \varepsilon \end{pmatrix}.$$

c) SELBSTBEFRUCHTUNG VON PFLANZEN:



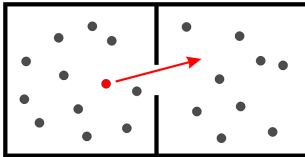
$$p = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 1 \end{pmatrix}$$

d) RANDOM WALK AUF  $S = \mathbb{Z}^d$ , ( $d \in \mathbb{N}$ ):



$$p(x, y) = \begin{cases} \frac{1}{2d} & \text{falls } |x - y| = 1, \\ 0 & \text{sonst.} \end{cases}$$

- e) URNENMODELL VON P. UND T. EHRENFEST (Austausch von Gasmolekülen in zwei Behältern):



Es seien  $N$  Kugeln auf zwei Urnen verteilt. Zu jedem Zeitpunkt  $t \in \mathbb{N}$  wechselt eine zufällig ausgewählte Kugel die Urne.

MAKROSKOPISCHES MODELL:

$$S = \{0, 1, 2, \dots, n\}.$$

$x \in S$  beschreibt die Anzahl Kugeln in der ersten Urne.

$$p(x, y) = \begin{cases} \frac{x}{n} & \text{falls } y = x - 1, \\ \frac{n-x}{n} & \text{falls } y = x + 1, \\ 0 & \text{sonst.} \end{cases}$$

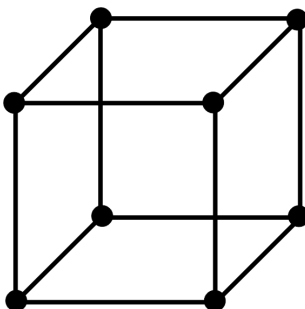
MIKROSKOPISCHES MODELL:

$$S = \{0, 1\}^n = \{(\sigma_1, \dots, \sigma_n) \mid \sigma_i \in \{0, 1\}\}.$$

Es ist  $\sigma_i = 1$  genau dann, wenn sich die  $i$ -te Kugel in Urne 1 befindet.

$$p(\sigma, \tilde{\sigma}) = \begin{cases} \frac{1}{N} & \text{falls } \sum_{i=1}^n |\sigma_i - \tilde{\sigma}_i| = 1, \\ 0 & \text{sonst.} \end{cases}$$

Die resultierende Markov-Kette ist ein Random Walk auf dem Hyperwürfel  $\{0, 1\}^n$ , d.h. sie springt in jedem Schritt von einer Ecke des Hyperwürfels zu einer zufällig ausgewählten benachbarten Ecke.



**Berechnung von Wahrscheinlichkeiten**

**Satz 2.4** (Markov-Eigenschaft). Für alle  $0 \leq k < l \leq n$  und  $x_0, \dots, x_l \in S$  mit  $P[X_0 = x_0, \dots, X_k = x_k] \neq 0$  gilt:

$$\begin{aligned} P[X_l = x_l \mid X_0 = x_0, \dots, X_k = x_k] &= P[X_l = x_l \mid X_k = x_k] \\ &= (p_{k+1} p_{k+2} \cdots p_l)(x_k, x_l), \end{aligned}$$

wobei

$$(pq)(x, y) := \sum_{z \in S} p(x, z) q(z, y)$$

das Produkt der Matrizen  $p$  und  $q$  ist.

**Bemerkung.** a) MARKOV-EIGENSCHAFT:

Die Weiterentwicklung hängt jeweils nur vom gegenwärtigen Zustand  $x_k$  ab, und nicht vom vorherigen Verlauf  $x_0, x_1, \dots, x_{k-1}$ .

b)  $n$ -SCHRITT-ÜBERGANGSWAHRSCHEINLICHKEITEN:

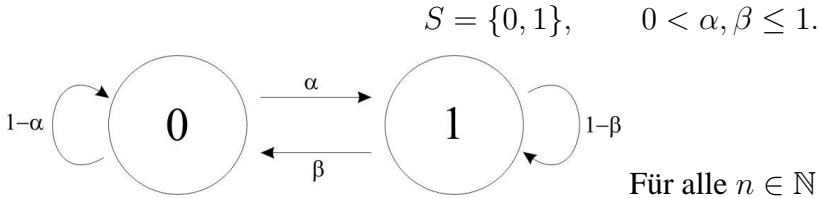
$$\begin{aligned} P[X_n = y \mid X_0 = x] &= (p_1 p_2 \cdots p_n)(x, y) \\ &= p^n(x, y) \quad \text{falls zeitlich homogen, d.h. } p_i \equiv p. \end{aligned}$$

*Beweis.*

$$\begin{aligned} P[X_l = x_l \mid X_0 = x_0, \dots, X_k = x_k] &= \frac{P[X_0 = x_0, \dots, X_k = x_k, X_l = x_l]}{P[X_0 = x_0, \dots, X_k = x_k]} \\ &= \frac{\sum_{x_{k+1}, \dots, x_{l-1}} p_0(x_0) p_1(x_0, x_1) \cdots p_l(x_{l-1}, x_l)}{p_0(x_0) p_1(x_0, x_1) \cdots p_k(x_{k-1}, x_k)} \\ &= \sum_{x_{k+1}} \cdots \sum_{x_{l-1}} p_{k+1}(x_k, x_{k+1}) p_{k+2}(x_{k+1}, x_{k+2}) \cdots p_l(x_{l-1}, x_l) \\ &= (p_{k+1} p_{k+2} \cdots p_l)(x_k, x_l). \\ P[X_l = x_l \mid X_k = x_k] &= \frac{P[X_k = x_k, X_l = x_l]}{P[X_k = x_k]} \\ &= \frac{\sum_{x_1, \dots, x_{k-1}} \sum_{x_{k+1}, \dots, x_{l-1}} p_0(x_0) p_1(x_0, x_1) \cdots p_l(x_{l-1}, x_l)}{\sum_{x_1, \dots, x_{k-1}} p_0(x_0) p_1(x_0, x_1) \cdots p_k(x_{k-1}, x_k)} \\ &= (p_{k+1} p_{k+2} \cdots p_l)(x_k, x_l). \end{aligned}$$

□

**Beispiel.**



$$\begin{aligned} p^n(0, 0) &= p^{n-1}(0, 0) \cdot p(0, 0) + p^{n-1}(0, 1) \cdot p(1, 0) \\ &= p^{n-1}(0, 0) \cdot (1 - \alpha) + (1 - p^{n-1}(0, 0)) \cdot \beta \\ &= (1 - \alpha - \beta) \cdot p^{n-1}(0, 0) + \beta. \end{aligned}$$

Daraus folgt mit Induktion:

$$\begin{aligned} p^n(0, 0) &= \frac{\beta}{\alpha + \beta} + \frac{\alpha}{\alpha + \beta} (1 - \alpha - \beta)^n, \quad \text{und} \\ p^n(0, 1) &= 1 - p^n(0, 0). \end{aligned}$$

Analoge Formeln erhält man für  $p^n(1, 0)$  und  $p^n(1, 1)$  durch Vertauschung von  $\alpha$  und  $\beta$ . Für die  $n$ -Schritt-Übergangsmatrix ergibt sich:

$$p^n = \underbrace{\begin{pmatrix} \frac{\beta}{\alpha + \beta} & \frac{\alpha}{\alpha + \beta} \\ \frac{\beta}{\alpha + \beta} & \frac{\alpha}{\alpha + \beta} \end{pmatrix}}_{\text{Gleiche Zeilen}} + \underbrace{(1 - \alpha - \beta)^n \begin{pmatrix} \frac{\alpha}{\alpha + \beta} & \frac{-\alpha}{\alpha + \beta} \\ \frac{-\beta}{\alpha + \beta} & \frac{\beta}{\alpha + \beta} \end{pmatrix}}_{\rightarrow 0 \text{ exponentiell schnell, falls } \alpha < 1 \text{ oder } \beta < 1}.$$

Insbesondere gilt  $p^n(0, \cdot) \approx p^n(1, \cdot)$  für große  $n \in \mathbb{N}$ . Die Kette »vergisst« also ihren Startwert exponentiell schnell (»Exponentieller Gedächtnisverlust«)!

### 2.3 Unabhängigkeit von Ereignissen

Sei  $(\Omega, \mathcal{A}, P)$  ein Wahrscheinlichkeitsraum. Hängen zwei Ereignisse  $A, B \in \mathcal{A}$  **nicht voneinander ab**, dann sollte gelten:

$$P[A|B] = P[A], \quad \text{falls } P[B] \neq 0,$$

sowie

$$\underbrace{P[B|A]}_{\frac{P[B \cap A]}{P[A]}} = P[B], \quad \text{falls } P[A] \neq 0,$$

also insgesamt

$$P[A \cap B] = P[A] \cdot P[B]. \tag{2.3.1}$$

**Definition.** i) Zwei Ereignisse  $A, B \in \mathcal{A}$  heißen **unabhängig** (bzgl.  $P$ ), falls (2.3.1) gilt.

ii) Eine beliebige Kollektion  $A_i, i \in I$ , von Ereignissen heißt **unabhängig** (bzgl.  $P$ ), falls

$$P[A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_n}] = \prod_{k=1}^n P[A_{i_k}]$$

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

**Beispiele.** a) Falls  $P[A] \in \{0, 1\}$  gilt, ist  $A$  unabhängig von  $B$  für alle  $B \in \mathcal{A}$ .

b) Wir betrachten das Modell für ZWEI FAIRE MÜNZWÜRFE, also  $\Omega = \{0, 1\}^2$  und  $P$  sei die Gleichverteilung. Die Ereignisse

$$A_1 = \{(1, 0), (1, 1)\}, \quad \text{»erster Wurf Zahl«},$$

$$A_2 = \{(0, 1), (1, 1)\}, \quad \text{»zweiter Wurf Zahl«},$$

$$A_3 = \{(0, 0), (1, 1)\}, \quad \text{»beide Würfe gleich«},$$

sind **paarweise unabhängig**, denn es gilt:

$$P[A_i \cap A_j] = \frac{1}{4} = P[A_i] \cdot P[A_j] \quad \text{für alle } i \neq j.$$

Allerdings ist die Kollektion  $A_1, A_2, A_3$  **nicht unabhängig**, denn es gilt

$$P[A_1 \cap A_2 \cap A_3] = \frac{1}{4} \neq P[A_1] \cdot P[A_2] \cdot P[A_3].$$

**Lemma 2.5.** Seien die Ereignisse  $A_1, \dots, A_n \in \mathcal{A}$  unabhängig,  $B_j = A_j$  oder  $B_j = A_j^C$  für alle  $j = 1, \dots, n$ . Dann sind die Ereignisse  $B_1, \dots, B_n$  unabhängig.

*Beweis.* Sei ohne Beschränkung der Allgemeinheit:

$$B_1 = A_1, \quad \dots, \quad B_k = A_k, \quad B_{k+1} = A_{k+1}^C, \quad \dots, \quad B_n = A_n^C$$

. Dann gilt unter Verwendung der Linearität des Erwartungswerts und der Unabhängigkeit von  $A_1, \dots, A_n$ :

$$\begin{aligned} P[B_1 \cap \dots \cap B_n] &= P[A_1 \cap \dots \cap A_k \cap A_{k+1}^C \cap \dots \cap A_n^C] \\ &= E[I_{A_1} \cdots I_{A_k} \cdot (1 - I_{A_{k+1}}) \cdots (1 - I_{A_n})] \\ &= E[I_{A_1} \cdots I_{A_k} \cdot \sum_{J \subseteq \{k+1, \dots, n\}} (-1)^{|J|} \prod_{j \in J} I_{A_j}] \\ &= \sum_{J \subseteq \{k+1, \dots, n\}} (-1)^{|J|} P[A_1 \cap \dots \cap A_k \cap \bigcap_{j \in J} A_j] \\ &= \sum_{J \subseteq \{k+1, \dots, n\}} (-1)^{|J|} P[A_1] \cdots P[A_k] \cdot \prod_{j \in J} P[A_j] \\ &= P[A_1] \cdots P[A_k] \cdot (1 - P[A_{k+1}]) \cdots (1 - P[A_n]) = P[B_1] \cdots P[B_n]. \end{aligned}$$

□



## Verteilungen für unabhängige Ereignisse

Seien  $A_1, A_2, \dots \in \mathcal{A}$  unabhängige Ereignisse (bzgl.  $P$ ) mit  $P[A_i] = p \in [0, 1]$ . Die Existenz von unendlich vielen unabhängigen Ereignissen auf einem geeigneten Wahrscheinlichkeitsraum setzen wir hier voraus – ein Beweis wird erst in der Vorlesung »Einführung in die Wahrscheinlichkeitstheorie« gegeben.

### Geometrische Verteilung

Die Wartezeit auf das erste Eintreten eines der Ereignisse ist

$$T(\omega) = \inf\{n \in \mathbb{N} \mid \omega \in A_n\}, \quad \text{wobei } \min \emptyset := \infty.$$

Mit Lemma 2.5 folgt:

$$\begin{aligned} P[T = n] &= P[A_1^C \cap A_2^C \cap \dots \cap A_{n-1}^C \cap A_n] \\ &= P[A_n] \cdot \prod_{i=1}^{n-1} P[A_i^C] \\ &= p \cdot (1 - p)^{n-1}. \end{aligned}$$

**Definition.** Sei  $p \in [0, 1]$ . Die Wahrscheinlichkeitsverteilung auf  $\mathbb{N}$  mit Massenfunktion

$$p(n) = p \cdot (1 - p)^{n-1}$$

heißt *geometrische Verteilung zum Parameter  $p$* .

**Bemerkung.** a) Für  $p \neq 0$  gilt:

$$\sum_{n=1}^{\infty} p \cdot (1 - p)^{n-1} = 1,$$

d.h. die geometrische Verteilung ist eine Wahrscheinlichkeitsverteilung auf den natürlichen Zahlen, und

$$P[T = \infty] = 0.$$

b) Allgemein gilt:

$$P[T > n] = P[A_1^C \cap \dots \cap A_n^C] = (1 - p)^n.$$

c) Es gilt:

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

(siehe Übung).

### Binomialverteilung

Die Anzahl der Ereignisse unter  $A_1, \dots, A_n$ , die eintreten, ist

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

Es gilt:

$$\begin{aligned} P[S_n = k] &= \sum_{\substack{I \subseteq \{1, \dots, n\} \\ |I|=k}} P\left[\bigcap_{i \in I} A_i \cap \bigcap_{i \in \{1, \dots, n\} \setminus I} A_i^C\right] \\ &= \sum_{\substack{I \subseteq \{1, \dots, n\} \\ |I|=k}} \prod_{i \in I} P[A_i] \cdot \prod_{i \in I^C} P[A_i^C] \\ &= \sum_{\substack{I \subseteq \{1, \dots, n\} \\ |I|=k}} \prod_{i \in I} p \cdot \prod_{i \in I^C} (1-p) \\ &= \sum_{\substack{I \subseteq \{1, \dots, n\} \\ |I|=k}} p^{|I|} \cdot (1-p)^{|I^C|} \\ &= \binom{n}{k} p^k (1-p)^{n-k}, \end{aligned}$$

d.h.  $S_n$  ist **Binomialverteilt mit Parametern  $n$  und  $p$** .

**Satz 2.6** (»Law of Averages«, Bernstein-Ungleichung). Für alle  $\varepsilon > 0$  und  $n \in \mathbb{N}$  gilt:

$$P\left[\frac{S_n}{n} \geq p + \varepsilon\right] \leq e^{-2\varepsilon^2 n}, \quad \text{und}$$

$$P\left[\frac{S_n}{n} \leq p - \varepsilon\right] \leq e^{-2\varepsilon^2 n}.$$

Insbesondere gilt:

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

d.h. die Wahrscheinlichkeit für eine Abweichung des Mittelwerts  $\frac{S_n}{n}$  vom Erwartungswert  $p$  um mehr als  $\varepsilon$  fällt exponentiell in  $n$ .

**Bemerkung.** a) Satz 2.6 ist eine erste Version des »Gesetzes der großen Zahlen«.

b) Der Satz liefert eine nachträgliche Rechtfertigung der frequentistischen Interpretation der Wahrscheinlichkeit als asymptotische relative Häufigkeit.

c) Anwendung auf Schätzen von  $p$ :

$$p \approx \frac{S_n}{n} = \text{relative Häufigkeit des Ereignisses bei } n \text{ unabhängigen Stichproben.}$$

d) Anwendung auf näherungsweise Monte Carlo-Berechnung von  $p$ :

Simuliere  $n$  unabhängige Stichproben,  $p \sim$  relative Häufigkeit.

*Beweis.* Sei  $q := 1 - p$ ,  $S_n \sim \text{Bin}(n, p)$ . Dann gilt für  $\lambda > 0$ :

$$\begin{aligned} P[S_n \geq n(p + \varepsilon)] &= \sum_{k \geq np + n\varepsilon} \binom{n}{k} p^k q^{n-k} \\ &\leq \sum_{k \geq np + n\varepsilon} \binom{n}{k} e^{\lambda k} p^k q^{n-k} e^{-\lambda(np + n\varepsilon)} \\ &\leq \sum_{k=0}^n \binom{n}{k} (pe^\lambda)^k q^{n-k} e^{-\lambda np} e^{-\lambda n\varepsilon} \\ &= (pe^\lambda + q)^n e^{-\lambda np} e^{-\lambda n\varepsilon} \leq (pe^{\lambda q} + qe^{-\lambda p})^n e^{-\lambda n\varepsilon}. \end{aligned}$$

Wir behaupten:

$$pe^{\lambda q} + qe^{-\lambda p} \leq e^{\frac{\lambda^2}{8}}.$$

Damit folgt:

$$P[S_n \geq n(p + \varepsilon)] \leq e^{n(\frac{\lambda^2}{8} - \lambda\varepsilon)}.$$

Der Exponent ist minimal für  $\lambda = 4\varepsilon$ . Für diese Wahl von  $\lambda$  folgt schließlich

$$P[S_n \geq n(p + \varepsilon)] \leq e^{-2n\varepsilon^2}.$$

*Beweis der Behauptung:*

$$f(\lambda) := \log(pe^{\lambda q} + qe^{-\lambda p}) = \log(e^{-\lambda p}(pe^\lambda + q)) = -\lambda p + \log(pe^\lambda + q).$$

Zu zeigen ist nun

$$f(\lambda) \leq \frac{\lambda^2}{8} \quad \text{für alle } \lambda \geq 0.$$

Es gilt:

$$\begin{aligned} f(0) &= 0, \\ f'(\lambda) &= -p + \frac{pe^\lambda}{pe^\lambda + q} = -p + \frac{p}{p + qe^{-\lambda}}, \quad f'(0) = 0, \\ f''(\lambda) &= \frac{pqe^{-\lambda}}{(p + qe^{-\lambda})^2} \leq \frac{1}{4}. \end{aligned}$$

Die letzte Ungleichung folgt aus::

$$(a + b)^2 = a^2 + b^2 + 2ab \geq 4ab$$

Damit folgt

$$\begin{aligned} f(\lambda) &= f(0) + \int_0^\lambda f'(x) dx \\ &= \int_0^\lambda \int_0^x f''(y) dy dx \leq \int_0^\lambda \frac{x}{4} dx \leq \frac{\lambda^2}{8} \quad \text{für alle } \lambda \geq 0. \end{aligned}$$

□

**Beispiel.** Im letzten Satz wurde gezeigt:

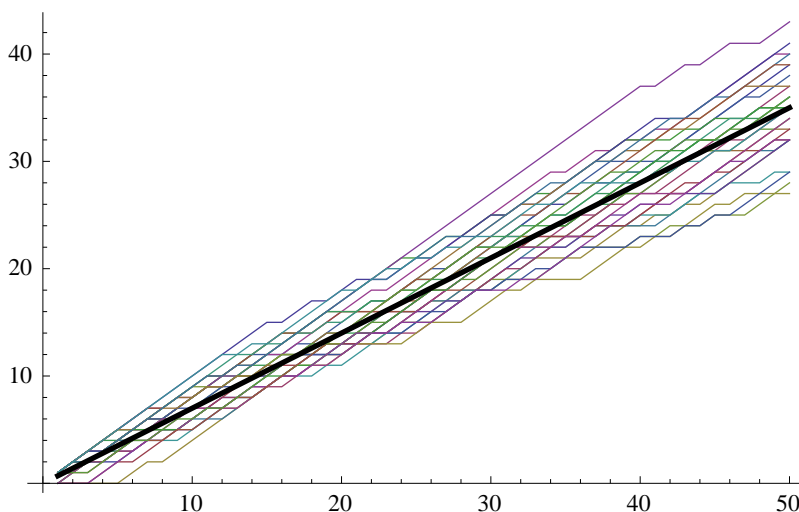
$$S_n = \sum_{i=1}^n I_{A_i}, \quad A_i \text{ unabhängig mit } P[A_i] = p \implies P\left[\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right] \longrightarrow 0 \quad \text{für } n \rightarrow \infty.$$

Zur Demonstration simulieren wir den Verlauf von  $S_n$  und  $\frac{S_n}{n}$  mehrfach ( $m$ -mal):

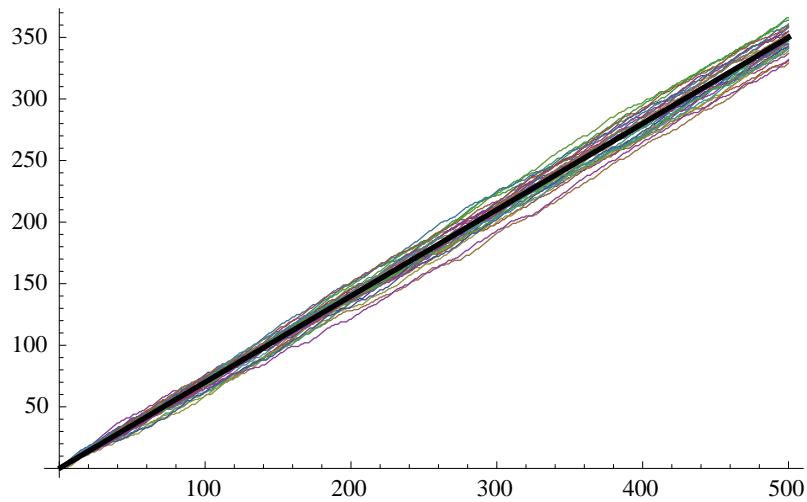
VERLAUF VON  $S_n$

```
m = 30; nmax = 1000; p = 0.7;
(Wir erzeugen  $m \times nmax$  Bernoulli-Stichproben mit Wahrscheinlichkeit p)
x = RandomChoice[{1 - p, p} -> {0, 1}, {nmax, m}]; s = Accumulate[x];
Das Feld s enthält m Verläufe von  $s_n = x_1 + \dots + x_n, n = 1, \dots, nmax$ 
Manipulate [Show[
  ListLinePlot[Transpose[s[[1 ;; n]]]],
  ListLinePlot[p*Range[n], PlotStyle -> {Black, Thick}]]
, {{n, 50}, 1, nmax, 1}]
(Vergleich der  $m$  Verläufe von  $s_n$  mit  $np$ )
```

- $n = 50$ :



- $n = 500$ :



VERLAUF VON  $\frac{S_n}{n}$

```
mean = s / Range[nmax];
```

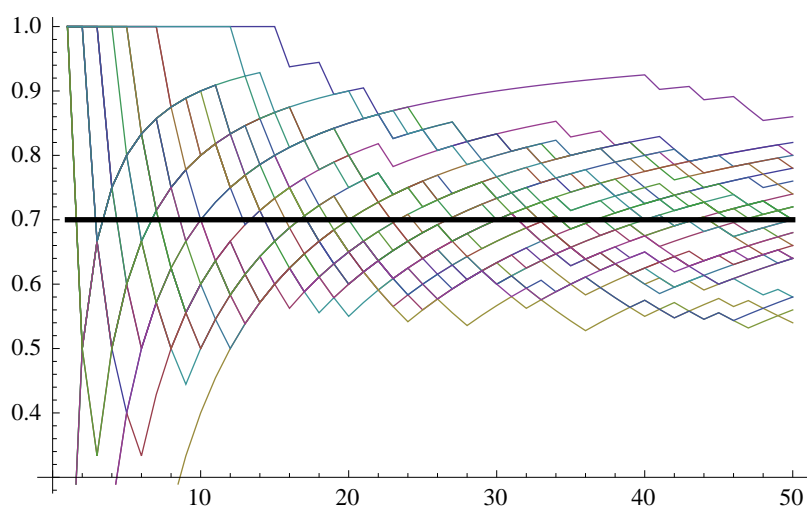
(Das Feld mean enthält m Verläufe der Werte von  $\frac{s_n}{n}$ )

```
Manipulate[Show[
```

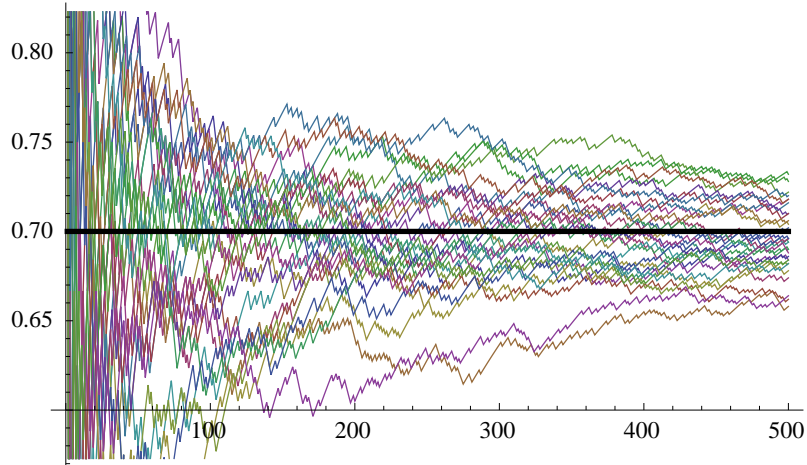
```
  ListLinePlot[Transpose[mean[[1 ;; n]]],
```

```
  ListLinePlot[ConstantArray[p, n], PlotStyle -> {Black, Thick}], {{n, 50}, 1, nmax, 1}]
```

- $n = 50$ :



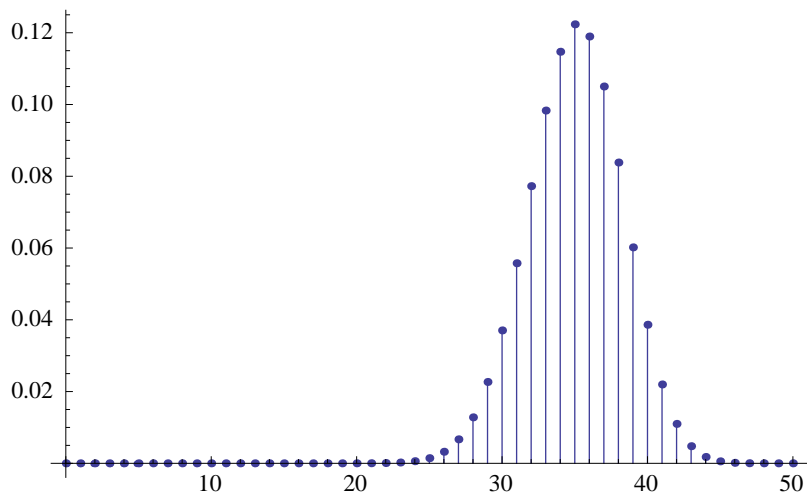
•  $n = 500$ :



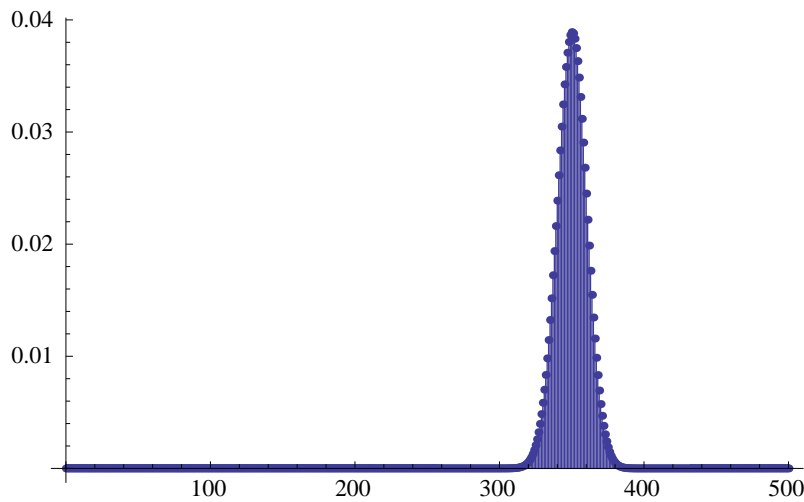
VERTEILUNG VON  $S_n$

```
Manipulate[
  ListPlot[Table[{k, PDF[BinomialDistribution[n, p], k]}, {k, 0, n}],
    PlotRange -> All, Filling -> Axis]
  , {{n, 50}, 1, nmax, 1}]
```

•  $n = 50$ :



- $n = 500$ :



## 2.4 Unabhängige Zufallsvariablen und Random Walk

### Unabhängigkeit von diskreten Zufallsvariablen

Seien  $X_i : \Omega \rightarrow S_i$ ,  $i = 1, \dots, n$ , diskrete Zufallsvariablen auf dem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ . Dann ist  $(X_1, \dots, X_n)$  eine Zufallsvariable mit Werten im Produktraum  $S_1 \times \dots \times S_n$ .

**Definition.** Die Verteilung  $\mu_{X_1, \dots, X_n}$  des Zufallsvektors  $(X_1, \dots, X_n)$  heißt **gemeinsame Verteilung** der Zufallsvariablen  $X_1, \dots, X_n$ . Die Massenfunktion der gemeinsamen Verteilung lautet

$$p_{X_1, \dots, X_n}(a_1, \dots, a_n) = P[X_1 = a_1, \dots, X_n = a_n].$$

**Definition.** Die diskreten Zufallsvariablen  $X_1, \dots, X_n$  heißen **unabhängig**, falls gilt:

$$P[X_1 = a_1, \dots, X_n = a_n] = \prod_{i=1}^n P[X_i = a_i] \quad \text{für alle } a_i \in S_i, \quad i = 1, \dots, n.$$

Die gemeinsame Verteilung enthält Informationen über den Zusammenhang zwischen den Zufallsgrößen  $X_j$ .

**Satz 2.7.** Die folgenden Aussagen sind äquivalent:

- (i)  $X_1, \dots, X_n$  sind unabhängig.
- (ii)  $p_{X_1, \dots, X_n}(a_1, \dots, a_n) = \prod_{i=1}^n p_{X_i}(a_i)$ .
- (iii)  $\mu_{X_1, \dots, X_n} = \otimes_{i=1}^n \mu_{X_i}$ .

(iv) Die Ereignisse  $\{X_1 \in A_1\}, \dots, \{X_n \in A_n\}$  sind unabhängig für alle  $A_i \subseteq S_i$ ,  $i = 1, \dots, n$ .

(v) Die Ereignisse  $\{X_1 = a_1\}, \dots, \{X_n = a_n\}$  sind unabhängig für alle  $a_i \in S_i$ ,  $i = 1, \dots, n$ .

*Beweis.*

• (i)  $\Leftrightarrow$  (ii) nach Definition von  $p_{X_1, \dots, X_n}$ .

• (ii)  $\Leftrightarrow$  (iii) nach Definition von  $\bigotimes_{i=1}^n \mu_{X_i}$ .

• (iii)  $\Rightarrow$  (iv):

Seien  $1 \leq i_1 < i_2 < \dots < i_m \leq n$  und  $A_{i_k} \subseteq S_{i_k}$ , ( $k = 1, \dots, m$ ). Wir setzen  $A_i := \Omega$  für  $i \notin \{i_1, \dots, i_m\}$ . Mit (iii) folgt dann nach Satz 2.2:

$$\begin{aligned} P[X_{i_1} \in A_{i_1}, \dots, X_{i_m} \in A_{i_m}] &= P[X_1 \in A_1, \dots, X_n \in A_n] \\ &= P[(X_1, \dots, X_n) \in A_1 \times \dots \times A_n] \\ &= \mu_{X_1, \dots, X_n}(A_1 \times \dots \times A_n) \\ &= \prod_{i=1}^n \mu_{X_i}(A_i) = \prod_{i=1}^n P[X_i \in A_i] \\ &= \prod_{i=1}^m P[X_{i_k} \in A_{i_k}]. \end{aligned}$$

• (iv)  $\Rightarrow$  (v)  $\Rightarrow$  (i) ist klar.

□

**Definition.** Eine beliebige Kollektion  $X_i: \Omega \rightarrow S_i$ ,  $i \in I$ , von diskreten Zufallsvariablen heißt **unabhängig**, falls die Ereignisse  $\{X_i = a_i\}$ ,  $i \in I$ , für alle  $a_i \in S_i$  unabhängig sind.

## Der Random Walk auf $\mathbb{Z}$

Seien  $X_1, X_2, \dots$  unabhängige identisch verteilte ( $\gg$ i.i.d. $\ll$  – independent and identically distributed) Zufallsvariablen auf dem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  mit

$$P[X_i = +1] = p, \quad P[X_i = -1] = 1 - p, \quad p \in (0, 1).$$



Die Existenz von unendlich vielen unabhängigen identisch verteilten Zufallsvariablen auf einem geeigneten Wahrscheinlichkeitsraum (unendliches Produktmodell) wird in der Vorlesung »Einführung in die Wahrscheinlichkeitstheorie« gezeigt. Sei  $a \in \mathbb{Z}$  ein fester Startwert. Wir betrachten die durch

$$\begin{aligned} S_0 &= a, \\ S_{n+1} &= S_n + X_{n+1}, \end{aligned}$$

definierte zufällige Bewegung (»Irrfahrt« oder »Random Walk«) auf  $\mathbb{Z}$ . Als Position zur Zeit  $n$  ergibt sich:

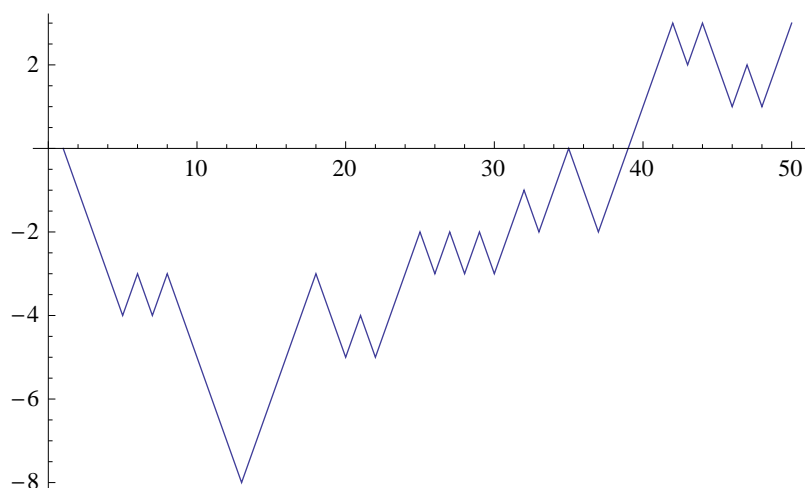
$$S_n = a + X_1 + X_2 + \cdots + X_n.$$

Irrfahrten werden unter anderem in primitiven Modellen für die Kapitalentwicklung beim Glücksspiel oder an der Börse (Aktienkurs), sowie die Brownsche Molekularbewegung (im Skalierungslimes Schrittweite  $\rightarrow 0$ ) eingesetzt.

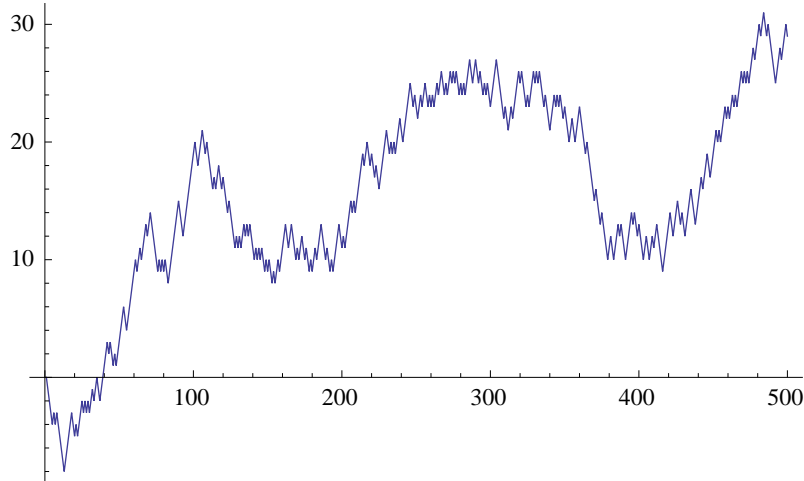
**Beispiel** (Symmetrischer Random Walk,  $p = \frac{1}{2}$ ).

```
zufall = RandomChoice[{-1, 1}, 10000];
randomwalk = FoldList[Plus, 0, zufall];
Manipulate[
  ListLinePlot[randomwalk[[1 ;; nmax]]], {nmax, 10, 10000, 10}]
```

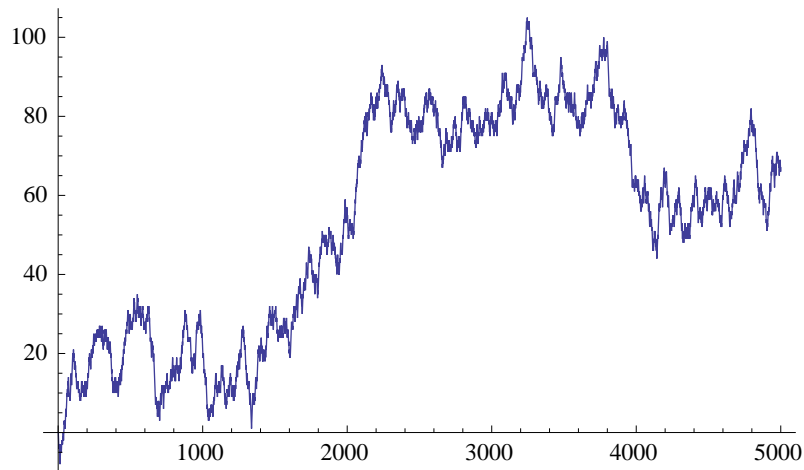
•  $nmax = 50$ :



- $n_{max} = 500$ :



- $n_{max} = 5000$ :



**Lemma 2.8** (Verteilung von  $S_n$ ). Für  $k \in \mathbb{Z}$  gilt

$$P[S_n = a + k] = \begin{cases} 0 & \text{falls } n + k \text{ ungerade oder } |k| > n, \\ \binom{n}{\frac{n+k}{2}} p^{\frac{n+k}{2}} (1-p)^{\frac{n-k}{2}} & \text{sonst.} \end{cases}$$

*Beweis.* Es gilt:

$$S_n = a + k \Leftrightarrow X_1 + \dots + X_n = k \Leftrightarrow \begin{cases} X_i = 1 & \text{genau } \frac{n+k}{2}\text{-mal,} \\ X_i = -1 & \text{genau } \frac{n-k}{2}\text{-mal.} \end{cases}$$

□

**Beispiel** (Rückkehrwahrscheinlichkeit zum Startpunkt). Mithilfe der **Stirlingschen Formel**

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \quad \text{für } n \rightarrow \infty.$$

folgt:

$$\begin{aligned} P[S_{2n+1} = a] &= 0, \\ P[S_{2n} = a] &= \binom{2n}{n} p^n (1-p)^n = \frac{(2n)!}{(n!)^2} p^n (1-p)^n \\ &\sim \frac{\sqrt{4\pi n}}{2\pi n} \frac{\left(\frac{2n}{e}\right)^{2n}}{\left(\frac{n}{e}\right)^{2n}} p^n (1-p)^n = \frac{1}{\sqrt{\pi n}} (4p(1-p))^n \quad \text{für } n \rightarrow \infty, \end{aligned}$$

wobei zwei Folgen  $a_n$  und  $b_n$  **asymptotisch äquivalent** heißen ( $a_n \sim b_n$ ), falls  $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 1$  gilt.

- Falls  $p \neq \frac{1}{2}$  gilt  $4p(1-p) < 1$  und  $P[S_{2n} = a]$  konvergiert exponentiell schnell gegen 0.
- Falls  $p = \frac{1}{2}$  konvergiert  $P[S_{2n} = a] \sim \frac{1}{\sqrt{\pi n}}$  nur langsam gegen 0.

### Symmetrischer Random Walk

Ab jetzt betrachten wir den **symmetrischen** Random Walk, d.h.  $p = \frac{1}{2}$ .

Sei  $\lambda \in \mathbb{Z}$ . Wir wollen die Verteilung der Zufallsvariable

$$T_\lambda(\omega) := \inf\{n \in \mathbb{N} \mid S_n(\omega) = \lambda\}, \quad (\min \emptyset := \infty),$$

bestimmen. Für  $\lambda \neq a$  ist  $T_\lambda$  die erste **Trefferzeit von**  $\lambda$ , für  $\lambda = a$  ist es die erste **Rückkehrzeit nach**  $a$ . Beschreibt der Random Walk beispielsweise die Kapitalentwicklung in einem Glücksspiel, dann kann man  $T_0$  als Ruinzeitpunkt interpretieren.

Sei  $n \in \mathbb{N}$ . Wir wollen nun die Wahrscheinlichkeit

$$P[T_\lambda \leq n] = P\left[\bigcup_{i=1}^n \{S_i = \lambda\}\right]$$

berechnen. Da das Ereignis  $\{T_\lambda \leq n\}$  von **mehreren** Positionen des Random Walks abhängt ( $S_1, S_2, \dots, S_n$ ), benötigen wir die **gemeinsame** Verteilung dieser Zufallsvariablen. Sei also

$$S(\omega) := (S_0(\omega), S_1(\omega), \dots, S_n(\omega))$$

der **Bewegungsverlauf bis zur Zeit**  $n$ . Dann ist  $S$  eine Zufallsvariable mit Werten im Raum

$$\widehat{\Omega}_a^{(n)} := \{(s_0, s_1, \dots, s_n) \mid s_0 = a, s_i \in \mathbb{Z}, \text{ so dass: } |s_i - s_{i-1}| = 1 \text{ für alle } i \in \{1, \dots, n\}\}$$

der möglichen Pfade des Random Walk. Sei  $\mu_a$  die gemeinsame Verteilung von  $S$  unter  $P$ .

**Lemma 2.9.**  $\mu_a$  ist die **Gleichverteilung** auf dem Pfadraum  $\widehat{\Omega}_a^{(n)}$ .

*Beweis.* Es gilt

$$\begin{aligned} \mu_a(\{(s_0, \dots, s_n)\}) &= P[S_0 = s_0, \dots, S_n = s_n] \\ &= P[S_0 = s_0, X_1 = s_1 - s_0, \dots, X_n = s_n - s_{n-1}] \\ &= \begin{cases} 0 & \text{falls } s_0 \neq a \text{ oder } |s_i - s_{i-1}| \neq 1 \text{ f\"ur ein } i \in \{1, \dots, n\}, \\ & \text{(d.h. } (s_0, \dots, s_n) \notin \widehat{\Omega}_a^{(n)}), \\ 2^{-n} & \text{sonst, d.h. falls } (s_0, \dots, s_n) \in \widehat{\Omega}_a^{(n)}. \end{cases} \end{aligned}$$

□

**Satz 2.10** (Reflektionsprinzip). *Seien  $\lambda, b \in \mathbb{Z}$ . Es gelte entweder ( $a < \lambda$  und  $b \leq \lambda$ ), oder ( $a > \lambda$  und  $b \geq \lambda$ ). Dann gilt:*

$$P[T_\lambda \leq n, S_n = b] = P[S_n = b^*],$$

wobei  $b^* := \lambda + (\lambda - b) = 2\lambda - b$  die **Spiegelung** von  $b$  an  $\lambda$  ist.

*Beweis.* Es gilt:

$$\begin{aligned} P[T_\lambda \leq n, S_n = b] &= \mu_a[\overbrace{\{(s_0, \dots, s_n) \mid s_n = b, s_i = \lambda \text{ f\"ur ein } i \in \{1, \dots, n\}\}}^{=:A}], \\ P[S_n = b^*] &= \mu_a[\underbrace{\{(s_0, \dots, s_n) \mid s_n = b^*\}}_{=:B}]. \end{aligned}$$

Die im Bild dargestellte Transformation (Reflektion des Pfades nach Treffen von  $\lambda$ ) definiert eine Bijektion von  $A$  nach  $B$ . Also gilt  $|A| = |B|$ . Da  $\mu_a$  die Gleichverteilung auf  $\widehat{\Omega}_a^{(n)}$  ist, folgt:

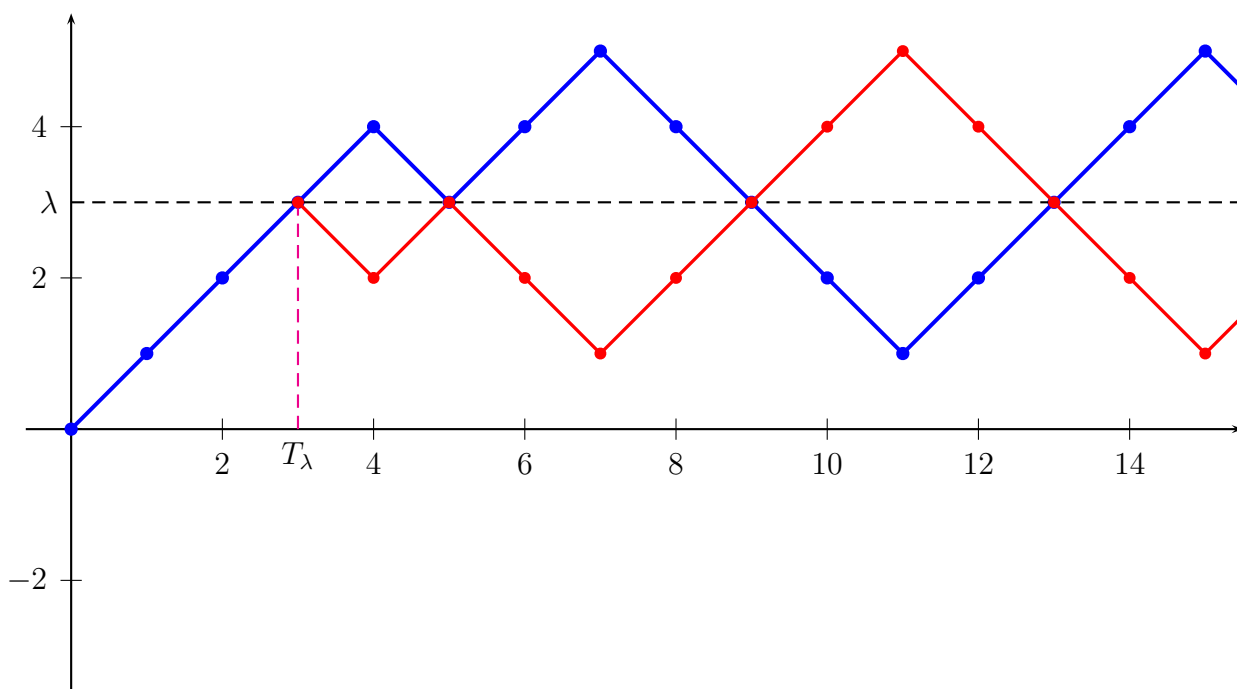
$$\mu_a(A) = \frac{|A|}{|\widehat{\Omega}_a^{(n)}|} = \frac{|B|}{|\widehat{\Omega}_a^{(n)}|} = \mu_a(B).$$

□

**Korollar** (Verteilung der Trefferzeiten). *Es gilt:*

i)

$$P[T_\lambda \leq n] = \begin{cases} P[S_n \geq \lambda] + P[S_n > \lambda], & \text{falls } \lambda > a, \\ P[S_n \leq \lambda] + P[S_n < \lambda], & \text{falls } \lambda < a. \end{cases}$$

Abbildung 2.2: Spiegelung des Random Walks an  $\lambda = 3$ 

ii)

$$\begin{aligned}
 P[T_\lambda = n] &= \begin{cases} \frac{1}{2}P[S_{n-1} = \lambda - 1] - \frac{1}{2}P[S_{n-1} = \lambda + 1], & \text{falls } \lambda > a, \\ \frac{1}{2}P[S_{n-1} = \lambda + 1] - \frac{1}{2}P[S_{n-1} = \lambda - 1], & \text{falls } \lambda < a. \end{cases} \\
 &= \begin{cases} \frac{\lambda - a}{n} \binom{n}{\frac{n + \lambda - a}{2}} 2^{-n} & \text{falls } \lambda > a, \\ \frac{a - \lambda}{n} \binom{n}{\frac{n + \lambda - a}{2}} 2^{-n} & \text{falls } \lambda < a. \end{cases}
 \end{aligned}$$

*Beweis.* Wir beweisen die Aussagen für  $\lambda > a$ , der andere Fall wird jeweils analog gezeigt.

i)

$$\begin{aligned}
 P[T_\lambda \leq n] &= \sum_{b \in \mathbb{Z}} \underbrace{P[T_\lambda \leq n, S_n = b]} \\
 &= \begin{cases} P[S_n = b] & \text{falls } b \geq \lambda, \\ P[S_n = b^*] & \text{falls } b < \lambda. \end{cases} \\
 &= \sum_{b \geq \lambda} P[S_n = b] + \underbrace{\sum_{b < \lambda} P[S_n = b^*]}_{= \sum_{b > \lambda} P[S_n = b]} = P[S_n \geq \lambda] + P[S_n > \lambda].
 \end{aligned}$$

ii)

$$P[T_\lambda = n] = P[T_\lambda \leq n] - P[T_\lambda \leq n - 1]$$

Mit i) folgt

$$= \underbrace{P[S_n \geq \lambda]}_{=:A} - \underbrace{P[S_{n-1} \geq \lambda]}_{=:B} + \underbrace{P[S_n \geq \lambda + 1]}_{=:II} - \underbrace{P[S_{n-1} \geq \lambda + 1]}_{=:I}$$

Wegen

$$P[A] - P[B] = P[A \setminus B] + P[A \cap B] - P[B \setminus A] - P[B \cap A] = P[A \setminus B] - P[B \setminus A]$$

erhalten wir für den ersten Term:

$$\begin{aligned} \mathbf{I} &= P[S_n \geq \lambda, S_{n-1} < \lambda] - P[S_{n-1} \geq \lambda, S_n < \lambda] \\ &= P[S_{n-1} = \lambda - 1, S_n = \lambda] - P[S_{n-1} = \lambda, S_n = \lambda - 1] \\ &= \frac{1}{2}P[S_{n-1} = \lambda - 1] - \frac{1}{2}P[S_{n-1} = \lambda]. \end{aligned}$$

Hierbei haben wir benutzt, dass

$$\begin{aligned} &|\{(s_0, \dots, s_n) \in \widehat{\Omega}_a^{(n)} \mid s_{n-1} = \lambda - 1\}| \\ &= |\{(s_0, \dots, s_n) \mid s_{n-1} = \lambda - 1 \text{ und } s_n = \lambda\}| \\ &\quad + |\{(s_0, \dots, s_n) \mid s_{n-1} = \lambda - 1 \text{ und } s_n = \lambda - 2\}| \\ &= 2 \cdot |\{(s_0, \dots, s_n) \mid s_{n-1} = \lambda - 1, s_n = \lambda\}| \end{aligned}$$

gilt. Mit einer analogen Berechnung für den zweiten Term erhalten wir insgesamt:

$$\begin{aligned} P[T_\lambda = n] &= \mathbf{I} + \mathbf{II} \\ &= \frac{1}{2} (P[S_{n-1} = \lambda - 1] - P[S_{n-1} = \lambda] \\ &\quad + P[S_{n-1} = (\lambda + 1) - 1] - P[S_{n-1} = \lambda + 1]) \\ &= \frac{1}{2} (P[S_{n-1} = \lambda - 1] - P[S_{n-1} = \lambda + 1]). \end{aligned}$$

□

Sei  $M_n := \max(S_0, S_1, \dots, S_n)$ .**Korollar** (Verteilung des Maximums). Für  $\lambda > a$  gilt:

$$P[M_n \geq \lambda] = P[T_\lambda \leq n] = P[S_n \geq \lambda] + P[S_n > \lambda].$$

## 2.5 Simulationsverfahren

Die Simulation von Stichproben verschiedener Wahrscheinlichkeitsverteilungen geht von auf  $[0, 1]$  gleichverteilten Pseudo-Zufallszahlen aus. In Wirklichkeit simuliert ein Zufallszahlengenerator natürlich nur auf  $\{k m^{-1} \mid k = 0, 1, \dots, m - 1\}$  gleichverteilte Zufallszahlen, wobei  $m^{-1}$  die Darstellungsgenauigkeit des Computers ist. Dieser Aspekt wird im folgenden ignoriert. Um Simulationsverfahren zu analysieren, benötigen wir noch den Begriff einer auf  $[0, 1]$  gleichverteilten reellwertigen Zufallsvariablen. Die Existenz solcher Zufallsvariablen auf einem geeigneten Wahrscheinlichkeitsraum wird hier vorausgesetzt, und kann erst in der Vorlesung »Analysis III« bzw. in der »Einführung in die Wahrscheinlichkeitstheorie« gezeigt werden.

Sei  $(\Omega, \mathcal{A}, P)$  ein Wahrscheinlichkeitsraum, und  $U: \Omega \rightarrow [0, 1]$  eine Abbildung.

**Definition.** i)  $U$  ist eine **reellwertige Zufallsvariable**, falls gilt:

$$\{\omega \in \Omega \mid U(\omega) \leq y\} \in \mathcal{A} \quad \text{für alle } y \in \mathbb{R}.$$

ii) Eine reellwertige Zufallsvariable  $U: \Omega \rightarrow [0, 1]$  ist **gleichverteilt auf**  $[0, 1]$ , falls

$$P[U \leq y] = y \quad \text{für alle } y \in [0, 1].$$

Wir notieren dies im folgenden als  $(U \sim \text{Unif}[0, 1])$ .

iii) Reellwertige Zufallsvariablen  $U_i: \Omega \rightarrow \mathbb{R}$ ,  $i \in I$ , heißen **unabhängig**, falls die Ereignisse  $\{U_i \leq y_i\}$ ,  $i \in I$ , für alle  $y_i \in \mathbb{R}$  unabhängig sind.

Ein **Zufallszahlengenerator** simuliert Stichproben  $u_1 = U_1(\omega)$ ,  $u_2 = U_2(\omega)$ , ... von auf  $[0, 1]$  gleichverteilten unabhängigen Zufallsvariablen. Wie erzeugt man daraus Stichproben von diskreten Verteilungen?

### Das direkte Verfahren

Sei  $S = \{a_1, a_2, \dots\}$  endlich oder abzählbar unendlich, und  $\mu$  eine Wahrscheinlichkeitsverteilung auf  $S$  mit Gewichten  $p_i = p(a_i)$ . Wir setzen

$$s_n := \sum_{i=1}^n p_i, \quad n \in \mathbb{N}, \quad \text{»kumulative Verteilungsfunktion«.}$$

Sei  $U: \Omega \rightarrow [0, 1)$  eine gleichverteilte Zufallsvariable. Wir setzen

$$X(\omega) := a_i, \quad \text{falls } s_{i-1} < U(\omega) \leq s_i, \quad i \in \mathbb{N}.$$

**Lemma 2.11.** Falls  $U \sim \text{Unif}[0, 1)$ , gilt  $X \sim \mu$ .

*Beweis.* Für alle  $i \in \mathbb{N}$  gilt:

$$P[X = a_i] = P[s_{i-1} < U \leq s_i] = P[U \leq s_i] - P[U \leq s_{i-1}] = s_i - s_{i-1} = p_i.$$

□

**Algorithmus 2.12** (Direkte Simulation einer diskreten Verteilung).

INPUT: Gewichte  $p_1, p_2, \dots$ ,

OUTPUT: Pseudozufallsstichprobe  $x$  von  $\mu$ .

$n := 1$

$s := p_1$

erzeuge Zufallszahl  $u \sim \text{Unif}[0, 1)$

**while**  $u > s$  **do**

$n := n + 1$

$s := s + p_n$

**end while**

**return**  $x := a_n$

**Bemerkung.** a) Die mittlere Anzahl von Schritten des Algorithmus ist

$$\sum_{n=1}^{\infty} n p_n = \text{Erwartungswert von Wahrscheinlichkeitsverteilung } (p_n) \text{ auf } \mathbb{N}.$$

b) Für große Zustandsräume  $S$  ist das direkte Verfahren oft nicht praktikabel, siehe Übung.

## Acceptance-Rejection-Verfahren

Sei  $S$  eine endliche oder abzählbare Menge,  $\mu$  eine Wahrscheinlichkeitsverteilung auf  $S$  mit Massenfunktion  $p(x)$ , und  $\nu$  eine Wahrscheinlichkeitsverteilung auf  $S$  mit Massenfunktion  $q(x)$ . Angenommen, wir können unabhängige Stichproben von  $\nu$  erzeugen. Wie können wir daraus Stichproben von  $\mu$  erhalten? IDEE: Erzeuge Stichprobe  $x$  von  $\nu$ , akzeptiere diese mit Wahrscheinlichkeit proportional zu  $\frac{p(x)}{q(x)}$ , sonst verwerfe die Stichprobe und wiederhole.

ANNAHME:

$$\text{es gibt ein } c \in [1, \infty) : \quad p(x) \leq c q(x) \quad \text{für alle } x \in S.$$

Aus der Annahme folgt:

$$\frac{p(x)}{c q(x)} \leq 1 \quad \text{für alle } x \in S,$$

d.h. wir können  $\frac{p(x)}{c q(x)}$  als **Akzeptanzwahrscheinlichkeit** wählen.



**Algorithmus 2.13** (Acceptance-Rejection-Verfahren).

INPUT: Gewichte  $p(y), q(y), c$  ( $y \in S$ ),

OUTPUT: Stichprobe  $x$  von  $\mu$ .

**repeat**

    erzeuge Stichprobe  $x \sim \nu$

    erzeuge Stichprobe  $u \sim \text{Unif}[0, 1]$

**until**  $\frac{p(x)}{cq(x)} \geq u$  { akzeptiere mit Wahrscheinlichkeit  $\frac{p(x)}{cq(x)}$  }

**return**  $x$

ANALYSE DES ALGORITHMUS

Für die verwendeten Zufallsvariablen gilt:

$$X_1, X_2, \dots \sim \nu, \quad (\text{Vorschläge}),$$

$$U_1, U_2, \dots \sim \text{Unif}[0, 1].$$

Es gilt Unabhängigkeit, d.h.

$$P[X_1 = a_1, \dots, X_n = a_n, U_1 \leq y_1, \dots, U_n \leq y_n] = \prod_{i=1}^n P[X_i = a_i] \cdot \prod_{i=1}^n P[U_i \leq y_i]$$

für alle  $n \in \mathbb{N}$ ,  $a_i \in S$  und  $y_i \in \mathbb{R}$ .

Seien

$$\begin{aligned} T &= \min \left\{ n \in \mathbb{N} \mid \frac{p(X_n)}{cq(X_n)} \geq U_n \right\} && \text{die »Akzeptanzzeit«, und} \\ X_T(\omega) &= X_{T(\omega)}(\omega) && \text{die ausgegebene Stichprobe.} \end{aligned}$$

des Acceptance-Rejection-Verfahrens. Wir erhalten:

**Satz 2.14.** i)  $T$  ist *geometrisch verteilt* mit Parameter  $1/c$ ,

ii)  $X_T \sim \mu$ .

**Bemerkung.** Insbesondere ist die mittlere Anzahl von Schritten bis Akzeptanz:

$$E[T] = c.$$

*Beweis.* i) Sei

$$A_n := \left\{ \frac{p(X_n)}{cq(X_n)} \geq U_n \right\}.$$

Aus der Unabhängigkeit der Zufallsvariablen  $X_1, U_1, X_2, U_2, \dots$  folgt, daß auch die Ereignisse  $A_1, A_2, \dots$  unabhängig sind. Dies wird in der Vorlesung »Einführung in die Wahrscheinlichkeitstheorie« bewiesen. Zudem gilt wegen der Unabhängigkeit von  $X_n$  und  $U_n$ :

$$\begin{aligned} P[A_n] &= \sum_{a \in S} P \left[ \left\{ U_n \leq \frac{p(a)}{c q(a)} \right\} \cap \{X_n = a\} \right] \\ &= \sum_{a \in S} P \left[ \left\{ U_n \leq \frac{p(a)}{c q(a)} \right\} \right] \cdot P[X_n = a] \\ &= \sum_{a \in S} \frac{p(a)}{c q(a)} \cdot q(a) = \frac{1}{c}. \end{aligned}$$

Also ist

$$T(\omega) = \min\{n \in \mathbb{N} \mid \omega \in A_n\}$$

geometrisch verteilt mit Parameter  $1/c$ .

ii)

$$\begin{aligned} P[X_T = a] &= \sum_{n=1}^{\infty} P[\{X_T = a\} \cap \{T = n\}] \\ &= \sum_{n=1}^{\infty} P[\{X_n = a\} \cap A_n \cap A_1^C \cap \dots \cap A_{n-1}^C] \\ &= \sum_{n=1}^{\infty} P[\{X_n = a\} \cap \left\{ \frac{p(a)}{c q(a)} \geq U_n \right\} \cap A_1^C \cap \dots \cap A_{n-1}^C] \\ &= \sum_{n=1}^{\infty} q(a) \frac{p(a)}{c q(a)} \left(1 - \frac{1}{c}\right)^{n-1} \\ &= \frac{p(a)}{c} \sum_{n=1}^{\infty} \left(1 - \frac{1}{c}\right)^{n-1} \\ &= \frac{p(a)}{c} \frac{1}{1 - (1 - \frac{1}{c})} = p(a). \end{aligned}$$

□

# Kapitel 3

## Konvergenzsätze und Monte Carlo Verfahren

Sei  $\mu$  eine Wahrscheinlichkeitsverteilung auf einer abzählbaren Menge  $S$ , und  $f : S \rightarrow \mathbb{R}$  eine reellwertige Zufallsvariable. Angenommen, wir wollen den Erwartungswert

$$\theta := E_{\mu}[f] = \sum_{x \in S} f(x) \mu(x)$$

berechnen, aber die Menge  $S$  ist zu groß, um die Summe direkt auszuführen. In einem **Monte Carlo-Verfahren** simuliert man eine große Anzahl unabhängiger Stichproben  $X_1(\omega), \dots, X_n(\omega)$  von  $\mu$ , und approximiert den Erwartungswert  $\theta$  durch den **Monte Carlo-Schätzer**

$$\hat{\theta}_n(\omega) := \frac{1}{n} \sum_{i=1}^n f(X_i(\omega)).$$

Wir wollen nun Methoden entwickeln, mit denen der Approximationsfehler  $|\hat{\theta}_n - \theta|$  abgeschätzt werden kann, und die Asymptotik des Approximationsfehlers für  $n \rightarrow \infty$  untersuchen. Nach dem Transformationssatz (1.7) und der Linearität des Erwartungswerts (1.8) gilt:

$$E[\hat{\theta}_n] = \frac{1}{n} \sum_{i=1}^n E[f(X_i)] = \frac{1}{n} \sum_{i=1}^n \sum_{x \in S} f(x) \mu(\{x\}) = E_{\mu}[f] = \theta,$$

d.h.  $\hat{\theta}_n$  ist ein **erwartungstreuer** Schätzer für  $\theta$ . Der mittlere quadratische Fehler (»MSE« – mean squared error) des Schätzers ist daher:

$$\text{MSE} = E[|\hat{\theta}_n - \theta|^2] = E[|\hat{\theta}_n - E[\hat{\theta}_n]|^2].$$

### 3.1 Varianz und Kovarianz

Sei  $(\Omega, \mathcal{A}, P)$  ein Wahrscheinlichkeitsraum und  $X : \Omega \rightarrow S$  eine Zufallsvariable auf  $(\Omega, \mathcal{A}, P)$ , so dass  $E[|X|]$  endlich ist.

**Definition.**

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

heißt **Varianz** von  $X$  und liegt in  $[0, \infty]$ .

$$\sigma(X) := \sqrt{\text{Var}(X)}$$

heißt **Standardabweichung** von  $X$ .

Die Varianz bzw. Standardabweichung kann als Kennzahl für die Größe der Fluktuationen (Streuung) der Zufallsvariablen  $X$  um den Erwartungswert  $E[X]$  und damit als Maß für das Risiko bei Prognose des Ausgangs  $X(\omega)$  durch  $E[X]$  interpretiert werden.

**Bemerkung.** (a) Die Varianz hängt nur von der Verteilung von  $X$  ab:

$$\text{Var}(X) = \sum_{a \in S} (a - m)^2 p_X(a), \quad \text{wobei} \quad m = E[X] = \sum_{a \in S} a p_X(a).$$

(b) Es gilt

$$\text{Var}(X) = 0 \quad \text{genau dann, wenn} \quad P[X = E[X]] = 1.$$

**Bemerkung** (Rechenregeln). i)

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

Insbesondere ist die Varianz von  $X$  genau dann endlich, wenn  $E[X^2]$  endlich ist.

ii)

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

**Beweis.** i) Nach der Linearität des Erwartungswerts gilt

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

ii) Wiederholte Anwendung der Linearität des Erwartungswerts liefert

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

□

**Beispiele.** a) Sei  $X = 1$  mit Wahrscheinlichkeit  $p$  und  $X = 0$  mit Wahrscheinlichkeit  $1 - p$ . Dann ist der Erwartungswert von  $X$ :

$$E[X^2] = E[X] = p,$$

und die Varianz von  $X$ :

$$\text{Var}(X) = p - p^2 = p(1 - p).$$

b) Sei  $T$  geometrisch verteilt ( $T \sim \text{Geom}(p)$ ) mit Parameter  $p \in (0, 1]$ . Der Erwartungswert von  $T$  beträgt:

$$E[T] = \sum_{k=1}^{\infty} k (1-p)^{k-1} p = -p \frac{d}{dp} \sum_{k=0}^{\infty} (1-p)^k = -p \frac{d}{dp} \frac{1}{p} = \frac{1}{p}.$$

Außerdem gilt:

$$\begin{aligned} E[T(T+1)] &= \sum_{k=1}^{\infty} k(k+1)(1-p)^{k-1} p \\ &= \sum_{k=1}^{\infty} k(k-1)(1-p)^{k-2} p = p \frac{d^2}{dp^2} \sum_{k=0}^{\infty} (1-p)^k = \frac{2}{p^2}. \end{aligned}$$

Die Varianz von  $T$  ist somit:

$$\text{Var}(T) = E[T^2] - E[T]^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

**Definition.**

$$\mathcal{L}^2(\Omega, \mathcal{A}, P) := \{X: \Omega \rightarrow \mathbb{R} \mid X \text{ ist diskrete Zufallsvariable mit } E[X^2] < \infty\}$$

**Lemma 3.1.** i) Für Zufallsvariablen  $X, Y \in \mathcal{L}^2$  gilt:

$$E[|XY|] \leq \sqrt{E[X^2]} \sqrt{E[Y^2]} < \infty.$$

ii)  $\mathcal{L}^2$  ist ein Vektorraum, und

$$(X, Y)_{\mathcal{L}^2} := E[XY]$$

ist eine **positiv semidefinite symmetrische Bilinearform** (»Skalarprodukt«) auf  $\mathcal{L}^2$ .

**Bemerkung.** i) Insbesondere gilt die **Cauchy-Schwarz-Ungleichung**:

$$E[XY]^2 \leq E[|XY|] \leq E[X^2] E[Y^2] \quad \text{für alle } X, Y \in \mathcal{L}^2.$$

ii) Für eine Zufallsvariable  $X \in \mathcal{L}^2$  gilt

$$E[|X|] \leq \sqrt{E[X^2]} \sqrt{E[1^2]} < \infty.$$

*Beweis.* i) Nach der Cauchy-Schwarz-Ungleichung gilt:

$$\begin{aligned} E[|XY|] &= \sum_{\substack{a \in X(\Omega) \\ b \in Y(\Omega)}} |ab| P[X = a, Y = b] \\ &= \sum_{\substack{a \in X(\Omega) \\ b \in Y(\Omega)}} |a| \sqrt{P[X = a, Y = b]} |b| \sqrt{P[X = a, Y = b]} \\ &\leq \sqrt{\sum_{a,b} a^2 P[X = a, Y = b]} \sqrt{\sum_{a,b} b^2 P[X = a, Y = b]} \\ &= \sqrt{\sum_a a^2 P[X = a]} \sqrt{\sum_b b^2 P[Y = b]} \\ &= \sqrt{E[X^2]} \sqrt{E[Y^2]}. \end{aligned}$$

ii) Seien  $X, Y \in \mathcal{L}^2$ ,  $a \in \mathbb{R}$ . Dann ist  $aX + Y$  eine diskrete Zufallsvariable, für die nach Monotonie und der Linearität des Erwartungswerts gilt:

$$E[(aX + Y)^2] = E[a^2 X^2 + 2aXY + Y^2] \leq 2a^2 E[X^2] + 2E[Y^2] < \infty.$$

$(X, Y)_{\mathcal{L}^2} = E[XY]$  ist bilinear, da  $E[\bullet]$  linear und symmetrisch ist, und positiv semidefinit, aufgrund von:

$$(X, X)_{\mathcal{L}^2} = E[X^2] \geq 0 \quad \text{für alle } X \in \mathcal{L}^2.$$

□

**Definition.** Seien  $X, Y \in \mathcal{L}^2$ .

i)

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

heißt **Kovarianz** von  $X$  und  $Y$ .

ii) Gilt  $\sigma(X), \sigma(Y) \neq 0$ , so heißt

$$\varrho(X, Y) := \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)}$$

**Korrelationskoeffizient** von  $X$  und  $Y$ .

iii)  $X$  und  $Y$  heißen **unkorreliert**, falls  $\text{Cov}(X, Y) = 0$ , d.h.

$$E[XY] = E[X] \cdot E[Y].$$

**Bemerkung.**  $\text{Cov} : \mathcal{L}^2 \times \mathcal{L}^2 \rightarrow \mathbb{R}$  ist eine symmetrische Bilinearform mit:

$$\text{Cov}(X, X) = \text{Var}(X) \geq 0 \quad \text{für alle } X \in \mathcal{L}^2.$$

**Satz 3.2** (Zusammenhang von Unabhängigkeit und Unkorreliertheit). Seien  $X : \Omega \rightarrow S$  und  $Y : \Omega \rightarrow T$  diskrete Zufallsvariablen auf  $(\Omega, \mathcal{A}, P)$ . Dann sind äquivalent:

(i)  $X$  und  $Y$  sind unabhängig, d.h.

$$P[X \in A, Y \in B] = P[X \in A] P[Y \in B] \quad \text{für alle } A, B \in \mathcal{A}.$$

(ii)  $f(X)$  und  $g(Y)$  sind unkorreliert für alle Funktionen  $f : S \rightarrow \mathbb{R}$  und  $g : T \rightarrow \mathbb{R}$  mit  $f(X), g(Y) \in \mathcal{L}^2$ .

*Beweis.* • (i)  $\Rightarrow$  (ii): Seien  $X$  und  $Y$  unabhängig, dann gilt:

$$\begin{aligned} E[f(X)g(Y)] &= \sum_{a \in S} \sum_{b \in T} f(a) g(b) P[X = a, Y = b] \\ &= \sum_{a \in S} f(a) P[X = a] \sum_{b \in T} g(b) P[Y = b] = E[f(X)] E[g(Y)] \end{aligned}$$

Somit folgt:

$$\text{Cov}(f(X), g(Y)) = 0.$$

• (ii)  $\Rightarrow$  (i): Aus (ii) folgt für alle  $a \in S, b \in T$ :

$$\begin{aligned} P[X = a, Y = b] &= E[I_{\{a\}}(X) I_{\{b\}}(Y)] \\ &= E[I_{\{a\}}(X)] E[I_{\{b\}}(Y)] = P[X = a] P[Y = b]. \end{aligned}$$

□

**Beispiel.** Sei  $X = +1, 0, -1$  jeweils mit Wahrscheinlichkeit  $\frac{1}{3}$ , und  $Y = X^2$ . Dann sind  $X$  und  $Y$  nicht unabhängig, aber unkorreliert:

$$E[XY] = 0 = E[X] E[Y].$$

**Satz 3.3** (Varianz von Summen). Für  $X_1, \dots, X_n \in \mathcal{L}^2$  gilt:

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

Falls  $X_1, \dots, X_n$  unkorreliert sind, folgt insbesondere:

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i).$$

*Beweis.* Nach Bilinearität der Kovarianz gilt:

$$\begin{aligned} \text{Var}(X_1 + \dots + X_n) &= \text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j\right) \\ &= \sum_{i,j=1}^n \text{Cov}(X_i, X_j) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{\substack{i,j=1 \\ i < j}}^n \text{Cov}(X_i, X_j). \end{aligned}$$

□

**Beispiel** (Varianz der Binomialverteilung). Sei

$$S_n = \sum_{i=1}^n X_i, \quad X_i = \begin{cases} 1 & \text{mit Wahrscheinlichkeit } p, \\ 0 & \text{mit Wahrscheinlichkeit } 1 - p, \end{cases}$$

mit unabhängigen Zufallsvariablen  $X_i$ . Mit Satz 3.2 folgt:

$$\text{Var}(S_n) = \sum_{i=1}^n \text{Var}(X_i) = n p (1 - p).$$

Analog gilt für den Random Walk:

$$\sigma(S_n) = O(\sqrt{n}).$$

## 3.2 Das schwache Gesetz der großen Zahlen

Seien  $X_1, X_2, \dots : \Omega \rightarrow \mathbb{R}$  Zufallsvariablen, die auf einem gemeinsamen Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  definiert sind (z.B. wiederholte Ausführungen desselben Zufallsexperiments), und sei

$$S_n(\omega) = X_1(\omega) + \dots + X_n(\omega).$$



Wir betrachten die empirischen Mittelwerte

$$\frac{S_n(\omega)}{n} = \frac{X_1(\omega) + \dots + X_n(\omega)}{n},$$

d.h. die arithmetischen Mittel der ersten  $n$  Beobachtungswerte  $X_1(\omega), \dots, X_n(\omega)$ . Gesetze der großen Zahlen besagen, dass sich unter geeigneten Voraussetzungen die zufälligen „Fluktuationen“ der  $X_i$  für große  $n$  wegmitteln, d.h. in einem noch zu präzisierenden Sinn gilt

$$\frac{S_n(\omega)}{n} \approx E\left[\frac{S_n}{n}\right] \quad \text{für große } n,$$

bzw.

$$\frac{S_n}{n} - \frac{E[S_n]}{n} \xrightarrow{n \rightarrow \infty} 0.$$

Ist insbesondere  $E[X_i] = m$  für alle  $i$ , dann sollten die empirischen Mittelwerte  $S_n/n$  gegen  $m$  konvergieren. Das folgende einfache Beispiel zeigt, dass wir ohne weitere Voraussetzungen an die Zufallsvariablen  $X_i$  kein Gesetz der großen Zahlen erwarten können.

**Beispiel.** Sind die Zufallsvariablen  $X_i$  alle gleich, d.h.  $X_1 = X_2 = \dots$ , so gilt  $\frac{S_n}{n} = X_1$  für alle  $n$ . Es gibt also kein Wegmitteln des Zufalls, somit kein Gesetz großer Zahlen.

Andererseits erwartet man ein Wegmitteln des Zufalls bei *unabhängigen* Wiederholungen desselben Zufallsexperiments.

Wir werden nun zeigen, dass sogar Unkorreliertheit und beschränkte Varianzen der Zufallsvariablen  $X_i$  genügen, um ein Gesetz der großen Zahlen zu erhalten. Dazu nehmen wir an, dass  $X_1, X_2, \dots$  diskrete Zufallsvariablen aus  $\mathcal{L}^2(\Omega, \mathcal{A}, P)$  sind, die folgende Voraussetzungen erfüllen:

ANNAHMEN:

- (i) Die Zufallsvariablen sind unkorreliert:

$$\text{Cov}(X_i, X_j) = 0 \quad \text{für alle } i \neq j.$$

- (ii) Die Varianzen sind beschränkt:

$$v := \sup_{i \in \mathbb{N}} \text{Var}(X_i) < \infty.$$

Es wird **keine Unabhängigkeit vorausgesetzt!**

**Satz 3.4** (Schwaches Gesetz der großen Zahlen). *Unter den Voraussetzungen (i) und (ii) gilt für alle  $\varepsilon > 0$ :*

$$P \left[ \left| \frac{S_n}{n} - \frac{E[S_n]}{n} \right| \geq \varepsilon \right] \leq \frac{v}{\varepsilon^2 n} \rightarrow 0 \quad \text{für } n \rightarrow \infty.$$

*Gilt außerdem  $E[X_i] = m$  für alle  $i \in \mathbb{N}$ , folgt  $\frac{E[S_n]}{n} = m$  und  $\frac{S_n}{n}$  **konvergiert stochastisch** gegen  $m$ .*

Zum Beweis benötigen wir:

**Lemma 3.5** (Čebyšev-Ungleichung). *Für  $X \in \mathcal{L}^2$  und  $c > 0$  gilt:*

$$P[|X - E[X]| \geq c] \leq \frac{1}{c^2} \text{Var}(X).$$

*Beweis.* Es gilt

$$I_{\{|X - E[X]| \geq c\}} \leq \frac{1}{c^2} (X - E[X])^2$$

$\frac{1}{c^2}(X - E[X])^2$  ist überall nichtnegativ und  $\geq 1$  auf  $\{|X - E[X]| \geq c\}$ . Durch Bilden des Erwartungswerts folgt:

$$P[|X - E[X]| \geq c] = E[I_{\{|X - E[X]| \geq c\}}] \leq E\left[\frac{1}{c^2}(X - E[X])^2\right] = \frac{1}{c^2} E[(X - E[X])^2]$$

□

*Beweis von Satz 3.4.* Nach der Čebyšev-Ungleichung und den Annahmen (i) und (ii) gilt für  $\varepsilon > 0$ :

$$P \left[ \left| \frac{S_n}{n} - \frac{E[S_n]}{n} \right| \geq \varepsilon \right] \leq \frac{1}{\varepsilon^2} \text{Var} \left( \frac{S_n}{n} \right) = \frac{1}{n^2 \varepsilon^2} \text{Var} \left( \sum_{i=1}^n X_i \right) = \frac{1}{n^2 \varepsilon^2} \sum_{i=1}^n \text{Var}(X_i) \leq \frac{v}{n \varepsilon^2}.$$

□

**Bemerkung** (Starkes Gesetz der großen Zahlen).

$$\frac{S_n(\omega)}{n} \rightarrow m \quad \text{mit Wahrscheinlichkeit 1.}$$

Dies wird in der Vorlesung »Einführung in die Wahrscheinlichkeitstheorie« bewiesen.

### 3.3 Monte Carlo-Verfahren

Sei  $S$  eine abzählbare Menge und  $\mu$  eine Wahrscheinlichkeitsverteilung auf  $S$ . Wir bezeichnen im folgenden die Massenfunktion ebenfalls mit  $\mu$ , d.h.

$$\mu(x) := \mu(\{x\}).$$

Sei  $f: S \rightarrow \mathbb{R}$  eine reellwertige Funktion mit:

$$E_\mu[f^2] = \sum_{x \in S} f(x)^2 \mu(x) < \infty.$$

Wir wollen den Erwartungswert

$$\theta := E_\mu[f] = \sum_{x \in S} f(x) \mu(x)$$

nherungsweise berechnen bzw. schtzen. Dazu approximieren wir  $\theta$  durch die **Monte Carlo-Schtzer**

$$\hat{\theta}_n := \frac{1}{n} \sum_{i=1}^n f(X_i), \quad n \in \mathbb{N},$$

wobei  $X_1, X_2, \dots$  unabhngige Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  mit Verteilung  $\mu$  sind. Nach der Abschtzung aus dem Gesetz der groen Zahlen ergibt sich:

**Korollar.**

$$P[|\hat{\theta}_n - \theta| \geq \varepsilon] \leq \frac{1}{n \varepsilon^2} \text{Var}_\mu[f] \longrightarrow 0 \quad \text{fur } n \rightarrow \infty,$$

d.h.  $\hat{\theta}_n$  ist eine **konsistente Schtzfolge** fur  $\theta$ .

*Beweis.* Da die Zufallsvariablen  $X_i$  unabhngig sind, sind  $f(X_i)$ ,  $i \in \mathbb{N}$ , unkorreliert. Zudem gilt

$$\begin{aligned} E[f(X_i)] &= \sum_{x \in S} f(x) \mu(x) = E_\mu[f] = \theta, & \text{und} \\ \text{Var}[f(X_i)] &= \sum_{x \in S} (f(x) - \theta)^2 \mu(x) = \text{Var}_\mu[f] < \infty \end{aligned}$$

nach Voraussetzung. Die Behauptung folgt nun aus Satz 3.4. □

**Bemerkung.** a)  $\hat{\theta}_n$  ist ein **erwartungstreuer Schtzer** fur  $\theta$ :

$$E[\hat{\theta}_n] = \frac{1}{n} \sum_{i=1}^n E[f(X_i)] = E_\mu[f] = \theta.$$

b) Fur den mittleren quadratischen Fehler des Schtzers ergibt sich nach a):

$$E[|\hat{\theta}_n - \theta|^2] = \text{Var}(\hat{\theta}_n) = \frac{1}{n} \text{Var}_\mu[f].$$

Insbesondere gilt:

$$\|\hat{\theta}_n - \theta\|_{\mathcal{L}^2} = \sqrt{E[|\hat{\theta}_n - \theta|^2]} = O(1/\sqrt{n}).$$

**Beispiele.** a) MONTE CARLO-SCHÄTZUNG VON  $\theta = \int_{[0,1]^d} f(x) dx$ :

Das mehrdimensionale Integral ist folgendermaßen definiert:

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

Der Wert von  $\theta$  kann mit dem folgenden Algorithmus geschätzt werden.

erzeuge Pseudozufallszahlen  $u_1, u_2, \dots, u_{nd} \in (0, 1)$

$$x^{(1)} := (u_1, \dots, u_d)$$

$$x^{(2)} := (u_{d+1}, \dots, u_{2d})$$

...

$$x^{(n)} := (u_{(n-1)d+1}, \dots, u_{nd})$$

$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n f(x^{(i)})$  ist Schätzwert für  $\theta$ .

b) MONTE CARLO-SCHÄTZUNG VON WAHRSCHEINLICHKEITEN:

Sei  $S$  abzählbar,  $B \subseteq S$ . Wir suchen:

$$p = \mu(B) = E_\mu[I_B]$$

Ein Monte Carlo-Schätzer ist

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^n I_B(X_i), \quad X_i \text{ unabhängig mit Verteilung } \mu.$$

FEHLERKONTROLLE:

- Mithilfe der Čebyšev-Ungleichung (Lemma 3.5) ergibt sich:

$$P[|\hat{p}_n - p| \geq \varepsilon] \leq \frac{1}{\varepsilon^2} \text{Var}(\hat{p}_n) = \frac{1}{n\varepsilon^2} \text{Var}_\mu(I_B) = \frac{p(1-p)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

Gilt beispielsweise  $n \geq \frac{5}{\varepsilon^2}$ , dann erhalten wir:

$$P[p \notin (\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon)] \leq 5\%, \quad \text{unabhängig von } p,$$

d.h. das zufällige Intervall  $(\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon)$  ist ein **95%-Konfidenzintervall** für den gesuchten Wert  $p$ .

- Mithilfe der Bernstein-Ungleichung (Chernoff-Abschätzung) erhalten wir für  $\delta > 0$  und  $S_n := \sum_{i=1}^n I_B(X_i)$ :

$$P[p \notin (\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon)] = P\left[\left|\frac{1}{n}S_n - p\right| \geq \varepsilon\right] \leq 2e^{-2n\varepsilon^2} \leq \delta, \quad \text{falls } n \geq \frac{\log(2/\delta)}{2\varepsilon^2}.$$

Für kleine  $\delta$  ist die erhaltene Bedingung an  $n$  wesentlich schwächer als eine entsprechende Bedingung, die man durch Anwenden der Čebyšev-Ungleichung erhält. Für den **relativen Schätzfehler**  $(\hat{p}_n - p)/p$  ergibt sich:

$$P[|\hat{p}_n - p| \geq \varepsilon p] \leq 2e^{-2n\varepsilon^2 p^2} \leq \delta, \quad \text{falls } n \geq \frac{\log(2/\delta)}{2\varepsilon^2 p^2}.$$

Die benötigte Anzahl von Stichproben für eine  $(\varepsilon, \delta)$ -Approximation von  $p$  ist also polynomiell in  $\varepsilon$ ,  $\log(1/\delta)$  und  $1/p$ . Mit einer etwas modifizierten Abschätzung kann man statt der Ordnung  $O(\frac{1}{p^2})$  sogar  $O(\frac{1}{p})$  erhalten, siehe Mitzenmacher und Upfal: »Probability and Computing«.

**Beispiel.** Wie viele Stichproben sind nötig, damit der **relative Fehler** mit 95% Wahrscheinlichkeit unterhalb von 10% liegt? Mithilfe der Čebyšev-Ungleichung (Lemma 3.5) ergibt sich:

$$P[|\hat{p}_n - p| \geq 0,1 p] \leq \frac{p(1-p)}{n(0,1p)^2} = \frac{100(1-p)}{np} \leq 0,05, \quad \text{falls } n \geq \frac{2000(1-p)}{p}.$$

So sind zum Beispiel für  $p = 10^{-5}$  ungefähr  $n \approx 2 \cdot 10^8$  Stichproben ausreichend. Dies ist nur eine obere Schranke, aber man kann zeigen, daß die tatsächlich benötigte Stichprobenzahl immer noch sehr groß ist. Für solch kleine Wahrscheinlichkeiten ist das einfache Monte Carlo-Verfahren ineffektiv, da die meisten Summanden von  $\hat{\theta}_n$  dann gleich 0 sind. Wir brauchen daher ein alternatives Schätzverfahren mit geringerer Varianz.

### Varianzreduktion durch Importance Sampling

Sei  $\nu$  eine weitere Wahrscheinlichkeitsverteilung auf  $S$  mit Massenfunktion  $\nu(x) = \nu(\{x\})$ . Es gelte  $\nu(x) > 0$  für alle  $x \in S$ . Dann können wir den gesuchten Wert  $\theta$  auch als Erwartungswert bzgl.  $\nu$  ausdrücken:

$$\theta = E_\mu[f] = \sum_{x \in S} f(x) \mu(x) = \sum_{x \in S} f(x) \frac{\mu(x)}{\nu(x)} \nu(x) = E_\nu[f \varrho],$$

wobei

$$\varrho(x) = \frac{\mu(x)}{\nu(x)}$$

der Quotient der beiden Massenfunktionen ist. Ein alternativer Monte Carlo-Schätzer für  $\theta$  ist daher

$$\tilde{\theta}_n = \frac{1}{n} \sum_{i=1}^n f(Y_i) \varrho(Y_i),$$

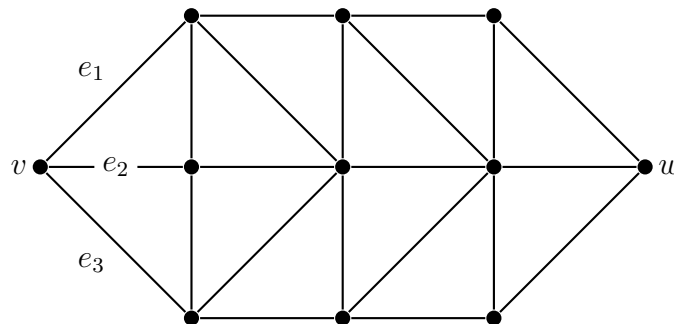


Abbildung 3.1: kleiner Beispielgraph für Perkolation

wobei die  $Y_i$  unabhängige Zufallsvariablen mit Verteilung  $\nu$  sind. Auch  $\tilde{\theta}_n$  ist erwartungstreu:

$$E_\nu[\tilde{\theta}_n] = E_\nu[f \varrho] = \theta.$$

Für die Varianz erhalten wir:

$$\text{Var}_\nu(\tilde{\theta}_n) = \frac{1}{n} \text{Var}_\nu(f \varrho) = \frac{1}{n} \left( \sum_{x \in S} f(x)^2 \varrho(x)^2 \nu(x) - \theta^2 \right).$$

Bei geeigneter Wahl von  $\nu$  kann die Varianz von  $\tilde{\theta}_n$  deutlich kleiner sein als die des Schätzers  $\hat{\theta}_n$ . Faustregel für eine gute Wahl von  $\nu$ :  $\nu(x)$  sollte groß sein, wenn  $|f(x)|$  groß ist.

»Importance Sampling«: Mehr Gewicht für die wichtigen  $x$ !

**Beispiel** (Zuverlässigkeit von Netzwerken; Perkolation). Gegeben sei ein endlicher Graph  $(V, E)$ , wobei  $V$  die Menge der Knoten und  $E$  die Menge der Kanten bezeichnet. Wir nehmen an, dass die Kanten unabhängig voneinander mit Wahrscheinlichkeit  $\varepsilon \ll 1$  ausfallen. Seien  $v, w \in E$  vorgegebene Knoten. Wir wollen die Wahrscheinlichkeit

$$p = P[\text{»}v \text{ nicht verbunden mit } w \text{ durch intakte Kanten«}]$$

approximativ berechnen. Sei

$$S = \{0, 1\}^E = \{(x_e)_{e \in E} \mid x_e \in \{0, 1\}\}$$

die Menge der Konfigurationen von intakten ( $x_l = 0$ ) bzw. defekten ( $x_l = 1$ ) Kanten und  $\mu$  die Wahrscheinlichkeitsverteilung auf  $S$  mit Massenfunktion

$$\mu(x) = \varepsilon^{k(x)} (1 - \varepsilon)^{|E| - k(x)}, \quad k(x) = \sum_{e \in E} x_e.$$

Sei

$$A = \{x \in S \mid v, w \text{ nicht verbunden durch Kanten } e \text{ mit } x_e = 0\}.$$

Dann ist

$$p = \mu(A) = E_\mu[I_A].$$

Der »klassische Monte Carlo-Schätzer« für  $p$  ist

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^n I_A(X_i), \quad X_i \text{ unabhängig mit Verteilung } \mu.$$

Fordern wir nun zum Beispiel

$$\sigma(\hat{p}_n) = \sqrt{\frac{p(1-p)}{n}} \stackrel{!}{\leq} \frac{p}{10},$$

dann benötigen wir eine Stichprobenanzahl

$$n \geq \frac{100(1-p)}{p},$$

um diese Bedingung zu erfüllen. Die Größenordnung von  $p$  für das in der obigen Graphik dargestellte Netzwerk mit  $\varepsilon = 1\%$  lässt sich wie folgt abschätzen:

$$\begin{aligned} 10^{-6} = \mu(\text{»}e_1, e_2, e_3 \text{ versagen«}) &\leq p \leq \mu(\text{»mindestens 3 Kanten versagen«}) \\ &= \binom{22}{3} \cdot 10^{-6} \approx 1,5 \cdot 10^{-3}. \end{aligned}$$

Es sind also eventuell mehrere Millionen Stichproben nötig!

Um die benötigte Stichprobenanzahl zu reduzieren, wenden wir ein Importance Sampling-Verfahren an. Sei

$$\nu(x) = t^{-k(x)} (1-t)^{|E|-k(x)}, \quad k(x) = \sum_{e \in E} x_e,$$

die Verteilung bei Ausfallwahrscheinlichkeit  $t := \frac{3}{22}$ . Da unter  $\nu$  im Schnitt 3 Kanten defekt sind, ist der Ausfall der Verbindung bzgl.  $\nu$  nicht mehr selten. Für den Schätzer

$$\tilde{p}_n = \frac{1}{n} \sum_{i=1}^n I_A(Y_i) \frac{\mu(Y_i)}{\nu(Y_i)}, \quad Y_i \text{ unabhängig mit Verteilung } \nu,$$

erhalten wir im Beispiel von oben:

$$\begin{aligned} \text{Var}(\tilde{p}_n) &= \frac{1}{n} \left( \sum_{x \in S} I_A(x)^2 \frac{\mu(x)^2}{\nu(x)} - p^2 \right) \\ &\leq \frac{1}{n} \sum_{k=3}^{22} \binom{|E|}{k} \left( \frac{\varepsilon^2}{t} \right)^k \left( \frac{(1-\varepsilon)^2}{1-t} \right)^{|E|-k} \leq 0,0053 \frac{p}{n}. \end{aligned}$$

Diese Abschätzung ist etwa um den Faktor 200 besser als die für den einfachen Monte Carlo-Schätzer erhaltene Abschätzung der Varianz.

### 3.4 Gleichgewichte von Markov-Ketten

Sei  $S$  eine abzählbare Menge,  $\nu$  eine Wahrscheinlichkeitsverteilung auf  $S$ , und  $p(x, y)$ , ( $x, y \in S$ ), eine **stochastische Matrix** bzw. **Übergangsmatrix**, d.h.  $p(x, y)$  erfüllt die folgenden Bedingungen:

- (i)  $p(x, y) \geq 0$  für alle  $x, y \in S$ ,
- (ii)  $\sum_{y \in S} p(x, y) = 1$  für alle  $x \in S$ .

Hier und im folgenden bezeichnen wir diskrete Wahrscheinlichkeitsverteilungen und die entsprechenden Massenfunktionen mit demselben Buchstaben, d.h.  $\nu(x) := \nu(\{x\})$ .

**Definition.** Eine Folge  $X_0, X_1, \dots: \Omega \rightarrow S$  von Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  heißt **zeitlich homogene Markov-Kette** mit Startverteilung  $\nu$  und Übergangsmatrix  $p$ , falls die folgenden Bedingungen erfüllt sind:

- (i) Für alle  $x_0 \in S$  gilt:

$$P[X_0 = x_0] = \nu(x_0)$$

- (ii) Für alle  $n \in \mathbb{N}$  und  $x_0, \dots, x_{n+1} \in S$  mit  $P[X_0 = x_0, \dots, X_n = x_n] \neq 0$  gilt:

$$P[X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n] = p(x_n, x_{n+1}).$$

**Bemerkung.** Die Bedingungen (i) und (ii) sind äquivalent zu:

$$P[X_0 = x_0, \dots, X_n = x_n] = \nu(x_0) p(x_0, x_1) \cdots p(x_{n-1}, x_n) \quad \text{für alle } n \in \mathbb{N}, x_i \in S.$$

#### Gleichgewichte und Stationarität

Für eine Wahrscheinlichkeitsverteilung  $\mu$  mit Massenfunktion  $\mu(x) = \mu(\{x\})$  und eine stochastische Matrix  $p$  auf  $S$  setzen wir

$$(\mu p)(y) := \sum_{x \in S} \mu(x) p(x, y), \quad (y \in S),$$

d.h.  $\mu p$  ist der Zeilenvektor, den wir erhalten, wenn wir den Zeilenvektor  $(\mu(x))_{x \in S}$  von links an die Matrix  $p$  multiplizieren.

**Lemma 3.6.** i) Die Verteilung zur Zeit  $n$  einer Markov-Kette mit Startverteilung  $\nu$  und Übergangsmatrix  $p$  ist  $\nu p^n$ .



ii) Gilt  $\nu p = \nu$ , dann folgt  $X_n \sim \nu$  für alle  $n \in \mathbb{N}$ . (»Stationarität«)

*Beweis.* i) Wie im Beweis von Satz 2.4 erhalten wir

$$P[X_n = y \mid X_0 = x] = p^n(x, y)$$

für alle  $n \in \mathbb{N}$  und  $x, y \in S$  mit  $P[X_0 = x] \neq 0$ , und damit:

$$\begin{aligned} P[X_n = y] &= \sum_{\substack{x \in S \\ P[X_0 = x] \neq 0}} P[X_n = y \mid X_0 = x] P[X_0 = x] \\ &= \sum_{\substack{x \in S \\ \nu(x) \neq 0}} p^n(x, y) \nu(x) = (\nu p^n)(y). \end{aligned}$$

ii) Aus  $\nu p = \nu$  folgt  $\nu p^n = \nu$  für alle  $n \in \mathbb{N}$ . □

**Definition.** i) Eine Wahrscheinlichkeitsverteilung  $\mu$  auf  $S$  heißt **Gleichgewichtsverteilung** (oder **stationäre Verteilung**) der Übergangsmatrix  $p$ , falls  $\mu p = \mu$ , d.h. falls:

$$\sum_{x \in S} \mu(x) p(x, y) = \mu(y) \quad \text{für alle } y \in S.$$

ii)  $\mu$  erfüllt die **Detailed Balance-Bedingung** bzgl. der Übergangsmatrix  $p$ , falls gilt:

$$\mu(x) p(x, y) = \mu(y) p(y, x) \quad \text{für alle } x, y \in S \quad (3.4.1)$$

**Satz 3.7.** Erfüllt  $\mu$  die Detailed Balance-Bedingung (3.4.1), dann ist  $\mu$  eine Gleichgewichtsverteilung von  $p$ .

*Beweis.* Aus der Detailed Balance-Bedingung folgt:

$$\sum_{x \in S} \mu(x) p(x, y) = \sum_{x \in S} \mu(y) p(y, x) = \mu(y).$$

□

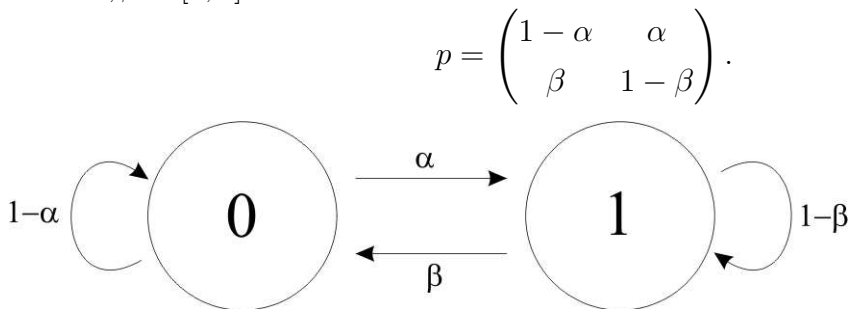
**Bemerkung.** Bei Startverteilung  $\mu$  gilt:

$$\mu(x) p(x, y) = P[X_0 = x, X_1 = y], \quad \text{»Fluss von } x \text{ nach } y\text{«.}$$

$$\begin{array}{lcl}
 \text{DETAILED BALANCE:} & \mu(x) p(x, y) & = \mu(y) p(y, x) \\
 & \text{»Fluss von } x \text{ nach } y\ll & = \text{»Fluss von } y \text{ nach } x\ll \\
 \text{GLEICHGEWICHT:} & \sum_{x \in S} \mu(x) p(x, y) & = \sum_{x \in S} \mu(y) p(y, x) \\
 & \text{»Gesamter Fluss nach } y\ll & \text{»Gesamter Fluss von } y\ll.
 \end{array}$$

**Beispiele.** a) MARKOV-KETTE AUF  $S = \{0, 1\}$ :

Seien  $\alpha, \beta \in [0, 1]$  und



Dann ist die Gleichgewichtsbedingung  $\mu p = \mu$  äquivalent zu den folgenden Gleichungen:

$$\mu(0) = \mu(0)(1 - \alpha) + \mu(1)\beta,$$

$$\mu(1) = \mu(0)\alpha + \mu(1)(1 - \beta).$$

Da  $\mu$  eine Wahrscheinlichkeitsverteilung ist, sind beide Gleichungen äquivalent zu

$$\beta(1 - \mu(0)) = \alpha\mu(0).$$

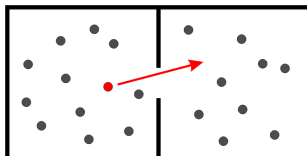
Die letzte Gleichung ist äquivalent zur Detailed Balance-Bedingung (3.4.1). Falls  $\alpha + \beta > 0$  gilt, ist  $\mu = \left(\frac{\beta}{\alpha + \beta}, \frac{\alpha}{\alpha + \beta}\right)$  die eindeutige Gleichgewichtsverteilung und erfüllt die Detailed Balance-Bedingung. Falls  $\alpha = \beta = 0$  gilt, ist jede Wahrscheinlichkeitsverteilung  $\mu$  eine Gleichgewichtsverteilung mit Detailed Balance-Bedingung.

b) ZYKLISCHER RANDOM WALK: Sei  $S = \mathbb{Z}/n\mathbb{Z}$  ein diskreter Kreis, und

$$p(k, k + 1) = p, \quad p(k, k - 1) = 1 - p.$$

Die Gleichverteilung  $\mu(x) = \frac{1}{n}$  ist ein Gleichgewicht. Die Detailed Balance-Bedingung ist dagegen nur für  $p = \frac{1}{2}$ , d.h. im symmetrischen Fall, erfüllt.

c) EHRENFEST-MODELL:



Sei  $S = \{0, 1, \dots, n\}$ ,

$$p(k, k - 1) = \frac{k}{n}, \quad p(k, k + 1) = \frac{n - k}{n}.$$

Man kann erwarten, daß sich im Gleichgewicht jede Kugel mit Wahrscheinlichkeit  $\frac{1}{2}$  in jeder der beiden Urnen befindet. Tatsächlich erfüllt die Binomialverteilung  $\mu(k) = \binom{n}{k} 2^{-n}$  mit Parameter  $p = \frac{1}{2}$  die Detailed Balance-Bedingung:

$$\mu(k-1)p(k-1, k) = \mu(k)p(k, k-1) \quad k = 1, \dots, n$$

ist äquivalent zu

$$2^{-n} \frac{n!}{(k-1)!(n-(k-1))!} \frac{n-(k-1)}{n} = 2^{-n} \frac{n!}{k!(n-k)!} \frac{k}{n} \quad k = 1, \dots, n$$

d) RANDOM WALKS AUF GRAPHEN:

Sei  $(V, E)$  ein endlicher Graph,  $S = V$  die Menge der Knoten.

- Sei

$$p(x, y) = \begin{cases} \frac{1}{\deg(x)} & \text{falls } \{x, y\} \in E, \\ 0 & \text{sonst.} \end{cases}$$

Die Detailed Balance-Bedingung lautet in diesem Fall:

$$\mu(x)p(x, y) = \mu(y)p(y, x).$$

Sie ist erfüllt, falls

$$\mu(x) = c \deg(x)$$

gilt, wobei  $c$  eine Konstante ist. Damit  $\mu$  eine Wahrscheinlichkeitsverteilung ist, muss  $c$  so gewählt werden, dass gilt:

$$\sum_{x \in B} \deg(x) = 2|E|.$$

Somit ist die Gleichgewichtsverteilung:

$$\mu(x) = \frac{\deg(x)}{2|E|}.$$

- Sei  $\Delta := \max_{x \in V} \deg(x)$ ,

$$p(x, y) = \begin{cases} \frac{1}{\Delta} & \text{falls } \{x, y\} \in E, \\ 1 - \frac{\deg(x)}{\Delta} & \text{sonst.} \end{cases}$$

Es gilt  $p(x, y) = p(y, x)$  und somit ist die Gleichverteilung auf  $V$  die stationäre Verteilung.

Ist  $\deg(x)$  konstant, dann stimmen die Random Walks in beiden Beispielen überein, und die Gleichverteilung ist ein Gleichgewicht.

Im nächsten Abschnitt zeigen wir:

**Satz** (Konvergenzsatz für Markov-Ketten). *Ist  $S$  endlich, und  $p$  eine irreduzible und aperiodische stochastische Matrix mit Gleichgewicht  $\mu$ , dann gilt für alle Wahrscheinlichkeitsverteilungen  $\nu$  auf  $S$ :*

$$\lim_{n \rightarrow \infty} (\nu p^n)(x) = \mu(x) \quad \text{für alle } x \in S.$$

Aufgrund des Konvergenzsatzes können wir Stichproben von einer Wahrscheinlichkeitsverteilung  $\mu$  näherungsweise erzeugen, indem wir eine Markov-Kette  $X_n$  mit Gleichgewicht  $\mu$  simulieren, und für großes  $n$  auswerten. Wie findet man eine Markov-Kette mit einer vorgegebenen stationären Verteilung?

## Metropolis-Algorithmus und Gibbs-Sampler

### Die Metropolis-Kette

Sei  $q(x, y)$  eine symmetrische stochastische Matrix, d.h.  $q(x, y) = q(y, x)$  für alle  $x, y \in S$ . Dann erfüllt die Gleichverteilung die Detailed Balance-Bedingung (3.4.1). Sei nun  $\mu$  eine beliebige Wahrscheinlichkeitsverteilung auf  $S$  mit  $\mu(x) > 0$  für alle  $x \in S$ . Wie können wir die Übergangsmatrix  $q$  so modifizieren, dass die Detailed Balance-Bedingung bzgl.  $\mu$  erfüllt ist?

**Algorithmus 3.8** (Metropolis-Algorithmus (Update  $x \rightarrow y$ )).  
 schlage Übergang  $x \rightarrow y$  mit Wahrscheinlichkeit  $q(x, y)$  vor  
 akzeptiere Übergang mit Wahrscheinlichkeit  $\alpha(x, y) \in [0, 1]$   
 sonst verwerfe Vorschlag und bleibe bei  $x$

ÜBERGANGSMATRIX:

$$p(x, y) := \begin{cases} \alpha(x, y) q(x, y) & \text{für } y \neq x, \\ 1 - \sum_{y \neq x} \alpha(x, y) q(x, y) & \text{für } y = x. \end{cases}$$

Die Detailed Balance-Bedingung lautet:

$$\mu(x) \alpha(x, y) = \mu(y) \alpha(y, x) \quad \text{für alle } x, y \in S.$$

Sie ist äquivalent dazu, dass

$$b(x, y) := \mu(x) \alpha(x, y)$$

symmetrisch in  $x$  und  $y$  ist. Was ist die größtmögliche Wahl von  $b(x, y)$ ?

Aus  $\alpha(x, y) \leq 1$  folgen

$$\begin{aligned} b(x, y) &\leq \mu(x), \\ b(x, y) &= b(y, x) \leq \mu(y), \end{aligned}$$

und somit

$$b(x, y) \leq \min(\mu(x), \mu(y)).$$

Der größtmögliche Wert  $b(x, y) = \min(\mu(x), \mu(y))$  entspricht gerade

$$\alpha(x, y) = \min\left(1, \frac{\mu(y)}{\mu(x)}\right) = \begin{cases} 1 & \text{falls } \mu(y) \geq \mu(x), \\ \frac{\mu(y)}{\mu(x)} & \text{falls } \mu(x) \geq \mu(y). \end{cases}$$

**Definition.** Die Markov-Kette mit Übergangsmatrix

$$p(x, y) = \min\left(1, \frac{\mu(y)}{\mu(x)}\right) \cdot q(x, y) \quad \text{für } y \neq x$$

heißt **Metropolis-Kette** mit Vorschlagsverteilung  $q(x, y)$  und Gleichgewicht  $\mu$ .

**Satz 3.9.**  $\mu$  erfüllt die Detailed Balance-Bedingung bzgl.  $p$ .

*Beweis.* siehe oben. □

### Der Gibbs-Sampler

Sei  $S = S_1 \times \dots \times S_d$  ein endlicher Produktraum,  $\mu(x_1, \dots, x_d)$  eine Wahrscheinlichkeitsverteilung auf  $S$  und

$$\mu_i(x_i \mid x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d) := \frac{\mu(x_1, \dots, x_d)}{\sum_{z \in S_i} \mu(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_d)}$$

die bedingte Verteilung der  $i$ -ten Komponente gegeben die übrigen Komponenten.

**Algorithmus 3.10** (Gibbs-Sampler (Update  $x \rightarrow y$ )).  $y := x$

**for**  $i := 1, \dots, d$  **do**

update  $y_i \sim \mu_i(\bullet \mid y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_d)$

**end for**

**return**  $y$

ÜBERGANGSMATRIX:

$$p = p_d p_{d-1} \cdots p_1,$$

wobei

$$p_i(x, y) = \begin{cases} \mu_i(y_i \mid y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_d) & \text{falls } y_k = x_k \text{ für alle } k \neq i, \\ 0 & \text{sonst.} \end{cases}$$

**Satz 3.11.** i)  $\mu$  erfüllt die Detailed Balance-Bedingung bzgl.  $p_i$  für alle  $i = 1, \dots, d$ .

ii)  $\mu$  ist ein Gleichgewicht von  $p$ .

*Beweis.* i) Der Beweis der ersten Aussage ist eine Übungsaufgabe.

ii) Nach der ersten Aussage ist  $\mu$  ein Gleichgewicht von  $p_i$  für alle  $i$ . Also gilt auch

$$\mu p = \mu p_d p_{d-1} \cdots p_1 = \mu.$$

□

**Bemerkung.** Zur Simulation von  $Y_n$ ,  $n \geq 0$ , genügt es, die Massenfunktion  $\mu(x)$  bis auf eine multiplikative Konstante zu kennen:

aus  $\mu(x) = C f(X)$  folgt

$$\alpha(x, y) = \min\left(1, \frac{f(y)}{f(x)}\right) \quad \text{unabhängig von } C.$$

**Beispiel (Rucksackproblem).** Gegeben:

$$\omega_1, \dots, \omega_d \in \mathbb{R}, \quad \text{»Gewichte«},$$

$$v_1, \dots, v_d \in \mathbb{R}, \quad \text{»Werte«}.$$

Rucksack mit maximalem Gewicht  $b > 0$ , packe soviel Wert wie möglich ein.

$$S = \{0, 1\}^d, \quad \text{alle Konfigurationen,}$$

$$S_b = \{(z_1, \dots, z_d) \in S : \sum_{i=1}^d z_i \omega_i \leq b\}, \quad \text{zulässige Konfigurationen,}$$

$$z_i = 1 : i\text{-ter Gegenstand im Rucksack.}$$

RUCKSACKPROBLEM:

$$\text{maximiere } V(z) = \sum_{i=1}^d z_i v_i \text{ unter Nebenbedingung } z \in S_b.$$

Das Rucksackproblem ist **NP-vollständig**, insbesondere ist keine Lösung in  $O(d^k)$  Schritten für ein  $k \in \mathbb{N}$  bekannt.

STOCHASTISCHER ZUGANG: SIMULATED ANNEALING

Für  $\beta > 0$  betrachten wir die Wahrscheinlichkeitsverteilung

$$\mu_\beta(z) = \begin{cases} \frac{1}{Z_\beta} e^{\beta V(z)} & \text{für } z \in S_b, \\ 0 & \text{für } z \in S \setminus S_b, \end{cases}$$

auf  $S$ , wobei  $Z_\beta = \sum_{z \in S_b} e^{\beta V(z)}$  eine Konstante ist, die  $\mu$  zu einer Wahrscheinlichkeitsverteilung normiert. Für  $\beta = 0$  ist  $\mu_\beta$  die Gleichverteilung auf  $S_b$ . Für  $\beta \rightarrow \infty$  konvergiert  $\mu_\beta$  gegen die Gleichverteilung auf der Menge der globalen Maxima von  $V$ , denn:

$$\mu_\beta(z) = \frac{e^{\beta V(z)}}{Z_\beta} = \frac{1}{\sum_{y \in S_b} e^{\beta(V(y)-V(z))}} \rightarrow \begin{cases} 0 & \text{falls } V(z) \neq \max V, \\ \frac{1}{|\{y \mid V(y) = \max V\}|} & \text{falls } V(z) = \max V. \end{cases}$$

IDEE: Simuliere Stichprobe  $z$  von  $\mu_\beta$  für  $\beta$  groß ( $\beta \rightarrow \infty$ ). Dann ist  $V(z)$  **wahrscheinlich** nahe dem Maximalwert.

METROPOLIS-ALGORITHMUS: Sei  $x^+ := \max(x, 0)$  der Positivteil von  $x$ . Wir wählen als Vorschlagsmatrix die Übergangsmatrix

$$q(z, w) := \begin{cases} \frac{1}{d} & \text{falls } z_i \neq w_i \text{ für genau ein } i \in \{1, \dots, d\}, \\ 0 & \text{sonst,} \end{cases}$$

des Random Walks auf  $\{0, 1\}^d$ . Für die Akzeptanzwahrscheinlichkeit ergibt sich

$$\alpha_\beta(z, w) = \min \left( 1, \frac{\mu_\beta(w)}{\mu_\beta(z)} \right) = \begin{cases} e^{-\beta(V(z)-V(w))} & \text{für } z, w \in S_b, \\ 0 & \text{für } z \in S_b, w \notin S_b. \end{cases}$$

Der Vorschlag  $w$  wird also mit Wahrscheinlichkeit 1 akzeptiert, wenn  $V(w) \geq V(z)$  gilt – andernfalls wird der Vorschlag nur mit Wahrscheinlichkeit  $\exp -\beta(V(z) - V(w))$  akzeptiert.

**Algorithmus 3.12** (Simulation einer Markov-Kette mit Gleichgewicht  $\mu_\beta$ ). initialisiere  $z^{(0)} \in S_b$

**for**  $n = 1, 2, \dots$  **do**

$z^{(n)} := w := z^{(n-1)}$

erzeuge  $i \sim \text{Unif}\{1, \dots, d\}$

$w_i := 1 - w_i$

**if**  $w \in S_b$  **then**

```

erzeuge  $u \sim \text{Unif}(0, 1)$ 
if  $u \leq \alpha_\beta(z, w)$  then
     $z^{(n)} := w$ 
end if
end if
end for

```

**Algorithmus 3.13** (Simulated Annealing). Wie Algorithmus 3.12 aber mit  $\beta = \beta(n) \rightarrow \infty$  für  $n \rightarrow \infty$ .

**Bemerkung.** a) PHYSIKALISCHE INTERPRETATIONEN:

$\mu_\beta$  ist die Verteilung im thermodynamischen Gleichgewicht für die Energiefunktion  $H(z) = -V(z)$  bei der Temperatur  $T = 1/\beta$ . Der Grenzwert  $\beta \rightarrow \infty$  entspricht  $T \rightarrow 0$  («simuliertes Abkühlen«).

b) Die beim Simulated Annealing-Verfahren simulierte zeitlich inhomogene Markov-Kette findet im allgemeinen nicht das globale Maximum von  $V$ , sondern kann in lokalen Maxima »steckenbleiben«. Man kann zeigen, daß die Verteilung der Markov-Kette zur Zeit  $n$  gegen die Gleichverteilung auf den Maximalstellen konvergiert, falls  $\beta(n)$  nur sehr langsam (logarithmisch) gegen  $+\infty$  geht. In praktischen Anwendungen wird der Algorithmus aber in der Regel mit einem schnelleren »Cooling schedule«  $\beta(n)$  verwendet. Das Auffinden eines globalen Maximums ist dann nicht garantiert – trotzdem erhält man ein oft nützliches **heuristisches** Verfahren.

### 3.5 Konvergenz ins Gleichgewicht

Sei  $S = \{x_1, \dots, x_m\}$  eine endliche Menge, und

$$\text{WV}(S) := \{\mu = (\mu(x_1), \dots, \mu(x_m)) \mid \mu(x_i) \geq 0, \sum_{i=1}^m \mu(x_i) = 1\} \subseteq \mathbb{R}^m$$

die Menge aller Wahrscheinlichkeitsverteilungen auf  $S$ . Geometrisch ist  $\text{WV}(S)$  ein Simplex im  $\mathbb{R}^m$ . Wir führen nun einen Abstandsbegriff auf  $\text{WV}(S)$  ein:

**Definition.** Die *Variationsdistanz* zweier Wahrscheinlichkeitsverteilungen  $\mu, \nu$  auf  $S$  ist:

$$d_{TV}(\mu, \nu) := \frac{1}{2} \|\mu - \nu\|_1 = \frac{1}{2} \sum_{x \in S} |\mu(x) - \nu(x)|.$$



**Bemerkung.** a) Für alle  $\mu, \nu \in \text{WV}(S)$  gilt:

$$d_{TV}(\mu, \nu) \leq \frac{1}{2} \sum_{x \in S} (\mu(x) + \nu(x)) = 1.$$

b) Seien  $\mu, \nu$  Wahrscheinlichkeitsverteilungen und  $A := \{x \in S \mid \mu(x) \geq \nu(x)\}$ . Dann gilt:

$$d_{TV}(\mu, \nu) = \sum_{x \in A} (\mu(x) - \nu(x)) = \max_{A \subseteq S} |\mu(A) - \nu(A)|.$$

Der Beweis dieser Aussage ist eine Übungsaufgabe.

Wir betrachten im folgenden eine stochastische Matrix  $p(x, y)$ , ( $x, y \in S$ ), mit Gleichgewicht  $\mu$ . Die Verteilung einer Markov-Kette mit Startverteilung  $\nu$  und Übergangsmatrix  $p$  zur Zeit  $n$  ist  $\nu p^n$ . Um Konvergenz ins Gleichgewicht zu zeigen, verwenden wir die folgende Annahme:

**MINORISIERUNGSBEDINGUNG:** Es gibt ein  $\delta \in (0, 1]$  und ein  $r \in \mathbb{N}$ , so dass für alle  $x, y \in S$  gilt:

$$p^r(x, y) \geq \delta \cdot \mu(y). \quad (3.5.1)$$

**Satz 3.14.** Gilt die Minorisierungsbedingung (3.5.1), dann konvergiert  $\nu p^n$  für jede Startverteilung  $\nu$  exponentiell schnell gegen  $\mu$ . Genauer gilt für alle  $n \in \mathbb{N}$  und  $\nu \in \text{WV}(S)$ :

$$d_{TV}(\nu p^n, \mu) \leq (1 - \delta)^{\lfloor n/r \rfloor}.$$

**Bemerkung.** Insbesondere ist  $\mu$  das **eindeutige** Gleichgewicht: Betrachte eine beliebige Wahrscheinlichkeitsverteilung  $\nu$  mit  $\nu p = \nu$ . Dann folgt für  $n \rightarrow \infty$ :

$$d_{TV}(\nu, \mu) = d_{TV}(\nu p^n, \mu) \rightarrow 0,$$

also  $d_{TV}(\mu, \nu) = 0$ , und somit  $\mu = \nu$ .

*Beweis.* 1. Durch die Zerlegung

$$p^r(x, y) = \delta \mu(y) + (1 - \delta) q(x, y)$$

der  $r$ -Schritt-Übergangswahrscheinlichkeiten wird eine **stochastische** Matrix  $q$  definiert, denn

(i) Aus der Minorisierungsbedingung (3.5.1) folgt  $q(x, y) \geq 0$  für alle  $x, y \in S$ .

(ii) Aus  $\sum_{y \in S} p^r(x, y) = 1$ ,  $\sum_{y \in S} \mu(y) = 1$  folgt  $\sum_{y \in S} q(x, y) = 1$  für alle  $x \in S$ .

Wir setzen im folgenden  $\lambda := 1 - \delta$ . Dann gilt für alle  $\nu \in \text{WV}(S)$ :

$$\nu p^r = (1 - \lambda) \mu + \lambda \nu q. \quad (3.5.2)$$

2. Wir wollen mit vollständiger Induktion zeigen:

$$\nu p^{kr} = (1 - \lambda^k) \mu + \lambda^k \nu q^k \quad \text{für alle } k \geq 0, \quad \nu \in \text{WV}(S). \quad (3.5.3)$$

Für  $k = 0$  ist die Aussage offensichtlich wahr. Gilt (3.5.3) für ein  $k \geq 0$ , dann erhalten wir durch Anwenden von Gleichung (3.5.2) auf  $\tilde{\nu} p^r$  mit  $\tilde{\nu} = \nu q^k$ :

$$\begin{aligned} \nu p^{(k+1)r} &= \nu p^{kr} p^r \\ &= ((1 - \lambda^k) \mu + \lambda^k \underbrace{\nu q^k}_{=\tilde{\nu}}) p^r \\ &= (1 - \lambda^k) \underbrace{\mu p^r}_{=\mu} + (1 - \lambda) \lambda^k \mu + \lambda^{k+1} \nu q^k q \\ &= (1 - \lambda^{k+1}) \mu + \lambda^{k+1} \nu q^{k+1}. \end{aligned}$$

3. Für  $n \in \mathbb{N}$ ,  $n = kr + i$ , ( $k \in \mathbb{N}$ ,  $0 \leq i < r$ ), folgt:

$$\nu p^n = \nu p^{kr} p^i = (1 - \lambda^k) \underbrace{\mu p^i}_{=\mu} + \lambda^k \nu q^k p^i,$$

also

$$\nu p^n - \mu = \lambda^k (\nu q^k p^i - \mu) \quad \text{für alle } \nu \in \text{WV}(S),$$

und damit

$$d_{TV}(\nu p^n, \mu) = \frac{1}{2} \|\nu p^n - \mu\|_1 = \lambda^k d_{TV}(\nu q^k p^i, \mu) \leq \lambda^k$$

nach der letzten Bemerkung.

□

Welche Übergangsmatrizen erfüllen die Minorisierungsbedingung?

**Definition.** i) Die stochastische Matrix  $p$  heißt **irreduzibel**, falls es für alle  $x, y \in S$  ein  $n \in \mathbb{N}$  gibt, so dass  $p^n(x, y) > 0$  gilt.

ii) Die **Periode** von  $x \in S$  ist definiert als

$$\text{Periode}(x) := \text{ggT}(\underbrace{\{n \in \mathbb{N} \mid p^n(x, x) > 0\}}_{=:R(x)}).$$

**Lemma 3.15.** i) Falls  $p$  irreduzibel ist, gilt  $\text{Periode}(x) = \text{Periode}(y)$  für alle  $x, y \in S$ .

ii) Falls  $p$  irreduzibel und aperiodisch (d.h.  $\text{Periode}(x) = 1$  für alle  $x \in S$ ) ist, gibt es ein  $r > 0$ , so dass  $p^r(x, y) > 0$  für alle  $x, y \in S$  gilt.

*Beweis.* Seien  $x, y \in S$ .

i) Sei  $p$  irreduzibel. Dann gibt es ein  $s$  und ein  $t \in \mathbb{N}$ , so dass gilt:

$$p^s(x, y) > 0 \quad \text{und} \quad p^t(y, x) > 0.$$

Für  $a := s + t$  folgt:

- $p^a(x, x) \geq p^s(x, y) p^t(y, x) > 0$ , also  $a \in R(x)$ .
- $p^{n+a}(x, x) \geq p^s(x, y) p^n(y, y) p^t(y, x) > 0$  für alle  $n \in R(y)$ , also  $n + a \in R(x)$  für alle  $n \in R(y)$ .

$\text{Periode}(x)$  ist ein gemeinsamer Teiler von  $R(x)$ , somit Teiler von  $a$  und  $n + a$ , also auch von  $n$  für alle  $n \in R(y)$ . Daher ist  $\text{Periode}(x)$  ein gemeinsamer Teiler von  $R(y)$  und somit gilt:

$$\text{Periode}(x) \leq \text{Periode}(y).$$

» $\geq$ « wird analog gezeigt. Es folgt:

$$\text{Periode}(x) = \text{Periode}(y).$$

ii)  $R(x)$  ist abgeschlossen unter Addition, denn falls  $s, t \in R(x)$  ist, gilt:

$$p^{s+t}(x, x) \geq p^s(x, x) p^t(x, x) > 0,$$

und somit  $s + t \in R(x)$ . Da  $p$  aperiodisch ist, folgt  $\text{ggT}(R(x)) = 1$  für alle  $x \in S$ . Nach einem Satz der Zahlentheorie gilt:

Da  $R(x)$  additiv abgeschlossen, gibt es für alle  $x$  ein  $r(x) \in \mathbb{N}$  mit  $n \in R(x)$  für alle  $n \geq r(x)$ .

$n \in R(x)$  impliziert  $p^n(x, x) > 0$ . Da  $p$  irreduzibel ist, folgt, dass es für alle  $x, y \in S$  ein  $r(x, y) \in \mathbb{N}$  gibt, so dass gilt:

$$p^n(x, y) > 0 \quad \text{für alle } n \geq r(x, y).$$

Für  $r \geq \max_{x, y \in S} r(x, y)$  folgt dann  $p^r(x, y) > 0$  für alle  $x, y \in S$ .

□

**Satz 3.16** (Konvergenzsatz für **endliche** Markov-Ketten). *Ist  $p$  irreduzibel und aperiodisch mit Gleichgewicht  $\mu$ , dann gilt:*

$$\lim_{n \rightarrow \infty} d_{TV}(\nu p^n, \mu) = 0 \quad \text{für alle } \nu \in \text{WV}(S).$$

*Beweis.* Da  $p$  irreduzibel und aperiodisch ist, gibt es ein  $r \in \mathbb{N}$  mit:

$$p^r(x, y) > 0 \quad \text{für alle } x, y \in S.$$

Daher gibt es ein  $r \in \mathbb{N}$  und ein  $\delta > 0$ , so dass gilt:

$$p^r(x, y) > \delta \mu(y) \quad \text{für alle } x, y \in S,$$

(z.B.  $\delta := \min_{x, y \in S} p^r(x, y)$ ). Mit Satz 3.14 folgt die Behauptung. □

**Beispiel** (Metropolis-Kette). Sei  $S$  endlich,  $\mu(x) > 0$  für alle  $x \in S$ , nicht konstant, und  $q(x, y)$  irreduzibel. Dann ist  $p(x, y)$  irreduzibel und aperiodisch. Somit folgt die Konvergenz ins Gleichgewicht nach Satz 3.16, allerdings evtl. sehr langsam!

#### ANWENDUNG: MARKOV-CHAIN-MONTE CARLO-VERFAHREN

Sei  $\mu \in \text{WV}(S)$ ,  $f: S \rightarrow \mathbb{R}$ .

GESUCHT:

$$\theta = E_\mu[f],$$

MARKOV-CHAIN-MONTE CARLO-SCHÄTZER:

$$\hat{\theta}_{n,b} = \frac{1}{n} \sum_{k=b+1}^{b+n} f(X_k),$$

wobei  $b \in \mathbb{N}$  eine feste Konstante (»burn-in-Zeit«) und  $(X_k)_{k \in \mathbb{N}}$  irreduzible Markov-Ketten mit Gleichgewicht  $\mu$  sind.

**Satz** (Ergodensatz / Gesetz der großen Zahlen für Markov-Ketten). : *Für alle  $b \in \mathbb{N}$  gilt:*

$$\lim_{n \rightarrow \infty} \hat{\theta}_{n,b} = \theta \quad \text{mit Wahrscheinlichkeit 1,}$$

*Beweis.* siehe Vorlesung »Stochastische Prozesse«. □

Die Analyse des Schätzfehler ist im Allgemeinen diffizil!

# Kapitel 4

## Stetige und Allgemeine Modelle

### 4.1 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

$$P \left[ \left| \frac{S_n}{n} - p \right| \geq \varepsilon \right] \leq 2 \cdot e^{-2\varepsilon^2 n}, \quad (4.1.1)$$

d.h. die Verteilung von  $\frac{S_n}{n}$  konzentriert sich für  $n \rightarrow \infty$  sehr rasch in der Nähe von  $p$ , siehe Abschnitt 2.3. 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.

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

**Definition.** (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 (4.1.1) sogar

$$\lim_{n \rightarrow \infty} \frac{S_n(\omega)}{n} = p \quad \text{für } P\text{-fast alle } \omega \in \Omega \quad (4.1.2)$$

folgt. Das relevante Ereignis

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

lässt sich offensichtlich nicht durch endlich viele der  $A_i$  beschreiben.

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

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

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

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

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  $\sigma$ -additiv ist. Der folgende Satz gibt eine alternative Charakterisierung der  $\sigma$ -Additivität:

**Satz 4.1** ( $\sigma$ -Additivität und monotone Stetigkeit). Sei  $\mathcal{A}$  eine  $\sigma$ -Algebra und  $P : \mathcal{A} \rightarrow [0, \infty]$  additiv, d.h.

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

(i)  $P$  ist  $\sigma$ -additiv genau dann, wenn:

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

(ii) Gilt  $P[\Omega] = 1$ , dann ist dies auch äquivalent zu:

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

*Beweis.* (i) Sei  $P$   $\sigma$ -additiv und  $A_1 \subseteq A_2 \subseteq \dots$ . Die Mengen  $B_1 := A_1$ ,  $B_2 := A_2 \setminus A_1$ ,  $B_3 := A_3 \setminus A_2, \dots$  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:

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

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

(ii) Gilt  $P[\Omega] = 1$ , dann folgt

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

Die Behauptung folgt nun aus (i). □

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

**Satz 4.2 ( $\sigma$ -Subadditivität).** Für beliebige Ereignisse  $A_1, A_2, \dots \in \mathcal{A}$  gilt:

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

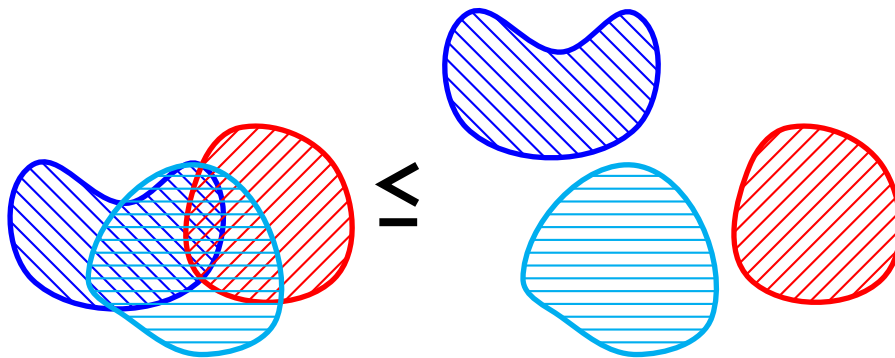


Abbildung 4.1: Darstellung von drei Mengen. Das Maß der Vereinigung von Mengen ist stets kleiner gleich als die Summe der Maße der einzelnen Mengen.

*Beweis.* Die Mengen

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

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.

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

**Satz 4.3 (1. Borel - Cantelli - Lemma).** Für Ereignisse  $A_1, A_2, \dots \in \mathcal{A}$  mit

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



*gilt:*

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

*Beweis.* Da die Folge  $\bigcup_{n \geq m} A_n =: B_m$  von Ereignissen aus  $\mathcal{A}$  monoton fallend ist, ergibt sich nach Satz 4.1 und 4.2:

$$\begin{aligned} P\left[\bigcap_{m \in \mathbb{N}} \bigcup_{n \geq m} A_n\right] &= P\left[\bigcap_{m \in \mathbb{N}} B_m\right] \\ &\stackrel{4.1}{=} \lim_{m \rightarrow \infty} P[B_m] \\ &= \lim_{m \rightarrow \infty} P\left[\bigcup_{n \geq m} A_n\right] \\ &\stackrel{4.2}{\leq} \sum_{n=m}^{\infty} P[A_n] \\ &\leq \liminf_{m \rightarrow \infty} \underbrace{\sum_{n=m}^{\infty} P[A_n]}_{\xrightarrow{m \rightarrow \infty} 0} = 0, \end{aligned}$$

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

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

*gilt:*

$$P[A_n \text{ unendlich oft}] = P\left[\bigcap_{m \in \mathbb{N}} \bigcup_{n \geq m} A_n\right] = 1$$

**Bemerkung.** Insbesondere ergibt sich ein **0-1 Gesetz**:

Sind  $A_1, A_2, \dots \in \mathcal{A}$  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 2.5. 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 4.1 gilt:

$$P\left[\bigcup_m \bigcap_{n \geq m} A_n^C\right] = \lim_{m \rightarrow \infty} P\left[\bigcap_{n \geq m} A_n^C\right] \quad (4.1.3)$$

Wegen der Unabhängigkeit der Ereignisse  $A_n^C$  erhalten wir zudem

$$\begin{aligned} P\left[\bigcap_{n \geq m} A_n^C\right] &\stackrel{\text{mon. Stetigkeit}}{=} \lim_{k \rightarrow \infty} P\left[\bigcap_{n=m}^k A_n^C\right] \\ &\stackrel{\text{unabh.}}{=} \lim_{k \rightarrow \infty} \prod_{n=m}^k \underbrace{P[A_n^C]}_{=1-P[A_n] \leq \exp(-P[A_n])} \\ &\leq \liminf_{k \rightarrow \infty} \prod_{n=m}^k e^{-P[A_n]} \\ &= \liminf_{k \rightarrow \infty} e^{-\sum_{n=m}^k P[A_n]} = 0, \end{aligned} \quad (4.1.4)$$

da  $\lim_{k \rightarrow \infty} \sum_{n=m}^k P[A_n] = \sum_{n=m}^{\infty} P[A_n] = \infty$  nach Voraussetzung.

Aus 4.1.3 und 4.1.4 folgt die Behauptung.  $\square$

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

**Satz 4.5 (Starkes Gesetz großer Zahlen I, Borel 1909, Hausdorff 1914, Cantelli 1917).** Sind  $A_1, A_2, \dots \in \mathcal{A}$  unabhängige Ereignisse mit Wahrscheinlichkeit  $P[A_n] = p$  für alle  $n \in \mathbb{N}$ , dann gilt für  $S_n = \sum_{i=1}^n I_{A_i}$ :

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

*Beweis.* Sei

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

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

Wegen

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

gilt:

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

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 4.2 selbst eine Nullmenge.  $\square$

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

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

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 4.5 folgt:

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

Für  $p \neq \frac{1}{2}$  wächst (bzw. fällt)  $Z_n$  also mit Wahrscheinlichkeit 1 asymptotisch linear (siehe Abbildung 4.2):

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

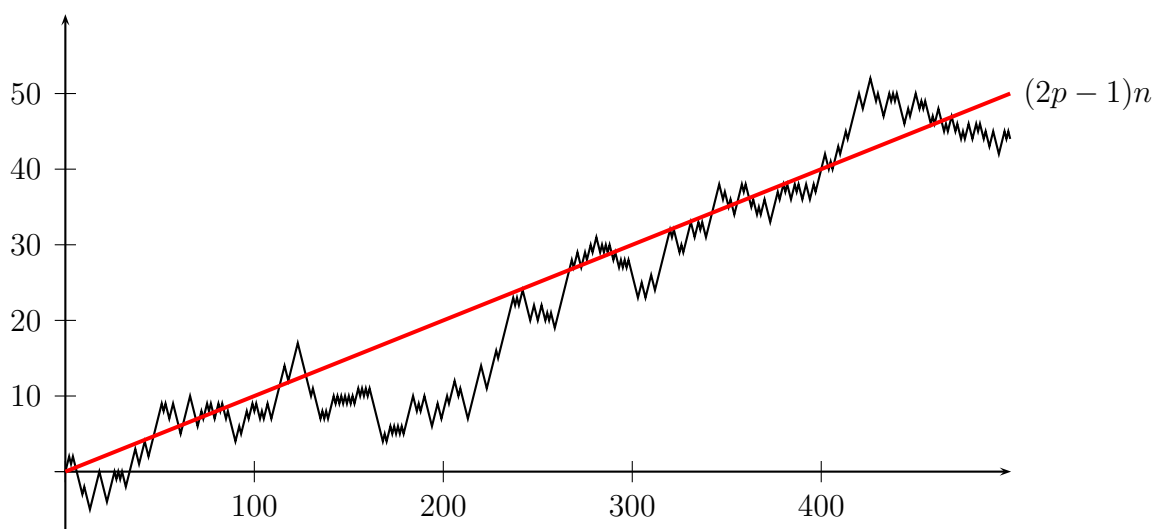


Abbildung 4.2: Random Walk mit Drift:  $p = 0.55, n = 500$

Für  $p = \frac{1}{2}$  dagegen wächst der Random Walk sublinear, d.h.  $\frac{Z_n}{n} \rightarrow 0$   $P$ -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 4.3).

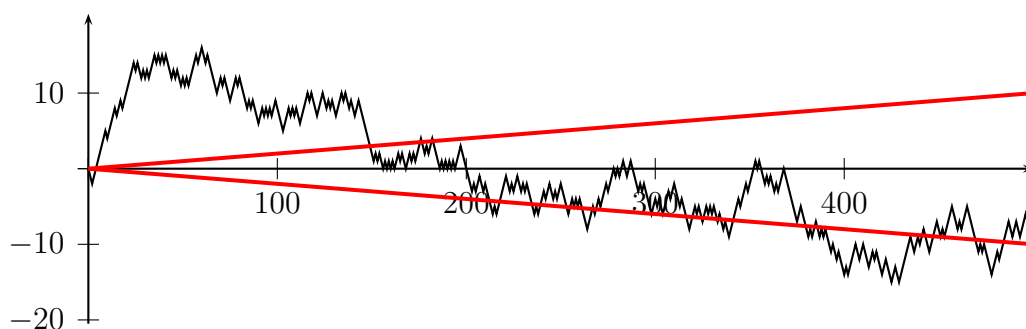


Abbildung 4.3: Random Walk ohne Drift:  $p = 0.5, n = 500$

Eine viel präzisere Beschreibung der Asymptotik des Random Walk liefert der **Satz vom iterierten Logarithmus**:

$$\limsup_{n \rightarrow \infty} \frac{S_n(\omega)}{\sqrt{n \log \log n}} = +1 \text{ } P\text{-fast sicher,}$$

$$\liminf_{n \rightarrow \infty} \frac{S_n(\omega)}{\sqrt{n \log \log n}} = -1 \text{ } P\text{-fast sicher}$$

Mehr dazu: siehe Vorlesung „Stochastische Prozesse.“

## 4.2 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.

### Beispiele von Wahrscheinlichkeitsräumen

Wir beginnen mit einer Auflistung von verschiedenen Wahrscheinlichkeitsräumen  $(\Omega, \mathcal{A}, P)$ , die wir gerne konstruieren würden:

#### Dirac-Maß

Sei  $\Omega$  beliebig,  $a \in \Omega$  fest,  $\mathcal{A} = \mathcal{P}(\Omega)$ ,  $P = \delta_a$ , wobei

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

Dies ist eine deterministische Verteilung mit:

$$P[\{\omega = a\}] = 1$$

### Diskrete Wahrscheinlichkeitsräume

Ist  $\Omega$  eine abzählbare Menge und  $p : \Omega \rightarrow [0, 1]$  eine Gewichtsfunktion mit  $\sum_{\omega \in \Omega} p(\omega) = 1$ , dann haben wir bereits gezeigt, dass eine eindeutige Wahrscheinlichkeitsverteilung  $P$  auf der Potenzmenge  $\mathcal{A} = \mathcal{P}(\Omega)$  existiert mit

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

Jede diskrete Wahrscheinlichkeitsverteilung ist eine Konvexkombination von Diracmaßen:

$$P = \sum_{a \in \Omega} p(a) \delta_a$$

### Endliche Produktmodelle

Auch die Konstruktion mehrstufiger diskreter Modelle ist auf diese Weise möglich: Ist beispielsweise

$$\Omega = \{(\omega_1, \dots, \omega_n) : \omega_i \in \Omega_i\} = \Omega_1 \times \dots \times \Omega_n$$

eine Produktmenge, und sind  $p_1, \dots, p_n$  Gewichtsfunktionen von Wahrscheinlichkeitsverteilungen  $P_1, \dots, P_n$  auf  $\Omega_1, \dots, \Omega_n$ , dann ist

$$p(\omega) = \prod_{i=1}^n p_i(\omega_i)$$

die Gewichtsfunktion einer Wahrscheinlichkeitsverteilung  $P = P_1 \otimes \dots \otimes P_n$  auf  $\Omega$ . Unter dieser Wahrscheinlichkeitsverteilung sind die Zufallsvariablen  $X_i(\omega) = \omega_i$  unabhängig.

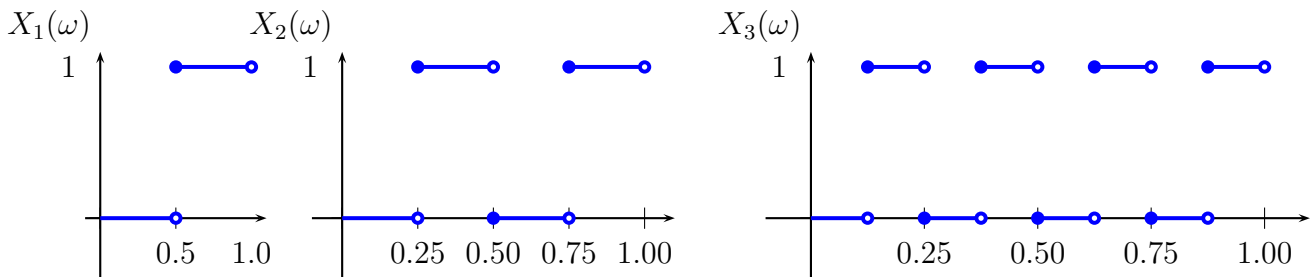
### Unendliches Produktmodell (z.B. Münzwurffolge)

Es stellt sich die Frage, ob wir auch unendlich viele unabhängige Zufallsvariablen auf einem ähnlichen Produktraum realisieren können. Im einfachsten Fall möchten wir eine Folge unabhängiger fairer Münzwürfe (0-1-Experimente) auf dem Grundraum

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

modellieren.  $\Omega$  ist überabzählbar, denn die Abbildung  $X : (0, 1) \rightarrow \Omega$ , die einer reellen Zahl die Ziffernfolge ihrer Binärdarstellung zuordnet, ist injektiv. Genauer ist eine injektive Abbildung  $X : (0, 1) \rightarrow \Omega$  definiert durch

$$X(\omega) = (X_1(\omega), X_2(\omega), X_3(\omega), \dots), \tag{4.2.1}$$

Abbildung 4.4: Darstellung der ersten drei  $X_i(\omega)$ .

wobei  $X_n(\omega) = I_{D_n}(\omega)$ ,  $D_n = \bigcup_{i=1}^{2^{n-1}} [(2i-1) \cdot 2^{-n}, 2i \cdot 2^{-n})$ .

Wir suchen eine Wahrscheinlichkeitsverteilung  $P$  auf  $\Omega$  mit

$$P[\{\omega \in \Omega : \omega_1 = a_1, \omega_2 = a_2, \dots, \omega_n = a_n\}] = 2^{-n} \quad (4.2.2)$$

Gibt es eine  $\sigma$ -Algebra  $\mathcal{A}$ , die alle diese Ereignisse enthält, und eine eindeutige Wahrscheinlichkeitsverteilung  $P$  auf  $\mathcal{A}$  mit (4.2.2)?

Wir werden in Abschnitt 5.3 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  $\Omega = [a, b]$ ,  $-\infty < a < b < \infty$ , sollte gelten:

$$P[(c, d)] = P[[c, d]] = \frac{d - c}{b - a} \quad \forall a \leq c < d \leq b. \quad (4.2.3)$$

Gibt es eine  $\sigma$ -Algebra  $\mathcal{B}$ , die alle Teilintervalle von  $[a, b]$  enthält, und eine Wahrscheinlichkeitsverteilung  $P$  auf  $\mathcal{B}$  mit (4.2.3)?

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

Tatsächlich sind die Probleme in den letzten beiden Abschnitten weitgehend äquivalent: die durch die Binärdarstellung (4.2.1) 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  $\rightarrow 0$  anscheinend eine irreguläre, aber stetige zufällige Bewegung in kontinuierlicher Zeit. Der entsprechende, 1923 von N. Wiener konstruierte stochastische Prozess heißt **Brown-**

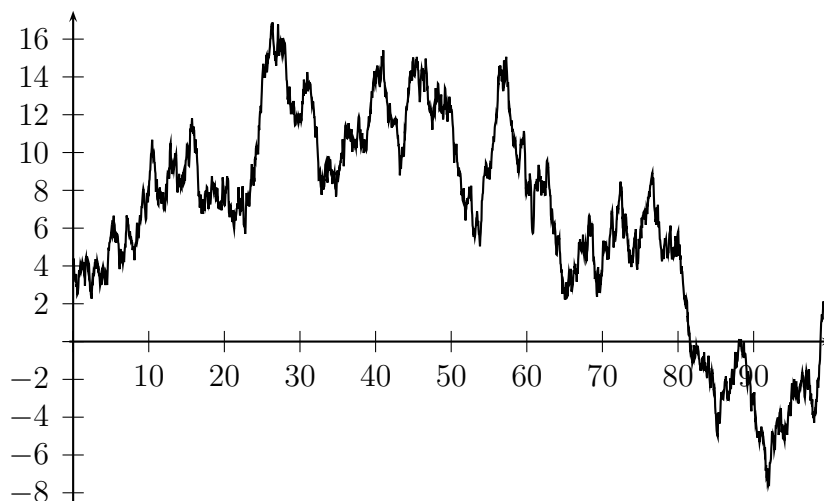


Abbildung 4.5: Graph einer Stichprobe der eindimensionalen Brownschen Bewegung

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

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

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

$$P[\{\omega \in \Omega : \omega(t) \in [a, b]\}] = \frac{1}{\sqrt{2\pi t}} \int_a^b e^{-\frac{x^2}{2t}} dx \quad \text{für alle } t > 0,$$

siehe zum Beispiel die Vorlesung „Stochastische Prozesse“ im Sommersemester.

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

### Konstruktion von $\sigma$ -Algebren

Sei  $\Omega$  eine beliebige Menge, und  $\mathcal{J} \subseteq \mathcal{P}(\Omega)$  eine Kollektion von Ereignissen, die auf jeden Fall in der zu konstruierenden  $\sigma$ -Algebra enthalten sein sollen (z.B. die Mengen aus den Beispielen zu unendlichen Produktmodellen und kontinuierlichen Gleichverteilungen auf Seite 110f).



**Definition.** Die Kollektion

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

von Teilmengen von  $\Omega$  heißt die von  $\mathcal{J}$ -erzeugte  $\sigma$ -Algebra.

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

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

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

heißt **Borel'sche  $\sigma$ -Algebra**. Man prüft leicht nach, dass  $\mathcal{B}(\mathbb{R})$  auch alle abgeschlossenen und halboffenen Intervalle enthält. Die Borel'sche  $\sigma$ -Algebra wird auch erzeugt von der Kollektion aller abgeschlossenen bzw. aller kompakten Intervall. Ebenso gilt:

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

Allgemeiner definieren wir:

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

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

heißt **Borel'sche  $\sigma$ -Algebra auf  $\Omega$** .

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,  $\sigma$ -Algebra.

**Bemerkung.** Nicht jede Teilmenge von  $\mathbb{R}$  ist in der Borelschen  $\sigma$ -Algebra  $\mathcal{B}(\mathbb{R})$  enthalten - ein Beispiel wird in den Übungen gegeben.

Trotzdem 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  $\sigma$ -Algebra auf  $\{0, 1\}^{\mathbb{N}}$ ).** Eine **Zylindermenge** auf dem Folgenraum

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

ist eine Teilmenge  $A$  von  $\Omega$  von der Form

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

In Beispiel 4.2 von oben betrachten wir die von der Kollektion  $\mathcal{C}$  aller Zylindermengen erzeugte  $\sigma$ -Algebra  $\mathcal{A} = \sigma(\mathcal{C})$  auf  $\{0, 1\}^{\mathbb{N}}$ .  $\mathcal{A}$  heißt **Produkt- $\sigma$ -Algebra** auf  $\Omega$ .

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.** Sind  $\mathcal{A}_i, i \in I$   $\sigma$ -Algebren auf  $\Omega_i$ , dann heißt die von der Kollektion  $\mathcal{C}$  aller Zylindermengen

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

$n \in \mathbb{N}, i_1, \dots, i_n \in I, A_{i_1} \in \mathcal{A}_{i_1}, \dots, A_{i_n} \in \mathcal{A}_{i_n}$ , erzeugte  $\sigma$ -Algebra

$$\mathcal{A} = \bigotimes_{i \in I} \mathcal{A}_i := \sigma(\mathcal{C})$$

**Produkt  $\sigma$ -Algebra auf  $\Omega$ .**

Man kann nachprüfen, dass die etwas anders definierte Produkt- $\sigma$ -Algebra aus Beispiel 4.2 ein Spezialfall dieser allgemeinen Konstruktion ist.

## Existenz und Eindeutigkeit von Wahrscheinlichkeitsverteilungen

Sei  $(\Omega, \mathcal{A})$  ein **messbarer Raum**, d.h.  $\Omega$  ist eine nichtleere Menge und  $\mathcal{A} \subseteq \mathcal{P}(\Omega)$  eine  $\sigma$ -Algebra. In der Regel sind 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.** (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)  $A \in \mathcal{J} \Rightarrow A^C \in \mathcal{J}$
- (c)  $A, B \in \mathcal{J} \Rightarrow A \cup B \in \mathcal{J}$ .

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  $\mathcal{J} = \{(-\infty, c] \mid c \in \mathbb{R}\}$ .

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

**Satz 4.6 (Eindeutigkeitsatz).** *Stimmen zwei Wahrscheinlichkeitsverteilungen  $P$  und  $\tilde{P}$  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 Eindeutigkeitsatz 4.6 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  $\sigma$ -additive Abbildung zu einem Maß auf der  $\sigma$ -Algebra fortsetzen lässt, falls  $\mathcal{J}$  eine Algebra ist.

**Satz 4.7 (Fortsetzungssatz von Carathéodory).** *Ist  $\mathcal{J}$  eine Algebra, und  $P : \mathcal{J} \rightarrow [0, \infty]$  eine  $\sigma$ -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.** Ist  $P[\Omega] = 1$ , bzw. allgemeiner  $P[\Omega] < \infty$ , dann ist die Maßfortsetzung nach Satz 4.6 eindeutig, denn eine Algebra ist durchschnittsstabil.

Als Konsequenz aus dem Fortsetzungssatz erhält man:

**Korollar 4.8 (Existenz und Eindeutigkeit der kontinuierlichen Gleichverteilung).** *Es existiert genau eine Wahrscheinlichkeitsverteilung  $\mathcal{U}_{(0,1)}$  auf  $\mathcal{B}((0, 1))$  mit*

$$\mathcal{U}_{(0,1)}[(a, b)] = b - a \quad \text{für alle } 0 < a \leq b < 1. \quad (4.2.4)$$

Zum Beweis ist noch zu zeigen, dass die durch (4.2.4) definierte Abbildung  $\mathcal{U}_{(0,1)}$  sich zu einer  $\sigma$ -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 auf  $\mathcal{A}_0$  aussieht, ist offensichtlich - der Beweis der  $\sigma$ -Additivität ist etwas aufwändiger. Wir verweisen dazu wieder auf die Analysisvorlesung, bzw. den Appendix A1 in Williams: „Probability with martingales.“

**Bemerkung.** (1). Auf ähnliche Weise folgt die Existenz und Eindeutigkeit des durch

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

eindeutig festgelegten Lebesguemaßes  $\lambda$  auf  $\mathcal{B}(\mathbb{R}^d)$ , siehe Analysis III. Man beachte, dass wegen  $\lambda[\mathbb{R}^d] = \infty$  eine Reihe von Aussagen, die wir für Wahrscheinlichkeitsverteilungen beweisen werden, nicht für das Lebesguemaß auf  $\mathbb{R}^d$  gelten!

(2). 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 unmittelbar aus der Gleichverteilung  $\mathcal{U}_{(0,1)}$  konstruieren.

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

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

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

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

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

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

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

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

**Bemerkung.** Für ein Dynkinsystem  $\mathcal{D}$  gilt:

$$A, B \in \mathcal{D}, A \subseteq B \Rightarrow B \setminus A = B \cap A^c = \underbrace{(B^c \cup A)}_{\text{disjunkt}}^c \in \mathcal{D}$$

**Lemma 4.9.** Jedes  $\cap$ -stabile Dynkinsystem  $\mathcal{D}$  ist eine  $\sigma$ -Algebra.

*Beweis.* Für  $A, B \in \mathcal{D}$  gilt:

$$A \cup B = A \cup \underbrace{(B \setminus (A \cap B))}_{\substack{\in \mathcal{D} \text{ falls } \cap\text{-stabil} \\ \in \mathcal{D} \text{ nach Bem.}}} \in \mathcal{D}.$$

$\uparrow$   
 disjunkt

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

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

und damit

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

□

**Lemma 4.10.** Ist  $\mathcal{J}$  ein  $\cap$ -stabiles Mengensystem, so stimmt das von  $\mathcal{J}$  erzeugte Dynkinsystem

$$\mathcal{D}(\mathcal{J}) := \bigcap_{\substack{\mathcal{D} \text{ Dynkinsystem} \\ \mathcal{D} \supseteq \mathcal{J}}} \mathcal{D}$$

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

Aus Lemma (4.10) folgt der Eindeutigkeitsatz, denn  $\{A \in \mathcal{A} \mid P[A] = \tilde{P}[A]\}$  ist ein Dynkinsystem, das  $\mathcal{J}$  enthält, und somit gilt nach dem Lemma

$$\{A \in \mathcal{A} \mid P[A] = \tilde{P}[A]\} \supseteq \mathcal{D}(\mathcal{J}) = \sigma(\mathcal{J}),$$

falls  $\mathcal{J}$  durchschnitts stabil ist.

*Beweis.* (von Lemma (4.10))

Jede  $\sigma$ -Algebra ist ein Dynkingsystem, also gilt  $\mathcal{D}(\mathcal{J}) \subseteq \sigma(\mathcal{J})$ .

Es bleibt zu zeigen, dass  $\mathcal{D}(\mathcal{J})$  eine  $\sigma$ -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:  $\mathcal{D}_B := \{A \in \mathcal{A} \mid A \cap B \in \mathcal{D}(\mathcal{J})\} \supseteq \mathcal{J}$  ist ein Dynkingsystem. Z.B. gilt

$$\begin{aligned} A \in \mathcal{D}_B &\Rightarrow A \cap B \in \mathcal{D}(\mathcal{J}) \\ &\Rightarrow A^C \cap B = \underbrace{B}_{\in \mathcal{D}(\mathcal{J})} \setminus \underbrace{(A \cap B)}_{\in \mathcal{D}(\mathcal{J})} \stackrel{\text{Bem.}}{\in} \mathcal{D}(\mathcal{J}) \\ &\Rightarrow A^C \in \mathcal{D}_B \text{ usw.} \end{aligned}$$

Also gilt  $\mathcal{D}_B \supseteq \mathcal{D}(\mathcal{J})$ , 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:  $\mathcal{D}_A := \{B \in \mathcal{A} \mid A \cap B \in \mathcal{D}(\mathcal{J})\} \supseteq \mathcal{J}$  nach Schritt 1. Zudem ist  $\mathcal{D}_A$  ein Dynkingsystem (Beweis analog zu Schritt 1), also gilt  $\mathcal{D}_A \supseteq \mathcal{D}(\mathcal{J})$ .  $\square$

### 4.3 Allgemeine Zufallsvariablen und ihre Verteilung

Sei  $(\Omega, \mathcal{A}, P)$  ein Wahrscheinlichkeitsraum. Wir wollen nun Zufallsvariablen  $X : \Omega \rightarrow S$  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  $\sigma$ -Algebra. Oft interessieren uns die Wahrscheinlichkeiten von Ereignissen der Form

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

„Der Wert der Zufallsgröße  $X$  liegt in  $B$ “

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

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

## Allgemeine Zufallsvariablen

**Definition.** Eine Abbildung  $X : \Omega \rightarrow S$  heißt **messbar bzgl.  $\mathcal{A}/S$** , falls

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

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

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

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

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

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

Stimmt die  $\sigma$ -Algebra  $\mathcal{S}$  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  $\sigma$ -Algebra  $\mathcal{S}$  zu überprüfen:

**Lemma 4.11.** Sei  $\mathcal{J} \subseteq \mathcal{P}(S)$  mit  $\mathcal{S} = \sigma(\mathcal{J})$ . Dann gilt (M) bereits, falls

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

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

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

$$\begin{aligned} \{X \leq c\} &= \{\omega \in \Omega \mid X(\omega) \leq c\} \in \mathcal{A} \quad \forall c \in \mathbb{R}, & \text{ bzw. wenn} \\ \{X < c\} &= \{\omega \in \Omega \mid X(\omega) < c\} \in \mathcal{A} \quad \forall c \in \mathbb{R}. \end{aligned}$$

*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  $\Omega$  und  $S$  topologische Räume, und  $\mathcal{A}, \mathcal{S}$  die Borelschen  $\sigma$ -Algebren. Dann gilt:

$$X : \Omega \rightarrow S \text{ stetig} \Rightarrow X \text{ messbar.}$$

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

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

folgt die Behauptung. □

Kompositionen von messbaren Abbildungen sind wieder messbar:

**Lemma 4.12.** Sind  $(\Omega_1, \mathcal{A}_1)$ ,  $(\Omega_2, \mathcal{A}_2)$  und  $(\Omega_3, \mathcal{A}_3)$  messbare Räume, und ist  $X_1 : \Omega_1 \rightarrow \Omega_2$  messbar bzgl.  $\mathcal{A}_1/\mathcal{A}_2$  und  $X_2 : \Omega_2 \rightarrow \Omega_3$  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  $X : \Omega \rightarrow \mathbb{R}$  eine reellwertige Zufallsvariable und  $f : \mathbb{R} \rightarrow \mathbb{R}$  eine messbare (z.B. stetige) Funktion, dann ist auch

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

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



(2). Sind  $X, Y : \Omega \rightarrow \mathbb{R}$  reellwertige Zufallsvariablen, dann ist  $(X, Y) : \omega \mapsto (X(\omega), Y(\omega))$  eine messbare Abbildung in den  $\mathbb{R}^2$  mit Borelscher  $\sigma$ -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

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

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

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

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  $\sigma$ -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.** Die Wahrscheinlichkeitsverteilung  $\mu_X$  auf  $(S, \mathcal{S})$  heißt **Bild von  $P$  unter  $X$  oder Verteilung (law) von  $X$  unter  $P$ .**

Für  $\mu_X$  werden häufig auch die folgenden Notationen verwendet:

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

### Charakterisierung der Verteilung

- Diskrete Zufallsvariablen:

Die Verteilung  $\mu_X$  einer diskreten Zufallsvariablen ist eindeutig durch die **Massenfunktion**

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

festgelegt.

- Reelle Zufallsvariablen

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

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

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

**Definition.** Die Funktion  $F_X : \mathbb{R} \rightarrow [0, 1]$ ,

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

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

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

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

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  $(a, b)$ .

**Beispiel** (Exponentialverteilung). Angenommen, wir wollen die Wartezeit auf das erste Eintreten eines unvorhersehbaren Ereignisses (radioaktiver Zerfall, Erdbeben, ...) mithilfe einer Zufallsvariable  $T : \Omega \rightarrow (0, \infty)$  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  $(\frac{(k-1)t}{n}, \frac{kt}{n}]$ ,  $1 \leq k \leq n$ .

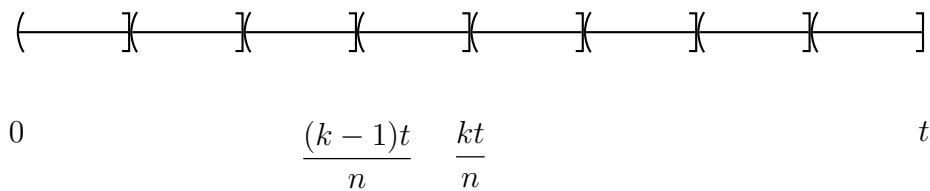


Abbildung 4.6: Unterteilung des Intervalls  $(0, t]$  in  $n$  Teile.

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  $n \rightarrow \infty$  immer genauer wird. Damit erhalten wir:

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

Für  $n \rightarrow \infty$  konvergiert die rechte Seite gegen  $e^{-\lambda t}$ .

Daher liegt folgende Definition nahe:

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

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

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

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

bzw. mit Verteilungsfunktion

$$F(t) = \mu[(-\infty, t]] = \begin{cases} 1 - e^{-\lambda t} & \text{für } t \geq 0 \\ 0 & \text{für } t < 0. \end{cases} \quad (4.3.1)$$

Nach dem Eindeutigkeitsatz ist die  $\text{Exp}(\lambda)$ -Verteilung durch (4.3.1) eindeutig festgelegt.

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

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

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

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

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

**Algorithmus 4.14 (Simulation einer exponentialverteilten Stichprobe).**

**Input:** Intensität  $\lambda > 0$

**Output:** Stichprobe  $x$  von  $\text{Exp}(\lambda)$

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

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

Wir werden in Abschnitt 4.5 zeigen, dass mit einem entsprechenden Verfahren beliebige reelle Zufallsvariablen konstruiert und simuliert werden können. Zum Abschluss dieses Abschnitts zeigen wir noch eine bemerkenswerte Eigenschaft exponentialverteilter Zufallsvariablen:

**Satz 4.15 (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.*

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

□

## 4.4 Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$

In diesem und im nächsten Abschnitt beschäftigen wir uns systematischer mit der Beschreibung, Konstruktion und Simulation reellwertiger Zufallsvariablen. Wir notieren dazu zunächst einige grundlegende Eigenschaften der Verteilungsfunktion

$$F(c) = P[X \leq c]$$

einer auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  definierten Zufallsvariable  $X : \Omega \rightarrow \mathbb{R}$ . Wir werden im nächsten Abschnitt sehen, dass umgekehrt jede Funktion mit den Eigenschaften (1)-(3) aus Satz 4.16 die Verteilungsfunktion einer reellen Zufallsvariable ist.

### Eigenschaften der Verteilungsfunktion

**Satz 4.16.** Für die Verteilungsfunktion  $F : \mathbb{R} \rightarrow [0, 1]$  einer reellwertigen Zufallsvariable  $X$  gilt:

- (1).  $F$  ist **monoton wachsend**,
- (2).  $\lim_{c \rightarrow -\infty} F(c) = 0$  und  $\lim_{c \rightarrow \infty} F(c) = 1$ ,
- (3).  $F$  ist **rechtsstetig**, d.h.  $F(c) = \lim_{y \searrow c} F(y)$  für alle  $c \in \mathbb{R}$ ,
- (4).  $F(c) = \lim_{y \nearrow c} F(y) + \mu_X[\{c\}]$ .

Insbesondere ist  $F$  stetig bei  $c$ , falls  $\mu_X[\{c\}] = 0$  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:

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

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

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

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

Im Folgenden betrachten wir einige Beispiele von eindimensionalen Verteilungen und ihren Verteilungsfunktionen.

## Diskrete Verteilungen

Die Verteilung  $\mu$  einer reellen Zufallsvariable  $X$  heißt diskret, wenn  $\mu[S] = 1$  für eine abzählbare Menge  $S$  gilt.

**Beispiele.** (1). BERNOULLI-VERTEILUNG MIT PARAMETER  $p \in [0, 1]$ :

$$\mu[\{1\}] = p, \quad \mu[\{0\}] = 1 - p.$$

Als Verteilungsfunktion ergibt sich

$$F(c) = \begin{cases} 0 & \text{für } c < 0 \\ 1 - p & \text{für } c \in [0, 1) \\ 1 & \text{für } c \geq 1. \end{cases}$$

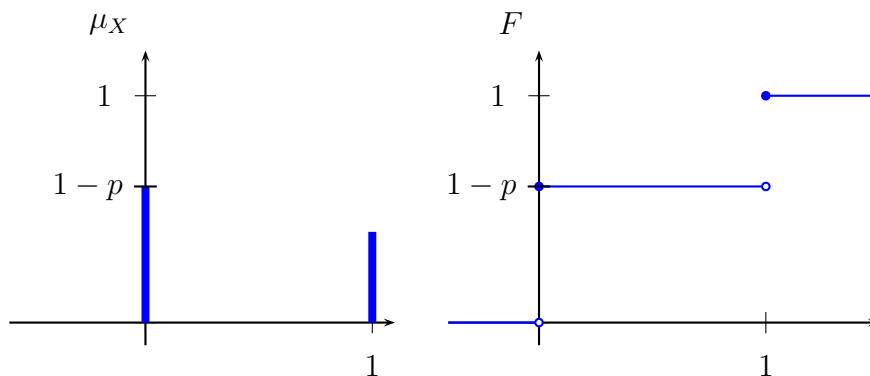


Abbildung 4.7: Massen- und Verteilungsfunktion einer  $Ber(p)$ -verteilten Zufallsvariablen.

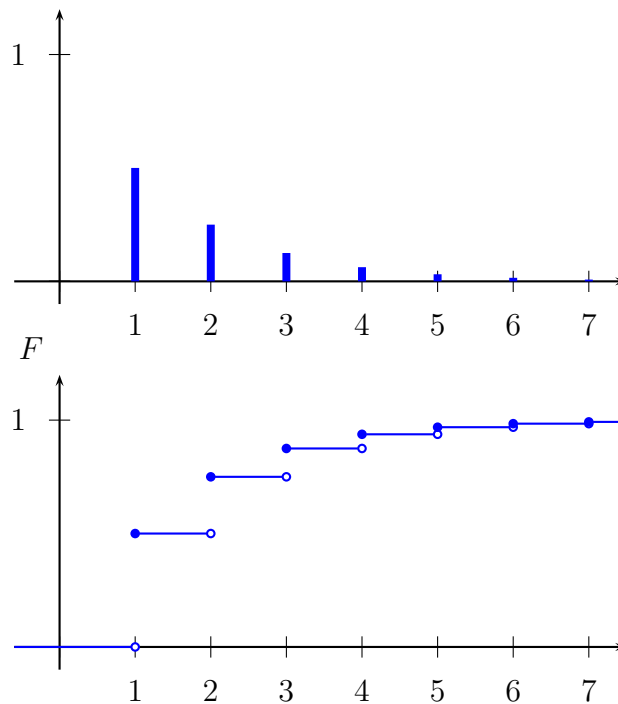
(2). GEOMETRISCHE VERTEILUNG MIT PARAMETER  $p \in [0, 1]$ :

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

Für eine geometrisch verteilte Zufallsvariable  $T$  gilt:

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

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

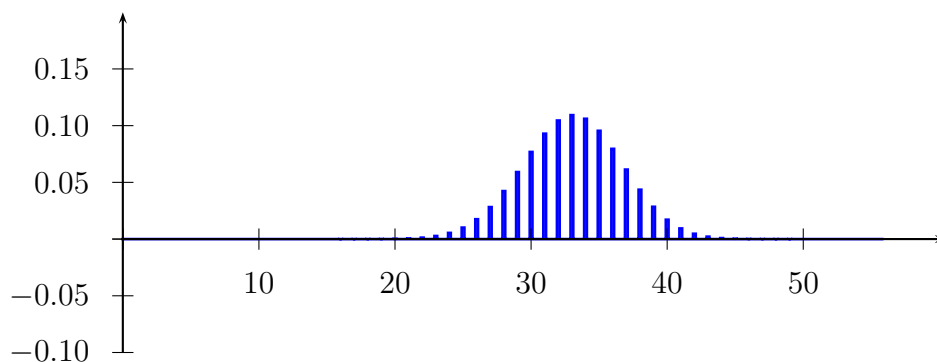
Abbildung 4.8: Massen- und Verteilungsfunktion einer  $Geom(\frac{1}{2})$ -verteilten Zufallsvariablen.

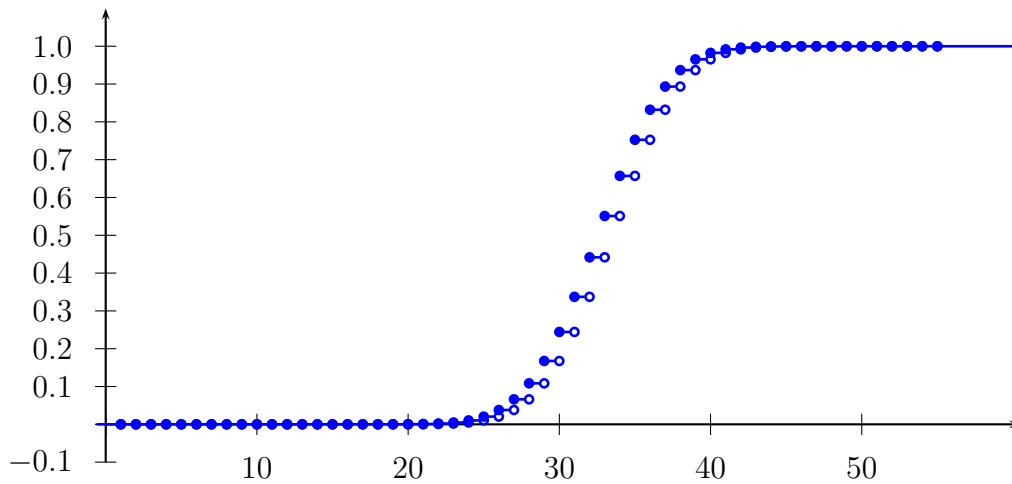
(3). BINOMIALVERTEILUNG MIT PARAMETERN  $n$  UND  $p$ :

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

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

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

Abbildung 4.9: Massenfunktion einer  $\text{Bin}(55, 0.6)$ -verteilten Zufallsvariable.

Abbildung 4.10: Verteilungsfunktion von  $\text{Bin}(55, 0.6)$ 

Allgemein sind die Unstetigkeitsstellen der Verteilungsfunktion  $F$  einer reellwertigen Zufallsvariable  $X$  nach Satz 4.16 (4) 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.

## Stetige Verteilungen

Die Verteilung  $\mu$  einer reellen Zufallsvariable  $X$  heißt **stetig**, bzw. **absolutstetig**, falls eine integrierbare Funktion  $f : \mathbb{R} \rightarrow [0, \infty)$  existiert mit

$$F(c) = P[X \leq c] = \mu[(-\infty, c]] = \int_{-\infty}^c f(x) \, dx \quad \text{für alle } c \in \mathbb{R}. \quad (4.4.1)$$

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

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

**Definition.** Eine Lebesgue-integrierbare Funktion  $f : \mathbb{R} \rightarrow [0, \infty)$  mit (4.4.1) heißt **Dichtefunktion** der Zufallsvariable  $X$  bzw. der Verteilung  $\mu$ .

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

$$F'(x) = f(x) \quad (4.4.2)$$



für alle  $x \in \mathbb{R}$ , falls  $f$  stetig ist. Im Allgemeinen gilt (4.4.2) für  $\lambda$ -fast alle  $x$ , wobei  $\lambda$  das Lebesguemaß auf  $\mathbb{R}$  ist.

(2). Aus (4.4.1) folgt aufgrund der Eigenschaften des Lebesgueintegrals (s. Kapitel 6 unten):

$$P[X \in B] = \mu_X[B] = \int_B f(x) dx, \quad (4.4.3)$$

für alle Mengen  $B \in \mathcal{B}(\mathbb{R})$ . Zum Beweis zeigt man, dass beide Seiten von (4.4.3) Wahrscheinlichkeitsverteilungen definieren, und wendet den Eindeutigkeitssatz an.

**Beispiele.** (1). GLEICHVERTEILUNG AUF  $(a, b)$  ( $-\infty < a < b < \infty$ ).

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

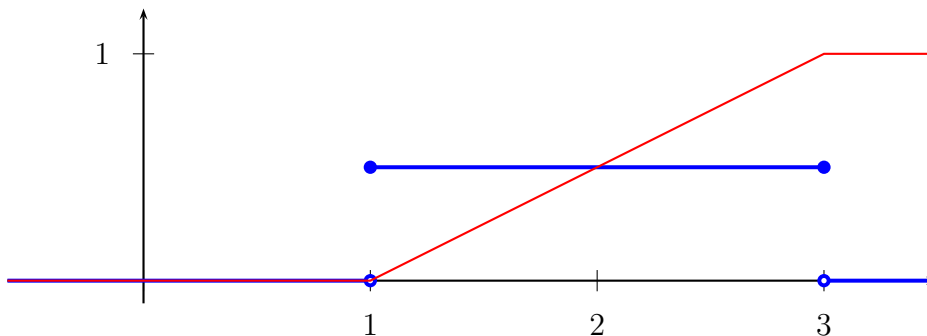


Abbildung 4.11: Dichte  $f(x) = \mathbb{1}_{[1,3]}(x)$  einer uniform auf  $[1, 3]$  verteilten Zufallsvariable (blau), und deren Verteilungsfunktion  $F(c)$  (rot)

Affine Funktionen von gleichverteilten Zufallsvariablen sind wieder gleichverteilt.

(2). EXPONENTIALVERTEILUNG MIT PARAMETER  $\lambda > 0$ .

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

$$F(c) = \mu[(-\infty, c]] = (1 - e^{-\lambda c})^+ = \int_c^\infty f(x) dx.$$

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

$$P[aT > c] = P\left[T > \frac{c}{a}\right] = e^{-\frac{\lambda}{a}c} \quad \text{für alle } c \geq 0.$$

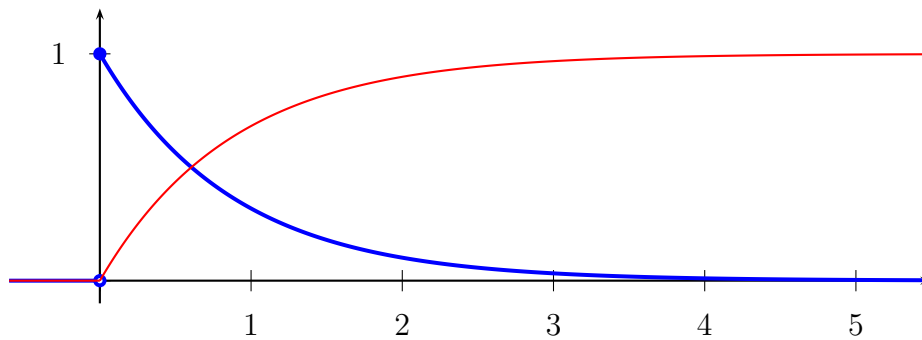


Abbildung 4.12: Dichte  $f(x) = \mathbb{1}_{[0,\infty)}(x) \cdot e^{-x}$  einer zum Parameter 1 exponentialverteilten Zufallsvariable (blau) und deren Verteilungsfunktion  $F(c)$  (rot)

### (3). NORMALVERTEILUNGEN

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

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

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

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

der Standardnormalverteilung ist i.A. 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

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

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

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

**Definition.** 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  $\sigma^2$** . 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.).

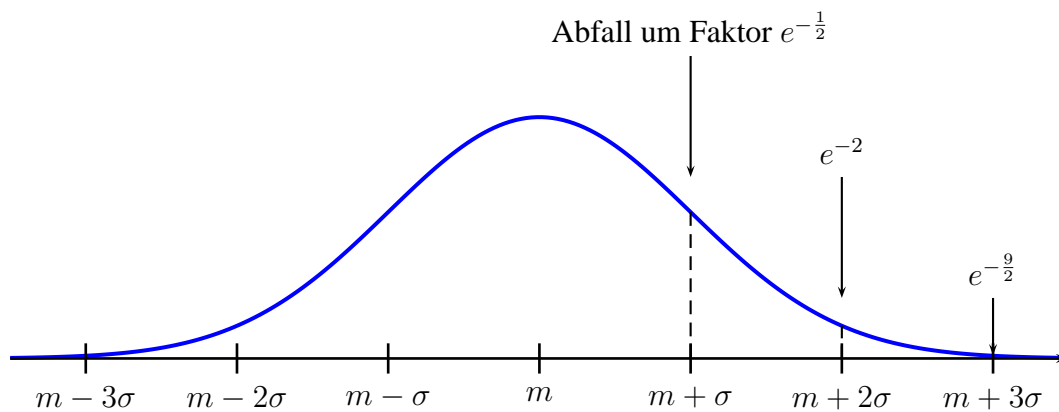


Abbildung 4.13: Dichte einer normalverteilten Zufallsvariable mit Mittelwert  $m$  und Varianz  $\sigma^2$ .

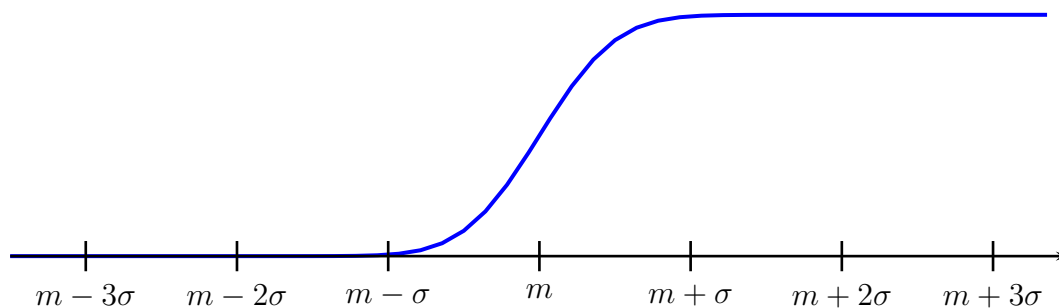


Abbildung 4.14: Verteilungsfunktion einer normalverteilten Zufallsvariable mit Mittelwert  $m$  und Varianz  $\sigma^2$ .

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

$$f_{m,\sigma}(m \pm \sigma) = \frac{f_{m,\sigma}(m)}{\sqrt{e}}$$

$$f_{m,\sigma}(m \pm 2\sigma) = \frac{f_{m,\sigma}(m)}{e^2}$$

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

Für die Wahrscheinlichkeit, dass eine normalverteilte Zufallsvariable Werte außerhalb der  $\sigma$ -,  $2\sigma$ - und  $3\sigma$ -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  $\sigma$  vom Mittelwert  $m$  ist also für eine normalverteilte Zufallsvariable relativ typisch, eine Abweichung der Größe  $3\sigma$  dagegen schon sehr selten.

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

**Lemma 4.17.** Für eine standardnormalverteilte Zufallsvariable  $Z$  gilt:

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

*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:

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

also

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

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

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

□

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

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

sowie eine entsprechende Abschätzung nach unten.

### Transformation von absolutstetigen Zufallsvariablen

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

**Satz 4.18 (Eindimensionaler Dichtetransformationssatz).** *Ist  $\Phi : I \rightarrow J$  einmal stetig differenzierbar mit  $\Phi'(x) \neq 0$  für alle  $x \in I$ , dann ist die Verteilung von  $\Phi(X)$  absolutstetig mit Dichte*

$$f_{\Phi(X)}(y) = \begin{cases} f_X(\Phi^{-1}(y)) \cdot |(\Phi^{-1})'(y)| & \text{für } y \in \Phi(I) \\ 0 & \text{sonst} \end{cases}. \quad (4.4.4)$$

*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  $\Phi$  streng monoton wachsend ist, also eine Bijektion von  $I$  nach  $\Phi(I)$ . Daher erhalten wir

$$F_{\Phi(X)}(c) = P[\Phi(X) \leq c] = P[X \leq \Phi^{-1}(c)] = F_X(\Phi^{-1}(c))$$

für alle  $c \in \Phi(I)$ . Nach der Kettenregel ist dann  $F_{\Phi(X)}$  für fast alle  $c \in \Phi(I)$  differenzierbar, und es gilt

$$F'_{\Phi(X)}(c) = f_X(\Phi^{-1}(c)) \cdot (\Phi^{-1})'(c).$$

Die Behauptung folgt hieraus nach dem Hauptsatz der Differential- und Integralrechnung, da

$$P[\Phi(x) \notin \Phi(I)] = 0.$$

□

**Beispiel (Geometrische Wahrscheinlichkeiten).** Sei  $\theta : \Omega \rightarrow [0, 2\pi)$  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 (4.4.4) 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 4.18 berechnen. Wegen

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

erhalten wir, dass  $\cos \theta$  eine sogenannte „Halbkreisverteilung“ mit Dichte

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

hat.

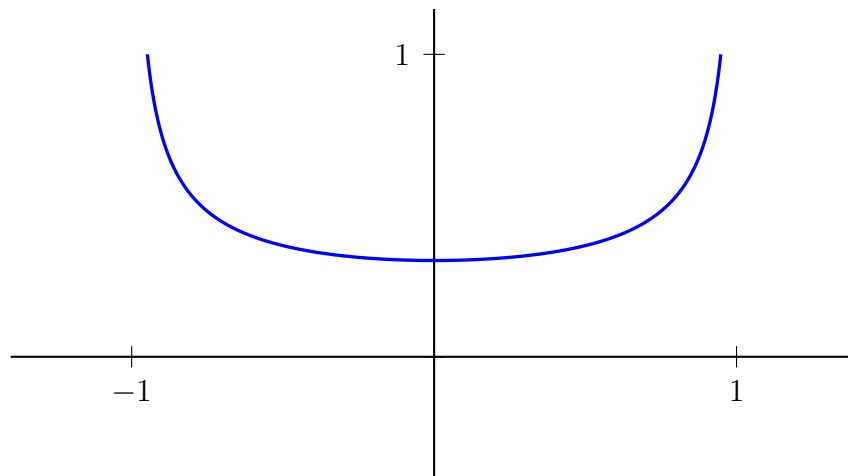


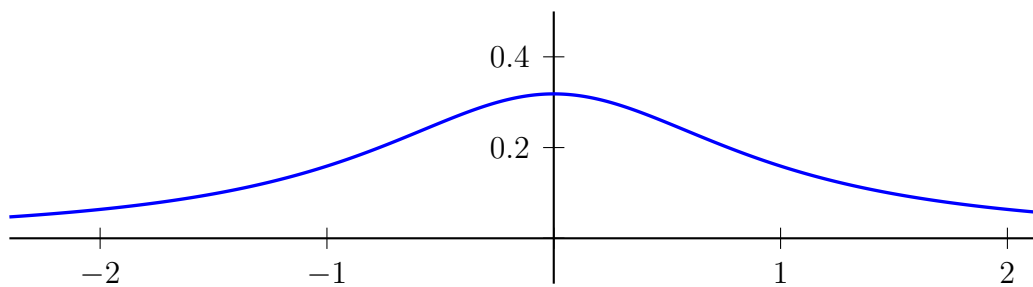
Abbildung 4.15: Abbildung der Dichtefunktion  $f_{\cos \theta}$

Anstelle von (4.4.4) gilt in diesem Fall

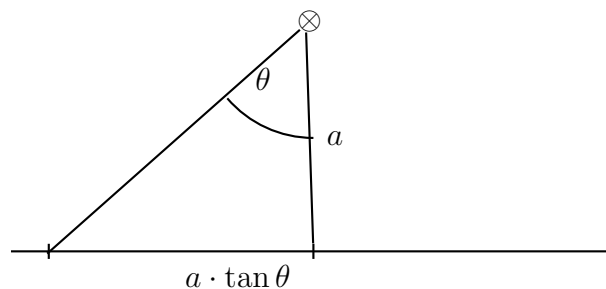
$$f_{\cos \theta}(x) = f_X(\psi_1(x)) \cdot |\psi_1'(x)| + f_X(\psi_2(x)) \cdot |\psi_2'(x)|,$$

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. Auf ähnliche Weise zeigt man für  $a > 0$  (Übung):

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

Abbildung 4.16: Abbildung der Dichtefunktion  $f_{a \tan \theta}$ 

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.



## 4.5 Quantile und Inversionsverfahren

Quantile sind Stellen, an denen die Verteilungsfunktion einen bestimmten Wert überschreitet. Mithilfe von Quantilen kann man daher verallgemeinerte Umkehrfunktionen der im Allgemeinen nicht bijektiven Verteilungsfunktion definieren. Diese Umkehrabbildungen werden wir nutzen, um reellwertige Zufallsvariablen mit einer gegebenen Verteilungsfunktion explizit zu konstruieren.

### Quantile

In praktischen Anwendungen (z.B. Qualitätskontrolle) müssen häufig Werte berechnet werden, sodass ein vorgegebener Anteil der Gesamtmasse einer Wahrscheinlichkeitsverteilung auf  $\mathbb{R}$  unterhalb dieses Wertes liegt. Sei  $X : \Omega \rightarrow \mathbb{R}$  eine Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  mit Verteilungsfunktion  $F$ .

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

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

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

Ist die Verteilungsfunktion nicht streng monoton wachsend, dann kann es mehrere  $u$ -Quantile zu einem Wert  $u$  geben.

**Beispiel (Stichprobenquantile).** Wir betrachten eine Stichprobe, die aus  $n$  reellwertigen Daten / Messwerten  $x_1, \dots, x_n$  mit  $x_1 \leq x_2 \leq \dots \leq x_n$  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

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

die relative Häufigkeit des Bereichs  $B$  unter den Messwerten  $x_i$ . Die empirische Verteilung ergibt sich, wenn wir zufällig ein  $i \in \{1, \dots, n\}$  wählen, und den entsprechenden Messwert betrachten.

Die Quantile der empirischen Verteilung bezeichnet man als **Stichprobenquantile**. Für  $u \in [0, 1]$  sei

$$k_u := 1 + (n - 1)u \in [1, n].$$

Ist  $k_u$  ganzzahlig, dann ist  $x_{k_u}$  das eindeutige  $u$ -Quantil der Stichprobe. Allgemein ist jedes  $q \in [x_{\lfloor k_u \rfloor}, x_{\lceil k_u \rceil}]$  ein  $u$ -Quantil der Stichprobe, d.h. für  $k_u \notin \mathbb{Z}$  gibt es mehrere  $u$ -Quantile.

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

$$\underline{G}(u) := \inf\{x \in \mathbb{R} | F(x) \geq u\} = \sup\{x \in \mathbb{R} | F(x) < u\}$$

und

$$\overline{G}(u) := \inf\{x \in \mathbb{R} | F(x) > u\} = \sup\{x \in \mathbb{R} | F(x) \leq u\}.$$

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

**Satz 4.19.** Für  $u \in (0, 1)$  und  $q \in \mathbb{R}$  sind die folgenden Aussagen äquivalent:

- (1).  $q$  ist ein  $u$ -Quantil.



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

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

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 (Übung!).  $\square$

## Konstruktion und Simulation reellwertiger Zufallsvariablen

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

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

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

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

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

eine Zufallsvariable mit Verteilung  $\mu$  definiert, denn

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

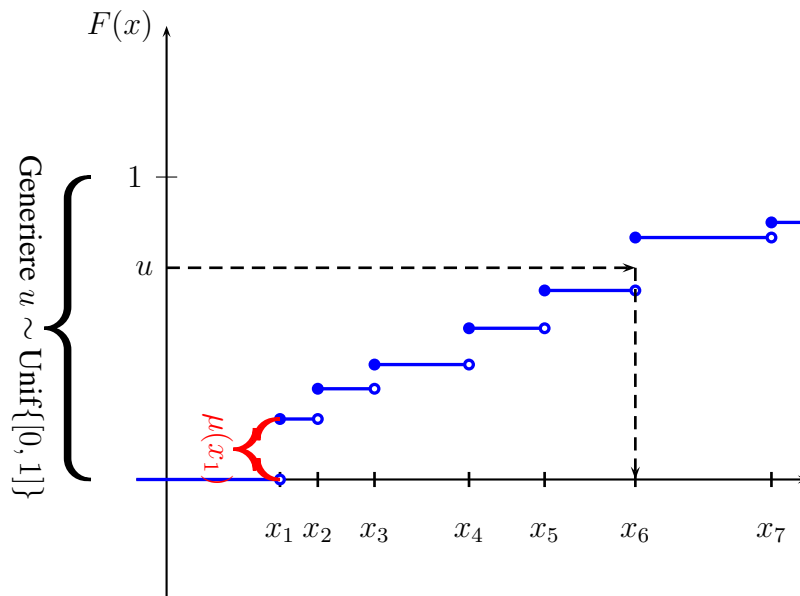


Abbildung 4.17: Wir generieren eine uniform auf  $(0, 1)$  verteilte Pseudozufallszahl  $u$ . Suche nun das minimale  $k \in \mathbb{N}$ , für das  $\sum_{i=1}^k \mu(x_i) > u$ . Dann ist  $x = x_k$  eine Pseudozufallsstichprobe von der Verteilung  $\mu$ .

**Allgemeiner Fall:** Wir wollen das Vorgehen nun verallgemeinern. Sei  $F : \mathbb{R} \rightarrow [0, 1]$  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 4.20.** Ist  $F : \mathbb{R} \rightarrow [0, 1]$  eine Funktion mit (1)-(3), und

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

die linksstetige verallgemeinerte Inverse, dann ist das Bild

$$\mu := \mathcal{U}_{(0,1)} \circ \underline{G}^{-1}$$

der Gleichverteilung auf  $(0, 1)$  unter  $\underline{G}$  eine Wahrscheinlichkeitsverteilung auf  $\mathbb{R}$  mit Verteilungsfunktion  $F$ .

Insbesondere gilt: Ist  $U : \Omega \rightarrow (0, 1)$  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

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

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:

$$\begin{aligned} P[\underline{G}(U) \leq c] &= \mathcal{U}_{(0,1)}[\{u \in (0, 1) \mid \underbrace{\underline{G}(u) \leq c}_{\iff F(c) \geq u}\}] \\ &= \mathcal{U}_{(0,1)}[(0, F(c))] = F(c). \end{aligned}$$

Also ist  $F$  die Verteilungsfunktion von  $\underline{G}(U)$  bzw. von  $\mu$ . □

**Bemerkung.** (1). Ist  $F$  eine Bijektion von  $\mathbb{R}$  nach  $(0, 1)$  (also stetig und streng monoton wachsend), dann ist  $\underline{G} = F^{-1}$ .

(2). Nimmt  $X$  nur endlich viele Werte  $x_1 < x_2 < \dots < x_n$  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  $\mu$  auf  $\mathbb{R}$  mit Verteilungsfunktion  $F$ .

Zudem erhalten wir einen expliziten Algorithmus zur Simulation einer Stichprobe von  $\mu$ :

**Algorithmus 4.21 (Inversionsverfahren zur Simulation einer Stichprobe  $x$  von  $\mu$ ).**

(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.

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

$$F = (1 - p) \cdot I_{[0, 1)} + 1 \cdot I_{[1, \infty)}$$

und  $\underline{G} = \mathbb{1}_{(1-p, 1)}$ , siehe Abbildung 4.18.

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

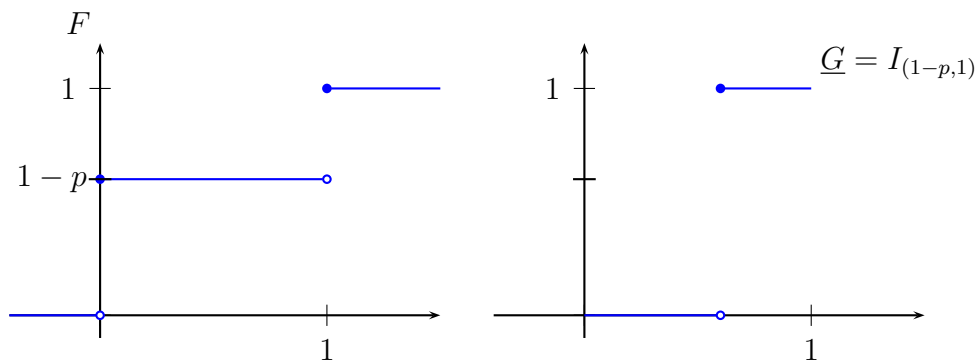


Abbildung 4.18:  $\underline{G}(U) = I_{\{U > 1-p\}}$  ist Bernoulli( $p$ )-verteilt.

(2). GLEICHVERTEILUNG AUF  $(a, b)$ :

$$F(c) = \frac{c - a}{b - a} \quad \text{für } c \in [a, b],$$

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

siehe Abbildung 4.19. Also ist  $a + (b - a)U$  für  $U \sim \mathcal{U}_{(0,1)}$  gleichverteilt auf  $(a, b)$ .

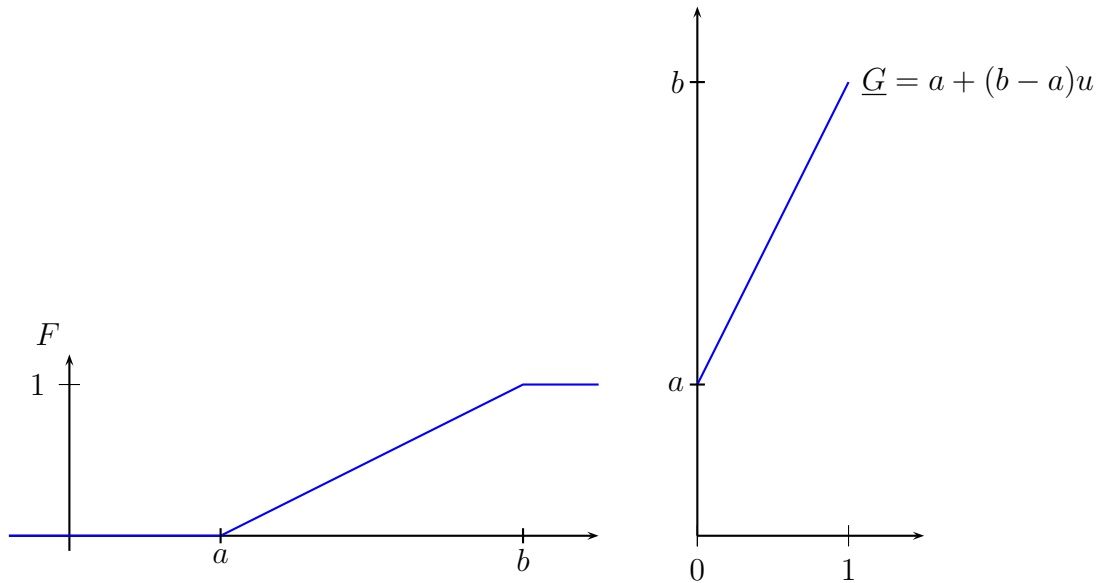


Abbildung 4.19:  $\underline{G}(u) = a + (b - a)u$  ist (für  $u \sim \text{unif}\{(0, 1)\}$ ) uniform auf  $(a, b)$  verteilt.

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

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

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

## 4.6 Normalapproximation der Binomialverteilung

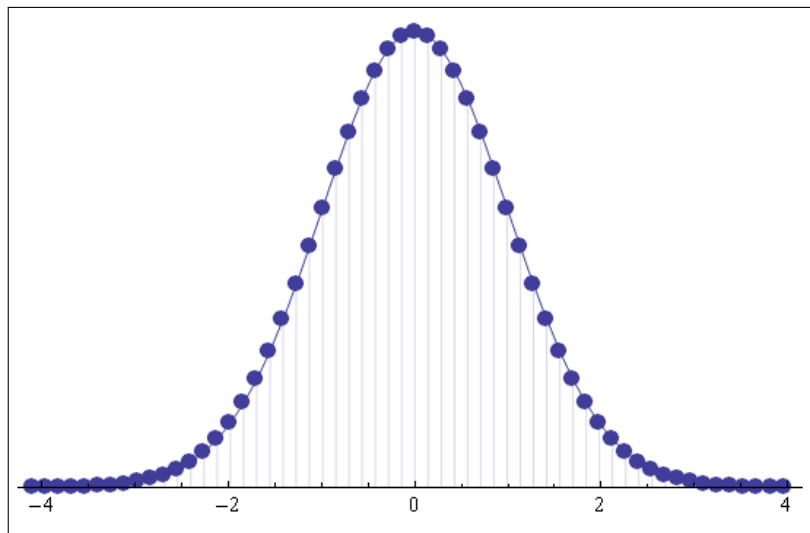
Die Binomialverteilung mit Parametern  $n$  und  $p$  beschreibt die Verteilung der Anzahl derjenigen unter  $n$  unabhängigen Ereignissen mit Wahrscheinlichkeit  $p$ , die in einem Zufallsexperiment eintreten. Viele Anwendungsprobleme führen daher auf die Berechnung von Wahrscheinlichkeiten bzgl. der 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 näherungsweise Berechnung nutzen:

Konvergiert  $n \rightarrow \infty$ , 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  $\lambda$  an:

$$b_{n,p_n}(k) = \binom{n}{k} p_n^k (1 - p_n)^{n-k} \rightarrow \frac{\lambda^k}{k!} e^{-\lambda} \quad (k = 0, 1, 2, \dots),$$

siehe Satz 1.5. Geht die Wahrscheinlichkeit  $p_n$  für  $n \rightarrow \infty$  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.

### Der Satz von De Moivre - Laplace

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

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

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 1$$

gilt.

#### Bemerkung.

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

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

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

#### Satz 4.22 (Stirlingsche Formel).

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

Zum Beweis nimmt man den Logarithmus, und schätzt die sich ergebende Summe mithilfe eines Integrals ab, siehe z.B. Forster: „Analysis I“.

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  $\text{Bin}(n, p)$ -verteilte Zufallsvariable auf  $(\Omega, \mathcal{A}, P)$ . Für den Erwartungswert und die Standardabweichung von  $S_n$  gilt:

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

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.

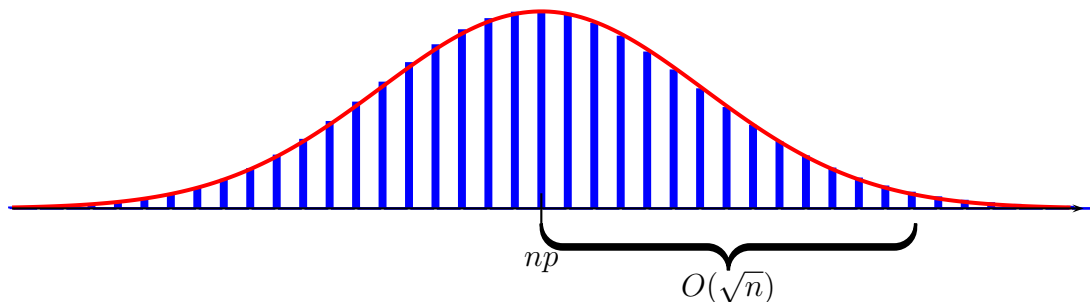


Abbildung 4.20: Die Gewichte der Binomialverteilung liegen für große  $n$  näherungsweise auf einer Glockenkurve mit Mittel  $np$  und Standardabweichung  $\sqrt{np(1-p)}$ .

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 einer Umgebung 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

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

falls

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

wobei

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

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 4.23 (de Moivre 1733, Laplace 1819).** Sei  $p \in (0, 1)$  und  $\sigma^2 = p(1-p)$ . Dann gilt:

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

$$(2). P\left[a \leq \frac{S_n - np}{\sqrt{n}} \leq b\right] \xrightarrow{n \nearrow \infty} \int_a^b \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}}_{\text{Gaußsche Glockenkurve}} dx$$



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

(a) Wir zeigen zunächst mithilfe der **Stirlingschen Formel**:

$$b_{n,p}(k) \approx \frac{1}{\sqrt{2\pi n \frac{k}{n} (1 - \frac{k}{n})}} \cdot \left(\frac{p}{\frac{k}{n}}\right)^k \cdot \left(\frac{1-p}{1 - \frac{k}{n}}\right)^{n-k} =: \bar{b}_{n,p}(k) \quad (4.6.1)$$

Es gilt

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

Für  $k \in U_{n,r}$  gilt

$$k \geq np - A \cdot \sqrt{n} \xrightarrow{n \rightarrow \infty} \infty,$$

also folgt

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

d.h.

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

Analog erhält man

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

und damit

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

(b) Wir zeigen nun mithilfe einer **Taylorapproximation**:

$$\bar{b}_{n,p}(k) \approx \tilde{b}_{n,p}(k) \quad (4.6.2)$$

Für  $k \in U_{n,r}$  gilt

$$\left| \frac{k}{n} - p \right| \leq r \cdot n^{-\frac{1}{2}},$$

woraus folgt:

$$\sqrt{2\pi \cdot n \cdot \frac{k}{n} \cdot \left(1 - \frac{k}{n}\right)} \approx \sqrt{2\pi \cdot n \cdot p \cdot (1-p)} = \sqrt{2\pi \cdot n \cdot \sigma^2}. \quad (4.6.3)$$

Um die Asymptotik der übrigen Faktoren von  $\bar{b}_{n,p}(k)$  zu erhalten, verwenden wir eine **Taylorapproximation für den Logarithmus**:

Wegen

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

gilt:

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

Für  $k \in U_{n,r}$  gilt:

$$\left| \frac{k}{n} - p \right|^3 \leq r^3 \cdot n^{-\frac{3}{2}}.$$

Also folgt:

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

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

$$\left( \frac{p}{\frac{k}{n}} \right)^k \left( \frac{1-p}{1-\frac{k}{n}} \right)^{n-k} \approx \exp \left( -\frac{1}{2\sigma^2} \left( \frac{k}{n} - p \right)^2 \right). \quad (4.6.4)$$

Aussage (4.6.2) folgt dann aus (4.6.3) und (4.6.4).

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

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

$$\begin{aligned} P \left[ a \leq \frac{S_n - np}{\sqrt{n}} \leq b \right] &= \sum_{\substack{k \in \{0,1,\dots,n\} \\ a \leq \frac{k-np}{\sqrt{n}} \leq b}} \underbrace{P[S_n = k]}_{=b_{n,p}(k) \approx \tilde{b}_{n,p}(k)} \\ &= \sum_{\substack{k \in \{0,1,\dots,n\} \\ a \leq \frac{k-np}{\sqrt{n}} \leq b}} \tilde{b}_{n,p}(k) (1 + \varepsilon_{n,p}(k)), \end{aligned}$$

wobei

$$\bar{\varepsilon}_{n,p} := \sup_{a \leq \frac{k-np}{\sqrt{n}} \leq b} |\varepsilon_{n,p}(k)| \longrightarrow 0 \quad \text{für } n \rightarrow \infty. \tag{4.6.5}$$

Wir zeigen nun

$$\lim_{n \rightarrow \infty} \sum_{\substack{k \in \{0,1,\dots,n\} \\ a \leq \frac{k-np}{\sqrt{n}} \leq b}} \tilde{b}_{n,p}(k) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \tag{4.6.6}$$

Aus (4.6.5) und (4.6.6) folgt dann die Behauptung, da

$$\left| \sum_{\substack{k \in \{0,1,\dots,n\} \\ a \leq \frac{k-np}{\sqrt{n}} \leq b}} \tilde{b}_{n,p}(k) \cdot \varepsilon_{n,p}(k) \right| \leq \underbrace{\bar{\varepsilon}_{n,p}}_{\rightarrow 0} \cdot \underbrace{\sum_{\substack{k \in \{0,1,\dots,n\} \\ a \leq \frac{k-np}{\sqrt{n}} \leq b}} \tilde{b}_{n,p}(k)}_{\rightarrow \int_a^b \dots dx < \infty} \xrightarrow{n \rightarrow \infty} 0$$

Zum Beweis von (4.6.6) sei

$$\Gamma_n := \left\{ \frac{k - np}{\sqrt{n}} \mid k = 0, 1, \dots, n \right\} \subseteq \mathbb{R}.$$

Dann ist  $\Gamma_n$  ein äquidistantes Gitter mit Maschenweite  $\Delta = \frac{1}{\sqrt{n}}$ , und es gilt

$$\sum_{\substack{k \in \{0,1,\dots,n\} \\ a \leq \frac{k-np}{\sqrt{n}} \leq b}} \tilde{b}_{n,p}(k) = \sum_{\substack{x \in \Gamma_n \\ a \leq x \leq b}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \Delta x.$$

Für  $n \rightarrow \infty$  folgt (4.6.6), da die rechte Seite eine Riemannsummenapproximation des Integrals ist.

□

Der Satz von de Moivre/Laplace besagt, dass die Verteilungen der Zufallsvariablen  $\frac{S_n - np}{\sqrt{n}}$  für  $n \rightarrow \infty$  **schwach** gegen die Normalverteilung  $N(0, \sigma^2)$  mit Varianz  $\sigma^2 = p(1-p)$  **konvergieren**. Die allgemeine Definition der schwachen Konvergenz einer Folge von Wahrscheinlichkeitsverteilungen wird in Abschnitt 8.3 unten gegeben. Ist  $Z$  eine standardnormalverteilte Zufallsvariable, dann gilt:

$$\frac{S_n - np}{\sqrt{n}} \xrightarrow{\mathcal{D}} \sigma Z,$$

bzw.

$$\frac{S_n - E[S_n]}{\sigma(S_n)} = \frac{S_n - np}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} Z, \quad (4.6.7)$$

wobei „ $\xrightarrow{\mathcal{D}}$ “ für schwache Konvergenz der Verteilungen der Zufallsvariablen steht (**Konvergenz in Verteilung**, s.u.).

**Bemerkung.** (1). Die Aussage (4.6.7) ist ein Spezialfall eines viel allgemeineren zentralen Grenzwertsatzes:

Sind  $X_1, X_2, \dots$  unabhängige, identisch verteilte Zufallsvariablen mit endlicher Varianz, und ist  $S_n = X_1 + \dots + X_n$ , 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 (4.6.7)

$$S_n \stackrel{\mathcal{D}}{\approx} np + \sqrt{np(1-p)} \cdot Z, \quad (4.6.8)$$

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

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

Entsprechende „Approximationen“ werden häufig in Anwendungen benutzt, sollten aber hinterfragt werden, da beim Übergang von (4.6.7) zu (4.6.8) 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  $X_1, X_2, \dots$  unabhängige Zufallsvariablen mit  $P[X_i = 0] = P[X_i = 1] = \frac{1}{2}$  und sei  $S_n = X_1 + \dots + X_n$  (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:**

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

Da  $\frac{S_{100} - E[S_{100}]}{\sigma(S_{100})}$  nach (4.6.7) näherungsweise  $N(0, 1)$ -verteilt ist, und  $\frac{10}{\sigma\sqrt{100}} = 2$ , folgt

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

(2). **16 faire Münzwürfe:**

$$\begin{aligned} P[S_{16} = 8] &= P[7.5 \leq S_{16} \leq 8.5] = P[|S_{16} - E[S_{16}]| \leq 0.5] \\ &= P\left[\left|\frac{S_{16} - E[S_{16}]}{\sigma(S_{16})}\right| \leq \frac{0.5}{\sigma\sqrt{16}}\right] \end{aligned}$$

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

$$P[S_{16} = 8] \approx P[|Z| \leq 1.4] = 0.1974\dots$$

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

## Approximative Konfidenzintervalle

Angenommen, wir wollen den Anteil  $p$  der Wähler einer Partei durch Befragung von  $n$  Wählern schätzen. Seien  $X_1, \dots, X_n$  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  $\bar{X}_n := \frac{S_n}{n}$ . Wie viele Stichproben braucht man, damit der tatsächliche Stimmenanteil mit 95% Wahrscheinlichkeit um höchstens  $\varepsilon = 1\%$  von Schätzwert abweicht?

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

$$P_p[p \notin [\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon]] \leq \alpha$$

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:

$$\begin{aligned} p \in [\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon] &\iff |\bar{X}_n - p| \leq \varepsilon \iff \bar{X}_n \in [p - \varepsilon, p + \varepsilon] \\ &\iff S_n \in [n(p - \varepsilon), n(p + \varepsilon)] \end{aligned}$$

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  $\text{Bin}(n, p)$  liegt.

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

#### Abschätzung mithilfe der Čebyšev-Ungleichung:

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

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:

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

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:

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

falls

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

Im Beispiel gilt

$$\Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \approx 1.96$$

und die Bedingung ist für  $n \geq 9604$  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  $p \rightarrow 0$  bzw.  $p \rightarrow 1$  degeneriert, ist die Niveaubedingung im Allgemeinen selbst für  $n \rightarrow \infty$  nicht erfüllt. Beispielsweise beträgt das Niveau von approximativen 99% Konfidenzintervallen asymptotisch tatsächlich nur 96.8%!

# Kapitel 5

## Unabhängigkeit und Produktmodelle

### 5.1 Unabhängigkeit in allgemeinen Modellen

#### Unabhängigkeit von Ereignissen

In Abschnitt 2.3 haben wir einen Unabhängigkeitsbegriff für Ereignisse eingeführt: Eine Kollektion  $A_i, i \in I$ , von Ereignissen aus derselben  $\sigma$ -Algebra  $\mathcal{A}$  heißt **unabhängig** bzgl. einer Wahrscheinlichkeitsverteilung  $P$ , falls

$$P[A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_n}] = \prod_{k=1}^n P[A_{i_k}] \quad (5.1.1)$$

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

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

für alle  $n \in \mathbb{N}, i_1, \dots, i_n \in I$  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. Allgemeiner:



**Satz 5.1.** Seien  $\mathcal{A}_i$  ( $i \in I$ ) unabhängige Mengensysteme. Jedes  $\mathcal{A}_i$  sei durchschnittsstabil. Dann gilt:

(1). Die  $\sigma$ -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  $\sigma$ -Algebren  $\sigma(\bigcup_{i \in I_k} \mathcal{A}_i)$  ( $k \in K$ ) unabhängige Mengensysteme.

**Beispiel.** Sind  $A_1, \dots, A_n$  unabhängige Ereignisse, dann sind die Mengensysteme

$$\mathcal{A}_1 = \{A_1\}, \quad \dots, \quad \mathcal{A}_n = \{A_n\}$$

unabhängig und durchschnittsstabil, also sind auch die  $\sigma$ -Algebren

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

unabhängige Mengensysteme, d.h es gilt

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

Dies kann man auch direkt beweisen, siehe Lemma 2.5 oben.

Ein Beispiel zum zweiten Teil der Aussage von Satz 5.1 werden wir im Anschluss an den Beweis des Satzes betrachten.

*Beweis.* (1). Seien  $i_1, \dots, i_n \in I$  ( $n \in \mathbb{N}$ ) paarweise verschieden. Wir müssen zeigen, dass

$$P[B_{i_1} \cap \dots \cap B_{i_n}] = P[B_{i_1}] \cdot \dots \cdot P[B_{i_n}] \quad (5.1.2)$$

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

(a) Die Aussage (5.1.2) gilt nach Voraussetzung für  $B_{i_1} \in \mathcal{A}_{i_1}, \dots, B_{i_n} \in \mathcal{A}_{i_n}$ .

(b) Für  $B_{i_2} \in \mathcal{A}_{i_2}, \dots, B_{i_n} \in \mathcal{A}_{i_n}$  betrachten wir das Mengensystem  $\mathcal{D}$  aller  $B_{i_1} \in \mathcal{A}$ , für die (5.1.2) gilt.  $\mathcal{D}$  ist ein Dynkinsystem, das  $\mathcal{A}_{i_1}$  nach (a) enthält. Da  $\mathcal{A}_{i_1}$  durchschnittsstabil ist, folgt

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

Also gilt (5.1.2) für alle  $B_{i_1} \in \sigma(\mathcal{A}_{i_1})$ .

(c) Für  $B_{i_1} \in \sigma(\mathcal{A}_{i_1})$  und  $B_{i_3} \in \sigma(\mathcal{A}_{i_3}), \dots, B_{i_n} \in \sigma(\mathcal{A}_{i_n})$  betrachten wir nun das Mengensystem aller  $B_{i_2} \in \mathcal{A}$ , für die (5.1.2) gilt. Wiederum ist  $\mathcal{D}$  ein Dynkinsystem, das  $\mathcal{A}_{i_2}$  nach (b) enthält. Wie im letzten Schritt folgt daher

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

d.h. (5.1.2) ist für alle  $B_{i_2} \in \sigma(\mathcal{A}_{i_2})$  erfüllt.

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

$$\mathcal{C}_k := \{B_{i_1} \cap \dots \cap B_{i_n} \mid n \in \mathbb{N}, i_1, \dots, i_n \in I_k \text{ paarw. verschieden}, B_{i_j} \in \mathcal{A}_{i_j}\}.$$

Die Mengensysteme  $\mathcal{C}_k$ ,  $k \in K$ , sind durchschnittsstabil und unabhängig, da jede Kollektion von Ereignissen  $B_i \in \mathcal{A}_i$ ,  $i \in I$ , nach Voraussetzung unabhängig ist. Also sind nach Teil (1) der Aussage auch die  $\sigma$ -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  $(a_1, \dots, a_n) \in \{0, 1\}^n$ ,  $n \in \mathbb{N}$ , gilt:

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

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

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

Zum Beweis bemerken wir, dass die Ereignisse

$$E_m = \{X_{mn+1} = a_1, X_{mn+2} = a_2, \dots, X_{mn+n} = a_n\}, \quad m \in \mathbb{N},$$

„Text steht im  $m$ -ten Block“

unabhängig sind. Nach Satz 5.1 sind nämlich die  $\sigma$ -Algebren

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

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:

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

□

## Unabhängigkeit von Zufallsvariablen

Wir betrachten nun Zufallsvariablen mit Werten in einem messbaren Raum  $(S, \mathcal{S})$ .

**Definition.** Seien  $X, X_i : \Omega \rightarrow S, i \in I$ , Abbildungen.

(1). Das Mengensystem

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

heißt **die von  $X$  erzeugte  $\sigma$ -Algebra auf  $\Omega$** .

(2). Allgemeiner heißt

$$\sigma(X_i | i \in I) := \sigma\left(\bigcup_{i \in I} \sigma(X_i)\right) = \sigma(\{X_i^{-1}(B) | B \in \mathcal{S}, i \in I\})$$

die von den Abbildungen  $X_i, i \in I$ , erzeugte  $\sigma$ -Algebra.

**Bemerkung.** (1). Man prüft leicht nach, dass  $\sigma(X)$  tatsächlich eine  $\sigma$ -Algebra ist.

(2). Eine Abbildung  $X : \Omega \rightarrow S$  ist messbar bzgl.  $\mathcal{A}/\mathcal{S}$  genau dann, wenn  $\sigma(X) \subseteq \mathcal{A}$  gilt. Somit ist  $\sigma(X)$  die **kleinste**  $\sigma$ -Algebra auf  $\Omega$ , bzgl. der  $X$  messbar ist.

(3). Entsprechend ist  $\sigma(X_i, i \in I)$  die kleinste  $\sigma$ -Algebra auf  $\Omega$ , bzgl. der alle Abbildungen  $X_i$  messbar sind.

**Beispiel (Produkt- $\sigma$ -Algebra).** Sei  $\Omega = \{0, 1\}^{\mathbb{N}} = \{\omega = (x_1, x_2, \dots) | x_i \in \{0, 1\}\}$ , oder ein allgemeiner Produktraum, und sei  $X_i(\omega) = x_i$  die Projektion auf die  $i$ -te Komponente. Dann ist die Produkt- $\sigma$ -Algebra  $\mathcal{A}$  auf  $\Omega$  gerade die von den Abbildungen  $X_i$  erzeugte  $\sigma$ -Algebra:

$$\begin{aligned} \mathcal{A} &= \sigma(\{X_1 = a_1, \dots, X_n = a_n\} | n \in \mathbb{N}, a_1, \dots, a_n \in \{0, 1\}) \\ &= \sigma(\{X_i = 1\} | i \in \mathbb{N}) \\ &= \sigma(X_1, X_2, \dots). \end{aligned}$$

Messbare Abbildungen auf  $(\Omega, \mathcal{A})$  sind z.B.

$$S_n(\omega) = X_1(\omega) + \dots + X_n(\omega),$$

$$\bar{L}(\omega) = \limsup_{n \rightarrow \infty} \frac{1}{n} S_n(\omega), \quad \underline{L}(\omega) = \liminf_{n \rightarrow \infty} \frac{1}{n} S_n(\omega), \quad \text{etc.}$$

Wir können nun einen Unabhängigkeitsbegriff für allgemeine Zufallsvariablen einführen, der kompatibel mit dem oben definierten Unabhängigkeitsbegriff für Mengensysteme ist.

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

(1). Eine endliche Kollektion  $X_1, \dots, X_n : \Omega \rightarrow S$  von Zufallsvariablen auf  $(\Omega, \mathcal{A}, P)$  heißt **unabhängig**, falls

$$P[X_1 \in B_1, \dots, X_n \in B_n] = \prod_{i=1}^n P[X_i \in B_i] \quad \forall B_i \in \mathcal{S} \quad (1 \leq i \leq n). \quad (5.1.3)$$

(2). Eine beliebige Kollektion  $X_i, i \in I$ , von Zufallsvariablen auf  $(\Omega, \mathcal{A}, P)$  heißt **unabhängig**, falls jede endliche Teilkollektion  $X_{i_1}, \dots, X_{i_n}$  ( $i_1, \dots, i_n \in I$  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 (5.1.3).

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

$$\sigma(X_i) = \{\{X_i \in B\} \mid B \in \mathcal{B}(\mathcal{S})\}, \quad i \in I,$$

unabhängige Mengensysteme sind.

Sei  $(\tilde{\Omega}, \tilde{\mathcal{S}})$  ein weiterer messbarer Raum. Eine sehr wichtige Konsequenz von Bemerkung (2) ist:

**Satz 5.2 (Funktionen von unabhängigen Zufallsvariablen sind unabhängig).** Sind  $X_i : \Omega \rightarrow S, i \in I$ , unabhängige Zufallsvariablen auf  $(\Omega, \mathcal{A}, P)$ , und sind  $h_i : S \rightarrow \tilde{S}$  messbare Abbildungen, dann sind auch die Zufallsvariablen  $Y_i := h_i(X_i), i \in I$ , unabhängig bzgl.  $P$ .

*Beweis.*

$$\sigma(Y_i) = \left\{ \underbrace{Y_i^{-1}(B)}_{X_i^{-1}(h_i^{-1}(B))} \mid B \in \tilde{\mathcal{S}} \right\}.$$

Da die  $\sigma$ -Algebren  $\sigma(X_i), i \in I$ , unabhängig sind, sind auch  $\sigma(Y_i), i \in I$ , unabhängige Mengensysteme.  $\square$

Aufgrund von Satz 5.1 kann man allgemeiner eine Kollektion  $X_i, i \in I$ , von unabhängigen Zufallsvariablen in disjunkte Gruppen  $X_i, i \in I_k, I = \dot{\bigcup}_k I_k$ , einteilen, und messbare Funktionen

$$Y_k = h_k(X_i | i \in I_k), \quad k \in K$$

von den Zufallsvariablen der verschiedenen Gruppen betrachten. Auch die  $Y_k$  sind dann wieder unabhängige Zufallsvariablen.

Für unabhängige reellwertige Zufallsvariablen  $X_i (i \in I)$  gilt insbesondere

$$P[X_{i_1} \leq c_1, \dots, X_{i_n} \leq c_n] = \prod_{k=1}^n P[X_{i_k} \leq c_k] \quad (5.1.4)$$

für alle  $n \in \mathbb{N}, i_1, \dots, i_n \in I$  paarweise verschieden, und  $c_i \in \mathbb{R}$ .

Tatsächlich werden wir im nächsten Abschnitt zeigen, dass Bedingung (5.1.4) äquivalent zur Unabhängigkeit der  $X_i$  ist. Als erste Anwendung betrachten wir Extrema von unabhängigen exponentialverteilten Zufallsvariablen.

**Beispiel (Maxima von exponentialverteilten Zufallsvariablen).** Seien  $T_1, T_2, \dots$  unabhängige  $\text{Exp}(1)$ -verteilte Zufallsvariablen. Wir wollen uns überlegen, wie sich die Extremwerte (Rekorde)

$$M_n = \max\{T_1, \dots, T_n\}$$

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

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

$$\limsup_{n \rightarrow \infty} \frac{T_n}{\log n} = 1 \quad P\text{-fast sicher.} \quad (5.1.5)$$

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

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

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

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

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

$$P \left[ \limsup_{n \rightarrow \infty} \frac{T_n}{\log n} > 1 \right] = \lim_{c \searrow 1} P \left[ \limsup_{n \rightarrow \infty} \frac{T_n}{\log n} > c \right] = 0. \quad (5.1.6)$$

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

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

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

$$P \left[ \limsup_{n \rightarrow \infty} \frac{T_n}{\log n} \geq 1 \right] = \lim_{c \nearrow 1} P \left[ \limsup_{n \rightarrow \infty} \frac{T_n}{\log n} \geq c \right] = 1 \quad (5.1.7)$$

Aus (5.1.6) und (5.1.7) folgt die Behauptung (5.1.5).

(2). Als nächstes folgern wir:

$$M_n \sim \log n, \quad \text{d.h.} \quad \lim_{n \rightarrow \infty} \frac{M_n}{\log n} = 1 \quad P\text{-f.s.} \quad (5.1.8)$$

Zum Beweis zeigen wir:

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

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

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

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

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

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

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

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

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

$$\sum_{n \in \mathbb{N}} P \left[ \frac{M_n}{\log n} \leq c \right] < \infty \quad (5.1.9)$$

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

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

und diese Folge ist für  $c < 1$  summierbar. Also gilt (5.1.9) 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:

$$P[M_n - \log n \leq c] \xrightarrow{n \rightarrow \infty} e^{-e^{-c}} \quad \text{für alle } c \in \mathbb{R}. \quad (5.1.10)$$

*Beweis.* Wegen

$$\begin{aligned} P[M_n \leq c] &= P[T_i \leq c \quad \forall i = 1, \dots, n] \\ &\stackrel{\text{i.i.d.}}{=} P[T_1 \leq c]^n \\ &= (1 - e^{-c})^n \quad \text{für alle } c \in \mathbb{R} \end{aligned}$$

folgt

$$P[M_n - \log n \leq c] = P[M_n \leq c + \log n] = \left(1 - \frac{1}{n} \cdot e^{-c}\right)^n \xrightarrow{n \rightarrow \infty} e^{-e^{-c}}$$

□

Aussage (5.1.10) 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

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

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

## Konfidenzintervalle für Quantile

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

**Annahme:**  $X_1, \dots, X_n$  unabhängig unter  $P_\mu$  mit stetiger Verteilung  $\mu$ .

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

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

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

**Satz 5.3 (Verteilung der Ordnungsstatistiken).** *Ist  $\mu$  eine absolutstetige Wahrscheinlichkeitsverteilung mit Verteilungsfunktion  $F$ , dann hat  $X_{(k)}$  die Verteilungsfunktion*

$$\begin{aligned} F_{(k)}(c) &= \text{Bin}(n, F(c))[\{k, k+1, \dots, n\}] \\ &= \sum_{j=k}^n \binom{n}{j} F(c)^j \cdot (1 - F(c))^{n-j}. \end{aligned} \quad (5.1.11)$$

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

$$\begin{aligned} F_{(k)}(c) = P_\mu[X_{(k)} \leq c] &= P_\mu[X_i \leq c \text{ für mindestens } k \text{ verschiedene } i \in \{1, \dots, n\}] \\ &= \text{Bin}(n, F(c))[\{k, k+1, \dots, n\}] \\ &= \sum_{j=k}^n \binom{n}{j} F(c)^j \cdot (1 - F(c))^{n-j}. \end{aligned}$$

□

Nach Satz 5.3 ist die Wahrscheinlichkeit, dass der Wert von  $X_{(k)}$  unterhalb eines  $u$ -Quantils der zugrundeliegenden Verteilung  $\mu$  liegt, für alle stetigen Verteilungen gleich! Damit folgt unmittelbar:

**Korollar 5.4 (Ordnungsintervalle).** *Sei  $u \in (0, 1)$  und  $0 \leq k < l \leq n$ . Dann ist das zufällige Intervall  $(X_{(k)}, X_{(l)})$  ein **Konfidenzintervall für das  $u$ -Quantil** der zugrundeliegenden Verteilung  $\mu$  zum **Konfidenzniveau***

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



d.h. für jede absolutstetige Wahrscheinlichkeitsverteilung  $\mu$  auf  $\mathbb{R}$ , und für jedes  $u$ -Quantil  $q_u(\mu)$  gilt:

$$P_\mu[X_{(k)} < q_u(\mu) < X_{(l)}] \geq \beta.$$

*Beweis.* Da die Verteilungen stetig sind, gilt  $F_\mu(q_u(\mu)) = u$  für jedes  $u$ -Quantil, und damit nach Satz 5.3:

$$\begin{aligned} P_\mu[X_{(k)} < q_u(\mu) < X_{(l)}] &= \text{Bin}(n, u)[\{k, k+1, \dots, n\}] - \text{Bin}(n, u)[\{l, l+1, \dots, n\}] \\ &= \text{Bin}(n, u)[\{k, k+1, \dots, l-1\}]. \end{aligned}$$

□

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

## 5.2 Gemeinsame Verteilungen und endliche Produktmodelle

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

### Wahrscheinlichkeitsverteilungen auf endlichen Produkträumen

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

$$\mathcal{S}_1 \otimes \dots \otimes \mathcal{S}_n = \sigma(\{B_1 \times \dots \times B_n \mid B_i \in \mathcal{S}_i \quad \forall 1 \leq i \leq n\}).$$

Bezeichnen wir mit  $\pi_i : S_1 \times \dots \times S_n \rightarrow S_i$ ,  $\pi_i(x_1, \dots, x_n) := x_i$ , die kanonische Projektion auf die  $i$ -te Komponente, so gilt

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

**Beispiel.** Für die Borelsche  $\sigma$ -Algebra auf  $\mathbb{R}^n$  gilt:

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

denn  $\mathcal{B}(\mathbb{R}^n)$  wird zum Beispiel von den offenen Quadern, also Produkten von offenen Intervallen, erzeugt. Ein anderes Erzeugendensystem von  $\mathcal{B}(\mathbb{R}^n)$  bilden die Produktmengen

$$(-\infty, c_1] \times (-\infty, c_2] \times \dots \times (-\infty, c_n], \quad c_1, \dots, c_n \in \mathbb{R}. \quad (5.2.1)$$

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

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

auf  $S_i$  (**eindimensionale Randverteilungen (marginals)**) von  $\mu$ . Wir werden in Kapitel 9 allgemeine Wahrscheinlichkeitsverteilungen auf Produkträumen konstruieren und systematisch untersuchen. Im Moment beschränken wir uns meist auf eine spezielle Klasse von solchen Verteilungen: die endlichen Produktmodelle.

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

$$\mu[B_1 \times \dots \times B_n] = \prod_{i=1}^n \mu_i[B_i] \quad \forall B_i \in \mathcal{S}_i, 1 \leq i \leq n, \quad (5.2.2)$$

gilt.

**Bemerkung.** Das Produktmaß  $\mu$  ist durch (5.2.2) *eindeutig* festgelegt, denn die Produktmengen bilden einen durchschnittsstabilen Erzeuger der  $\sigma$ -Algebra  $\mathcal{S}_1 \otimes \dots \otimes \mathcal{S}_n$ . Die Existenz von Produktmaßen folgt aus dem Satz von Fubini, den wir in Abschnitt 9.1 beweisen. Für Wahrscheinlichkeitsverteilungen auf  $\mathbb{R}$  zeigen wir die Existenz von Produktmaßen im nächsten Abschnitt.

Das nach der Bemerkung eindeutige Produktmaß der Wahrscheinlichkeitsverteilungen  $\mu_1, \dots, \mu_n$  bezeichnen wir mit  $\mu_1 \otimes \dots \otimes \mu_n$ . Die eindimensionalen Randverteilungen eines Produktmaßes sind gerade die Faktoren  $\mu_i$ .

**Lemma 5.5.** *Unter  $\mu = \mu_1 \otimes \dots \otimes \mu_n$  sind die Projektionen*

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

*unabhängig mit Verteilung  $\mu_i$ .*

*Beweis.* Für  $B_i \in \mathcal{S}_i, 1 \leq i \leq n$ , gilt:

$$\begin{aligned}\mu[\pi_i \in B_i] &= \mu[S_1 \times \dots \times S_{i-1} \times B_i \times S_{i+1} \times \dots \times S_n] \\ &= \mu_i[B_i] \cdot \prod_{\substack{j \neq i \\ =1}} \underbrace{\mu_j[S_j]}_{=1} = \mu_i[B_i],\end{aligned}$$

und

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

□

Sind die Mengen  $S_1, \dots, S_n$  abzählbar, dann gilt  $\mu = \mu_1 \otimes \dots \otimes \mu_n$  genau dann, wenn die Massenfunktion von  $\mu$  das Produkt der einzelnen Massenfunktionen ist, d.h.

$$\mu(x_1, \dots, x_n) = \prod_{i=1}^n \mu_i(x_i) \quad \text{für alle } x_i \in S_i, 1 \leq i \leq n.$$

Im Fall  $S_1 = \dots = S_n = \mathbb{R}$  mit Borelscher  $\sigma$ -Algebra bilden die Mengen aus (5.2.1) einen durchschnittsstabilen Erzeuger der Produkt- $\sigma$ -Algebra  $\mathcal{B}(\mathbb{R}^n)$ . Also ist  $\mu = \mu_1 \otimes \dots \otimes \mu_n$  genau dann, wenn

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

gilt. Die linke Seite ist die Verteilungsfunktion  $F_\mu(c_1, \dots, c_n)$  der multivariaten Verteilung  $\mu$ , die rechte Seite das Produkt der Verteilungsfunktionen der  $\mu_i$ .

**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  $\mu = \mu_1 \otimes \dots \otimes \mu_n$  die Gleichverteilung auf dem Quader  $S = \prod_{i=1}^n (a_i, b_i)$ , denn für  $c_1, \dots, c_n \in S$  gilt:

$$\begin{aligned}\mu \left[ \prod_{i=1}^n (-\infty, c_i] \right] &= \prod_{i=1}^n \mu_i[(-\infty, c_i]] \\ &= \prod_{i=1}^n \frac{c_i - a_i}{b_i - a_i} \\ &= \lambda^n \left[ \prod_{i=1}^n (a_i, c_i] \right] / \lambda^n[S].\end{aligned}$$

## Absolutstetigkeit von multivariaten Verteilungen

### Absolutstetigkeit von endlichen Produktmodellen

Der Satz von Fubini, den wir in Abschnitt 9.1 in größerer Allgemeinheit beweisen werden, besagt unter anderem, dass das  $n$ -dimensionale Lebesgueintegral einer beliebigen Borel-messbaren nicht-negativen Funktion  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  existiert, und als Hintereinanderausführung von eindimensionalen Integralen nach den Koordinaten  $x_1, \dots, x_n$  berechnet werden kann:

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

Hierbei können die eindimensionalen Integrationen in beliebiger Reihenfolge ausgeführt werden.

Für den Beweis verweisen wir auf die Analysisvorlesung bzw. auf Abschnitt 9.1 unten.

In Analogie zum eindimensionalen Fall heißt eine Wahrscheinlichkeitsverteilung  $\mu$  auf  $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$  *stetig* oder *absolutstetig*, falls eine  $\mathcal{B}(\mathbb{R}^n)$ -messbare *Dichtefunktion*  $f : \mathbb{R}^n \rightarrow [0, \infty)$  existiert mit

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

für jeden Quader, bzw. allgemeiner für jede Borelmenge  $B \subseteq \mathbb{R}^n$ . Endliche Produkte von eindimensionalen absolutstetigen Verteilungen sind wieder absolutstetig, und die Dichte ist das Produkt der einzelnen Dichten:

**Lemma 5.6.** *Sind  $\mu_1, \dots, \mu_n$  absolutstetige Wahrscheinlichkeitsverteilungen auf  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  mit Dichtefunktionen  $f_1, \dots, f_n$ , dann ist das Produkt  $\mu = \mu_1 \otimes \dots \otimes \mu_n$  eine absolutstetige Wahrscheinlichkeitsverteilung auf  $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$  mit Dichtefunktion*

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

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

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

□

Die Dichtefunktion der Gleichverteilung auf dem Quader  $S = (a_1, b_1) \times \dots \times (a_n, b_n)$  ist beispielsweise

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

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

**Beispiel (Standardnormalverteilung im  $\mathbb{R}^n$ ).** Die Wahrscheinlichkeitsverteilung

$$\mu = \bigotimes_{i=1}^n N(0, 1)$$

auf  $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$  heißt  **$n$ -dimensionale Standardnormalverteilung**. Die mehrdimensionale Standardnormalverteilung ist absolutstetig mit Dichte

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

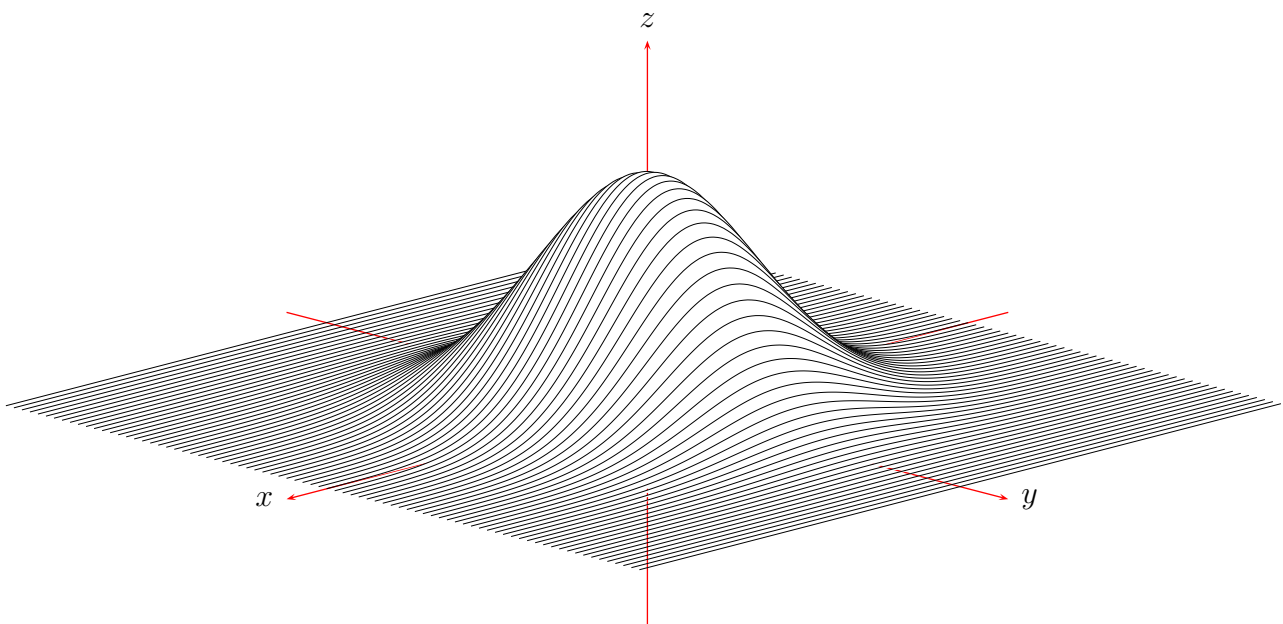


Abbildung 5.1: Dichte der Standardnormalverteilung in  $\mathbb{R}^2$ .

## Gemeinsame Verteilungen

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

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

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

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

Wie zuvor im diskreten Fall (s. Abschnitt 2.4) definieren wir:

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

Der folgende Satz gilt analog zum diskreten Fall:

**Satz 5.7.** Die folgenden Aussagen sind äquivalent:

- (1). Die Zufallsvariablen  $X_1, \dots, X_n$  sind unabhängig.
- (2). Die gemeinsame Verteilung  $\mu_{X_1, \dots, X_n}$  ist ein Produktmaß.
- (3).  $\mu_{X_1, \dots, X_n} = \mu_{X_1} \otimes \dots \otimes \mu_{X_n}$ .

*Beweis.* „(1)  $\implies$  (3)“: folgt direkt aus der Definition der Unabhängigkeit und der gemeinsamen Verteilung: Sind  $X_1, \dots, X_n$  unabhängig, dann gilt

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

für alle  $B_i \in \mathcal{S}_i, 1 \leq i \leq n$ .

„(3)  $\implies$  (2)“: Die Implikation ist offensichtlich, und „(2)  $\implies$  (1)“ folgt aus Lemma 5.5: Ist  $\mu_{X_1, \dots, X_n}$  ein Produktmaß, dann sind die kanonischen Projektionen  $\pi_1, \dots, \pi_n$  unabhängig unter  $\mu_{X_1, \dots, X_n}$ . Also gilt für  $B_i \in \mathcal{S}_i$ :

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

□

Wir wenden die Aussage von Satz 5.7 nun speziell auf diskrete und reellwertige Zufallsvariablen an:

### Diskrete Zufallsvariablen

Sind die Wertebereiche  $S_1, \dots, S_n$  der Zufallsvariablen  $X_1, \dots, X_n$  abzählbar, dann wird die gemeinsame Verteilung vollständig durch die gemeinsame Massenfunktion

$$p_{X_1, \dots, X_n}(a_1, \dots, a_n) = P[X_1 = a_1, \dots, X_n = a_n], \quad (a_1, \dots, a_n) \in S_1 \times \dots \times S_n$$

beschrieben. Die Zufallsvariablen  $X_1, \dots, X_n$  sind genau dann unabhängig, wenn die gemeinsame Massenfunktion das Produkt der einzelnen Massenfunktionen ist, s. Satz 2.7. Als Konsequenz aus Satz 5.7 ergibt sich zudem:

**Korollar 5.8.** Sind  $X_i : \Omega \rightarrow S_i$ ,  $1 \leq i \leq n$ , diskrete Zufallsvariablen, und hat die gemeinsame Massenfunktion eine Darstellung

$$p_{X_1, \dots, X_n}(a_1, \dots, a_n) = c \cdot \prod_{i=1}^n g_i(a_i) \quad \forall (a_1, \dots, a_n) \in S_1 \times \dots \times S_n$$

in Produktform mit einer Konstanten  $c \in \mathbb{R}$ , und Funktionen  $g_i : S_i \rightarrow [0, \infty)$ , dann sind  $X_1, \dots, X_n$  unabhängig mit Massenfunktion

$$p_{X_i}(a_i) = \frac{g_i(a_i)}{\sum_{a \in S_i} g_i(a)}$$

*Beweis.* Die Werte

$$\tilde{g}_i(a_i) = \frac{g_i(a_i)}{\sum_{a \in S_i} g_i(a)}, \quad a_i \in S_i,$$

sind die Gewichte eine Wahrscheinlichkeitsverteilung  $\mu_i$  auf  $S_i$ . Nach Voraussetzung gilt

$$\begin{aligned} \mu_{X_1, \dots, X_n}[\{a_1\} \times \dots \times \{a_n\}] &= p_{X_1, \dots, X_n}(a_1, \dots, a_n) \\ &= \tilde{c} \cdot \prod_{i=1}^n \tilde{\mu}_{X_i}[\{a_i\}] \quad \forall (a_1, \dots, a_n) \in S_1 \times \dots \times S_n \end{aligned} \quad (5.2.3)$$

mit einer reellen Konstante  $\tilde{c}$ . Da auf beiden Seiten von (5.2.3) bis auf den Faktor  $\tilde{c}$  die Massenfunktionen von Wahrscheinlichkeitsverteilungen stehen, gilt  $\tilde{c} = 1$ , und damit

$$\mu_{X_1, \dots, X_n} = \bigotimes_{i=1}^n \mu_i.$$

Also sind die  $X_i$  unabhängig mit Verteilung  $\mu_i$ , d.h. mit Massenfunktion  $\tilde{g}_i$ .  $\square$

**Beispiel (Zwei Würfel).** Seien  $X, Y : \Omega \rightarrow \{1, 2, 3, 4, 5, 6\}$  gleichverteilte Zufallsvariablen. Für die Gewichte der gemeinsamen Verteilung von  $X$  und  $Y$  gibt es dann beispielsweise folgende Möglichkeiten:

(1).  $X, Y$  unabhängig.

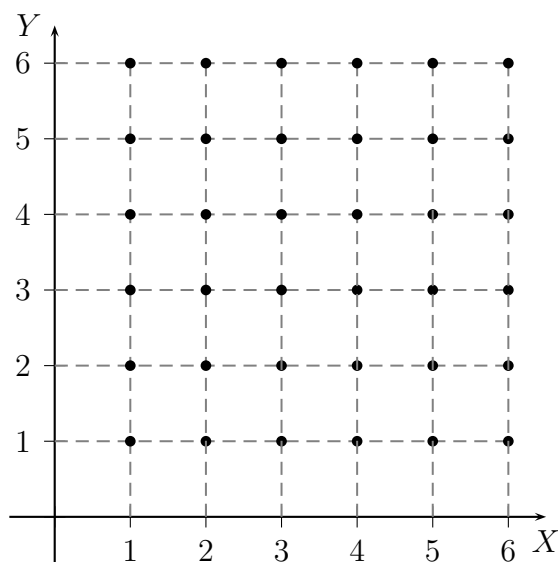


Abbildung 5.2:  $X, Y$  unabhängig;  $\mu_{X,Y} = \mu_X \otimes \mu_Y$ . Gewichte der Punkte sind jeweils  $\frac{1}{36}$ .

(2).  $X, Y$  deterministisch korreliert, z.B.  $Y = (X + 1) \bmod 6$ .

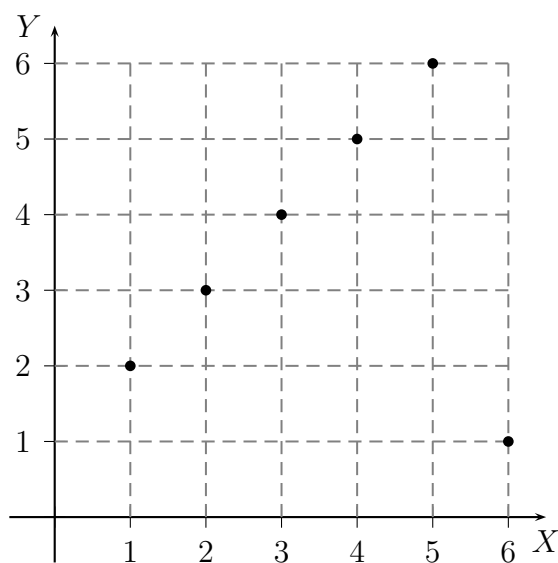


Abbildung 5.3:  $Y = (X + 1) \bmod 6$ . Das Gewicht eines einzelnen Punktes ist  $\frac{1}{6}$ .

(3).  $Y = (X + Z) \bmod 6$ ,  $Z$  unabhängig von  $X$ ,  $Z = 0, \pm 1$  mit Wahrscheinlichkeit  $\frac{1}{3}$ .



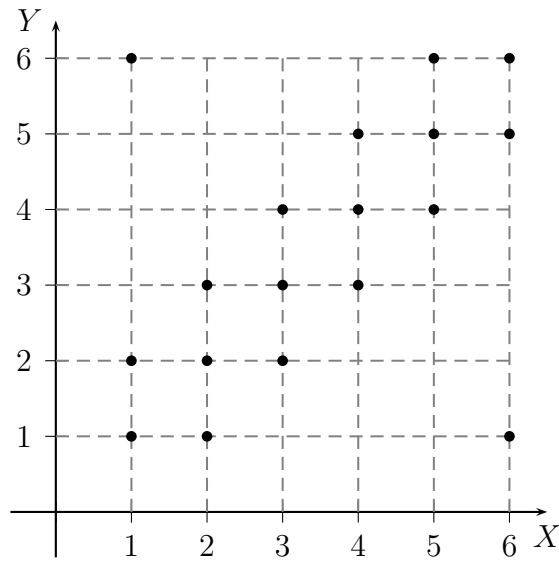


Abbildung 5.4:  $Y = (X + Z) \bmod 6$ ;  $Z \sim \text{unif}\{-1, 0, 1\}$ . Das Gewicht eines einzelnen Punktes ist  $\frac{1}{18}$

### Reelle Zufallsvariablen

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

$$\begin{aligned} F_{X_1, \dots, X_n}(c_1, \dots, c_n) &:= \mu_{X_1, \dots, X_n}[(-\infty, c_1] \times \dots \times (-\infty, c_n)] \\ &= P[X_1 \leq c_1, \dots, X_n \leq c_n], \quad (c_1, \dots, c_n) \in \mathbb{R}^n, \end{aligned}$$

beschrieben. Die Funktion  $F_{X_1, \dots, X_n} : \mathbb{R}^n \rightarrow [0, 1]$  heißt **gemeinsame Verteilungsfunktion**. Insbesondere sind  $X_1, \dots, X_n$  genau dann unabhängig, wenn

$$F_{X_1, \dots, X_n}(c_1, \dots, c_n) = \prod_{i=1}^n F_{X_i}(c_i) \quad \forall (c_1, \dots, c_n) \in \mathbb{R}^n$$

gilt. In Analogie zu Korollar 5.8 erhalten wir zudem:

**Korollar 5.9.** Seien  $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$  reellwertige Zufallsvariablen.

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

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i) \quad \forall x \in \mathbb{R}^n.$$

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

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = c \cdot \prod_{i=1}^n g_i(x_i) \quad \forall x \in \mathbb{R}^n$$

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

$$f_{X_i}(x_i) = \frac{g_i(x_i)}{\int_{\mathbb{R}} g_i(t) dt}.$$

Der Beweis verläuft ähnlich wie der von Korollar 5.8, und wird dem Leser zur Übung überlassen.

**Beispiel (Zufällige Punkte in der Ebene).** Seien  $X$  und  $Y$  unabhängige Zufallsvariablen,  $N(0, \sigma^2)$ -verteilt auf  $(\Omega, \mathcal{A}, P)$  mit  $\sigma > 0$ . Dann ist die gemeinsame Verteilung  $\mu_{X,Y}$  absolutstetig mit Dichte

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma^2} \cdot \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \quad (x, y) \in \mathbb{R}^2.$$

es gilt  $(X, Y) \neq (0, 0)$   $P$ -fast sicher. Wir definieren den Radial- und Polaranteil

$$R : \Omega \rightarrow (0, \infty), \quad \Phi : \Omega \rightarrow [0, 2\pi)$$

durch

$$X = R \cdot \cos \Phi \quad \text{und} \quad Y = R \cdot \sin \Phi,$$

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  $(R, \Phi)$  in beliebiger Weise, sodass sich messbare Funktionen ergeben. Wir berechnen nun die gemeinsame Verteilung von  $R$  und  $\Phi$ :

$$\begin{aligned} P[R \leq r_0, \Phi \leq \phi_0] &= P[(X, Y) \in \text{„Kuchenstück“ mit Winkel } \phi_0 \text{ und Radius } r_0] \\ &= \int \int_{\text{Kuchenstück}} f_{X,Y}(x, y) dx dy \\ &= \int_0^{r_0} \int_0^{\phi_0} f_{X,Y}(r \cos \phi, r \sin \phi) \underbrace{r}_{\substack{\text{Jacobideterminante} \\ \text{der Koordinatentrans. } f}} d\phi dr \\ &= \int_0^{r_0} \int_0^{\phi_0} \frac{r}{2\pi\sigma^2} e^{-r^2/(2\sigma^2)} d\phi dr. \end{aligned}$$

Hierbei haben wir im 3. Schritt den Transformationssatz (Substitutionsregel) für mehrdimensionale Integrale verwendet - der Faktor  $r$  ist die Jacobideterminante der Koordinatentransformation (s. Analysis). 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  $\Phi$  unabhängig. Die Randverteilung  $\mu_\Phi$  ist absolutstetig mit Dichte

$$f_\Phi(\phi) = \text{const.} = \frac{1}{2\pi} \quad (0 \leq \phi < 2\pi),$$

d.h.  $\Phi$  ist gleichverteilt auf  $[0, 2\pi)$ . Somit ist  $\mu_R$  absolutstetig mit Dichte

$$\phi_R(r) = \frac{r}{\sigma^2} \cdot e^{-r^2/(2\sigma^2)} \quad (r > 0).$$

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  $\Phi$  ist unabhängig von  $R$  und gleichverteilt auf  $[0, 2\pi)$ . Wir können Zufallsvariablen mit der entsprechenden gemeinsamen Verteilung erzeugen, indem wir

$$\begin{aligned} \Phi &:= 2\pi U_1, \\ R &:= \sqrt{-2 \log(1 - U_2)} \quad \left( \text{bzw.} = \sqrt{-2 \log U_2} \right), \end{aligned}$$

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

$$X := R \cos \Phi \quad \text{und} \quad Y := R \cdot \sin \Phi$$

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 5.10 (Box-Muller-Verfahren).** **Input:**  $m \in \mathbb{R}, \sigma > 0$

**Output:** unabhängige Stichproben  $\tilde{x}, \tilde{y}$  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.  $\tilde{x} := \sigma x + m, \tilde{y} = \sigma y + m$

**Beispiel (Ordnungsstatistiken).** Für die gesamte Verteilung der Ordnungsstatistiken  $X_{(1)} \leq \dots \leq X_{(n)}$ , unabhängiger, identisch verteilter, stetiger Zufallsvariablen  $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$  gilt aus Symmetriegründen und wegen  $P[X_i = X_j] = 0$  für  $i \neq j$ :

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

Also ist die gemeinsame Verteilung von  $X_{(1)}, \dots, X_{(n)}$  absolutstetig mit Dichte

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

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

$$\begin{aligned} P[X_{(k)} \leq c] &= n! \int_{\mathbb{R}} \dots \int_{\mathbb{R}} I_{\{y_1 < y_2 < \dots < y_n\}} f(y_1) \dots f(y_n) \cdot I_{Y_k \leq c} dy_1 \dots dy_n \\ &= \int_{-\infty}^c f_{(k)}(y_k) dy_k \end{aligned}$$

mit

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

Dasselbe Resultat hätte man auch mit etwas Rechnen aus Satz 5.3 herleiten können.

**Bemerkung (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

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

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

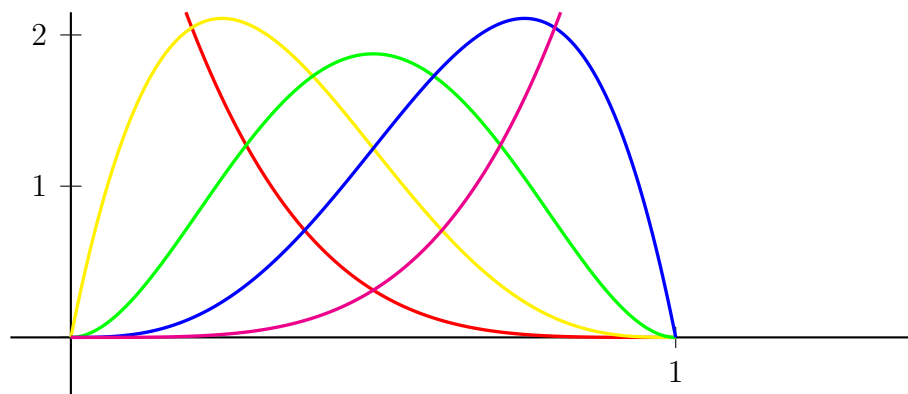


Abbildung 5.5: Abbildung der Dichtefunktionen der zugehörigen Verteilungen von  $X_{(1)}, \dots, X_{(5)}$  bei  $n = 5$  in (rot, gelb, grün, blau, magenta).

## 5.3 Unendliche Produktmodelle

### Konstruktion von unabhängigen Zufallsvariablen

Seien  $\mu_1, \mu_2, \dots$  vorgegebene Wahrscheinlichkeitsverteilungen auf  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ . Wir werden nun explizit unabhängige Zufallsvariablen  $X_k, k \in \mathbb{N}$ , mit Verteilungen  $\mu_k$  konstruieren. Als Konsequenz ergibt sich die Existenz des unendlichen Produktmaßes  $\bigotimes_{k=1}^{\infty} \mu_k$  als gemeinsame Verteilung der Zufallsvariablen  $X_k$ . Die Zufallsvariablen  $X_i$  können wir sogar auf den Raum  $\Omega = (0, 1)$  mit Gleichverteilung realisieren:

**Satz 5.11.** Auf dem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{B}((0, 1)), \mathcal{U}_{(0,1)})$  existieren unabhängige Zufallsvariablen  $X_k : \Omega \rightarrow \mathbb{R}, k \in \mathbb{N}$ , mit Verteilungen

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

*Beweis.* Wir verfahren in drei Schritten:

(1). Wir konstruieren die Zufallsvariablen im Fall

$$\mu_k = \text{Bernoulli} \left( \frac{1}{2} \right) = \mathcal{U}_{(0,1)} \quad \forall k \in \mathbb{N},$$

d.h. im fairen Münzwurfmodell. Dazu verwenden wir die schon in Abschnitt 4.2 eingeführte Transformation  $X : (0, 1) \rightarrow \{0, 1\}^{\mathbb{N}}$ , die einer reellen Zahl die Ziffernfolge ihrer Binärdarstellung zuordnet, d.h. wir setzen

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

siehe Abbildung 4.4. Die Abbildungen  $X_k : (0, 1) \rightarrow \{0, 1\}$  sind messbar, und es gilt

$$P[X_1 = a_1, \dots, X_n = a_n] = 2^{-n} \quad \forall n \in \mathbb{N}, a_1, \dots, a_n \in \{0, 1\}, \quad (5.3.1)$$

da die Menge  $\{\omega \in \Omega : X_1(\omega) = a_1, \dots, X_n(\omega) = a_n\}$  gerade aus den Zahlen in  $(0, 1)$  besteht, deren Binärdarstellung mit den Ziffern  $a_1, \dots, a_n$  beginnt, und damit ein Intervall der Länge  $2^{-n}$  ist. Nach (5.3.1) sind  $X_1, \dots, X_n$  für alle  $X_k, k \in \mathbb{N}$ , unabhängig mit Verteilung  $\mu_k$ .

(2). Wir konstruieren die Zufallsvariablen im Fall

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

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:

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

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

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

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

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

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

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

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

Da die dyadischen Intervalle ein durchschnittsstabiles Erzeugendensystem der Borelschen  $\sigma$ -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 4.19: Sind  $\mu_k, k \in \mathbb{N}$ , beliebige Wahrscheinlichkeitsverteilungen auf  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ , und

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

die linksstetigen verallgemeinerten Inversen der Verteilungsfunktionen

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

dann setzen wir

$$Y_k(\omega) := \underline{G}_k(U_k(\omega)), \quad k \in \mathbb{N}, \omega \in \Omega.$$

Da die Zufallsvariablen  $U_k, k \in \mathbb{N}$ , unabhängig sind, sind nach Satz 5.2 auch die  $Y_k, k \in \mathbb{N}$ , wieder unabhängig. Zudem gilt nach Satz 4.19:

$$P \circ Y_k^{-1} = \mu_k \quad \text{für alle } k \in \mathbb{N}.$$

□

**Bemerkung.** (1). Der Beweis von Satz 5.11 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  $\mu_k$  auf den Borelschen  $\sigma$ -Algebren  $\mathcal{B}(S_k)$  zeigen. Sind beispielsweise  $\phi_k : \mathbb{R} \rightarrow S_k$  Bijektionen, sodass  $\phi_k$  und  $\phi_k^{-1}$  messbar sind, und sind  $\tilde{X}_k : \Omega \rightarrow \mathbb{R}$  unabhängige reellwertige Zufallsvariablen mit Verteilungen  $P[\tilde{X}_k \in B] = \mu_k[\phi_k(B)]$ , dann sind die transformierten Zufallsvariablen

$$X_k = \phi_k(\tilde{X}_k) : \Omega \rightarrow S_k, \quad \forall k \in \mathbb{N},$$

unabhängig mit Verteilungen  $\mu_k$ .

**Beispiel (Random Walks im  $\mathbb{R}^d$ ).** Sei  $\mu$  eine Wahrscheinlichkeitsverteilung auf  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ , und seien  $X_i, i \in \mathbb{N}$ , unabhängige Zufallsvariablen mit identischer Verteilung  $X_i \sim \mu$ . Der durch

$$S_n = a + \sum_{i=1}^n X_i, \quad n = 0, 1, 2, \dots,$$

definierte stochastische Prozess heißt *Random Walk mit Startwert  $a \in \mathbb{R}^d$  und Inkrementverteilung  $\mu$* .

Im Fall  $d = 1$  können wir Stichproben von den Zufallsvariablen  $X_i$ , und damit vom Random Walk, beispielsweise mithilfe der Inversionsmethode, simulieren.

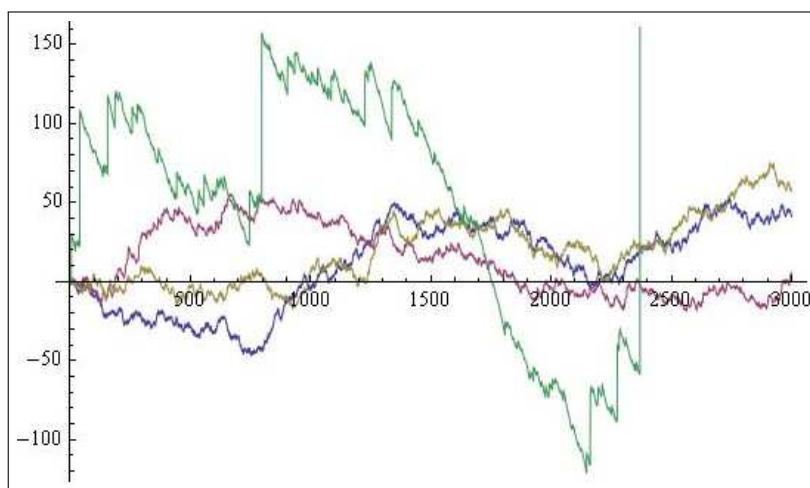


Abbildung 5.6: Grafiken von Trajektorien des Random Walks mit verschiedenen Inkrementverteilungen.

Abbildung 5.6 zeigt Grafiken von Trajektorien des Random Walks mit den Inkrementverteilungen

$$\mu = \frac{1}{2}(\delta_1 + \delta_{-1}) \quad (\text{klassischer Random Walk (SSRW)}),$$

$$\mu = N(0, 1) \quad (\text{diskrete Brownsche Bewegung}),$$



$\mu$  mit Dichte

$$f(x) = e^{-(x+1)} I_{(-1, \infty)}(x) \quad (\text{zentrierte Exp}(1)\text{-Verteilung})$$

und  $\mu$  mit Dichte

$$f(x) = 3 \cdot 2^{-5/2} \cdot \left(x + \frac{3}{2}\right)^{-5/2} \cdot I_{(\frac{1}{2}, \infty)}\left(x + \frac{3}{2}\right) \quad (\text{zentrierte Pareto}(\alpha - 1, \alpha)\text{-Verteilung mit } \alpha = \frac{3}{2}).$$

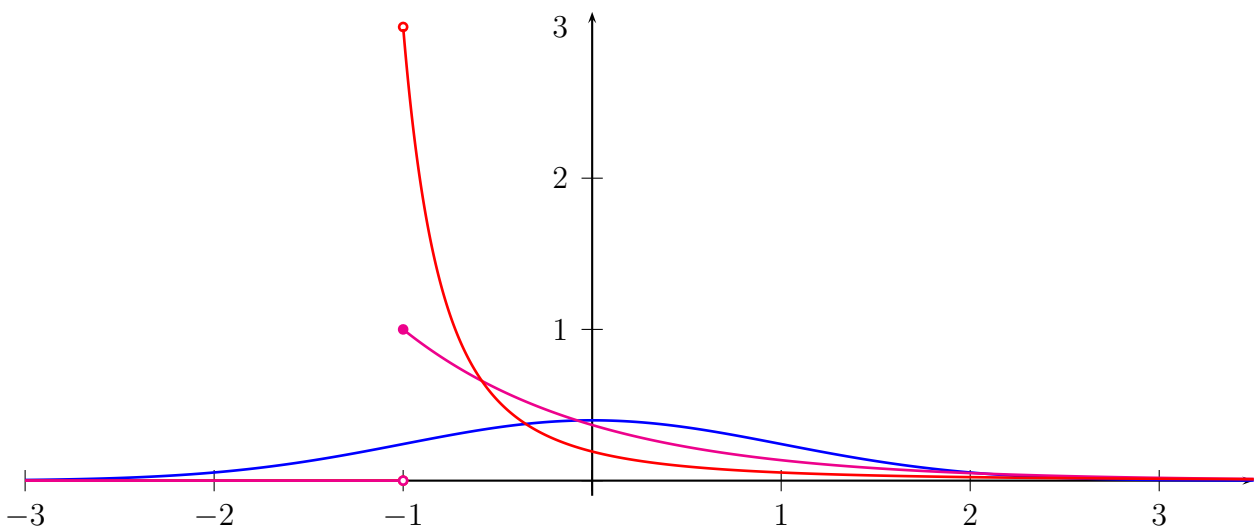


Abbildung 5.7: Dichten der drei stetigen Verteilungen aus Abbildung 5.6:  $f_{N(0,1)}$  in Blau,  $f_{\text{Exp}(1)-1}$  in Magenta und  $f_{\text{Pareto}(\alpha-1, \alpha)}$  in Rot.

Im Gegensatz zu den anderen Verteilungen fällt die Dichte der Pareto-Verteilung für  $x \rightarrow \infty$  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).

## 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

$$\mathbb{R}^{\mathbb{N}} = \{(x_1, x_2, \dots) \mid x_k \in \mathbb{R}, \forall k \in \mathbb{N}\}$$

mit der Produkt- $\sigma$ -Algebra

$$\bigotimes_{k \in \mathbb{N}} \mathcal{B}(\mathbb{R}) = \sigma(\mathcal{C}) = \sigma(\pi_k \mid k \in \mathbb{N}),$$

die von der Kollektion  $\mathcal{C}$  aller Zylindermengen

$$\{\pi_1 \in B_1, \dots, \pi_n \in B_n\} = \{x = (x_k) \in \mathbb{R}^{\mathbb{N}} \mid x_1 \in B_1, \dots, x_n \in B_n\},$$

$n \in \mathbb{N}, B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$ , von den Koordinatenabbildungen  $\pi_k : \mathbb{R}^{\mathbb{N}} \rightarrow \mathbb{R}, \pi_k(x) = x_k$ .

**Korollar 5.12 (Existenz von unendlichen Produktmaßen).** *Zu beliebigen Wahrscheinlichkeitsverteilungen  $\mu_k$  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*

$$\mu[\pi_1 \in B_1, \dots, \pi_n \in B_n] = \mu[B_1] \cdot \dots \cdot \mu_n[B_n] \quad (5.3.2)$$

für alle  $n \in \mathbb{N}$  und  $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$ .

**Definition.** Die Wahrscheinlichkeitsverteilung  $\mu$  mit (5.3.2) 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- $\sigma$ -Algebra bilden.

Zum Beweis der Existenz: betrachten wir die Abbildung  $X : \Omega \rightarrow \mathbb{R}^{\mathbb{N}}$  mit

$$X(\omega) = (X_1(\omega), X_2(\omega), \dots),$$

wobei  $X_k$  unabhängige Zufallsvariablen mit Verteilung  $\mu_k$  sind.  $X$  ist messbar bzgl.  $\bigotimes_{k \in \mathbb{N}} \mathcal{B}(\mathbb{R})$ , denn

$$X^{-1}[\{x \in \mathbb{R}^{\mathbb{N}} | (x_1, \dots, x_n) \in B\}] = \{\omega \in \Omega | (X_1(\omega), \dots, X_n(\omega)) \in B\} \in \mathcal{A}$$

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

$$\begin{aligned} \mu[\pi_1 \in B_1, \dots, \pi_m \in B_m] &= \mu[\{x \in \mathbb{R}^{\mathbb{N}} | x_1 \in B_1, \dots, x_n \in B_n\}] \\ &= P[X_1 \in B_1, \dots, X_n \in B_n] \\ &= \prod_{k=1}^n \mu_k[B_k] \end{aligned}$$

für alle  $n \in \mathbb{N}$  und  $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$ . Also ist  $\mu$  das gesuchte Produktmaß.  $\square$

**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  $\sigma$ -Algebren  $\mathcal{S}_k$ . Das Produktmaß sitzt auf dem Produktraum

$$\left( \prod_{k \in \mathbb{N}} S_k, \bigotimes_{k \in \mathbb{N}} \mathcal{S}_k \right).$$

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  $\pi_k : \prod_{i \in \mathbb{N}} S_i \rightarrow S_k$  unter dem Produktmaß  $\bigotimes_{i \in I} \mu_i$  unabhängig und  $\mu_k$ -verteilt. Man nennt den Produktraum

$$(\Omega, \mathcal{A}, P) = \left( \prod S_i, \bigotimes \mathcal{S}_i, \bigotimes \mu_i \right)$$

daher auch das **kanonische Modell** für unabhängige  $\mu_i$ -verteilte Zufallsvariablen.

## 5.4 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  $A \in \sigma(X_i \mid i \in I)$  heißt **asymptotisches Ereignis (tail event)**, falls

$$A \in \sigma(X_i \mid i \in I \setminus I_0) \quad \text{für jede **endliche** Teilmenge } I_0 \subseteq I \text{ gilt.}$$

Die Menge

$$\tau = \bigcap_{I_0 \subseteq I \text{ endlich}} \sigma(X_i \mid i \in I \setminus I_0)$$

aller asymptotischen Ereignisse ist eine  $\sigma$ -Algebra.  $\tau$  heißt **asymptotische  $\sigma$ -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  $A \in \sigma(X_n, n \in \mathbb{N})$ :

$$A \text{ asymptotisch} \quad \Leftrightarrow \quad A \in \underbrace{\sigma(X_{n+1}, X_{n+2}, \dots)}_{\text{Zukunft ab } n} \quad \text{für alle } n.$$

Beispiele für asymptotische Ereignisse von reellwertigen Zufallsvariablen sind

$$\{X_n > 5n \text{ unendlich oft}\}, \quad \left\{ \limsup_{n \rightarrow \infty} X_n < c \right\}, \quad \left\{ \exists \lim_{n \rightarrow \infty} X_n \right\}, \quad \left\{ \exists \lim_{n \rightarrow \infty} \frac{1}{n} S_n = m \right\},$$

wobei  $S_n = X_1 + \dots + X_n$ . Die Ereignisse

$$\left\{ \sup_{n \in \mathbb{N}} X_n = 3 \right\} \quad \text{und} \quad \left\{ \lim S_n = 5 \right\}$$

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  $X_i : \Omega \rightarrow \{0, 1\}$ , 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.

### Das 0-1-Gesetz von Kolmogorov

**Satz 5.13 ( 0-1-Gesetz von Kolmogorov ).** Sind  $X_i$  ( $i \in I$ ) unabhängige Zufallsvariablen auf  $(\Omega, \mathcal{A}, P)$ , dann gilt

$$P[A] \in \{0, 1\} \text{ für alle } A \in \tau.$$

„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:  $X_1, X_2, \dots$  unabhängige Zufallsvariablen

$\implies \sigma(X_1), \sigma(X_2), \dots, \sigma(X_n), \sigma(X_{n+1}), \sigma(X_{n+2}), \dots$  unabhängige Mengensysteme

$\implies \sigma(X_1, \dots, X_n), \sigma(X_{n+1}, X_{n+2}, \dots)$  sind unabhängig für alle  $n \in \mathbb{N}$

$\implies \sigma(X_1, \dots, X_n)$  und  $\tau$  sind unabhängig für alle  $n \in \mathbb{N}$

$\implies \tau$  unabhängig von  $\sigma(X_1, X_2, \dots) \supseteq \tau$

$\implies$  Ereignisse  $A \in \tau$  sind unabhängig von sich selbst

$\implies P[A] \in \{0, 1\} \quad \forall A \in \tau$ .

Hierbei gilt die zweite Implikation nach Satz 5.1 (2), und die vierte nach Satz 5.1 (1) □

### Anwendungen auf Random Walks und Perkulationsmodelle

**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

$$P[X_i = 1] = p, \quad P[X_i = -1] = 1 - p.$$

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  $n \rightarrow \infty$  dieser Wahrscheinlichkeiten. Aus der **Stirlingschen Formel**

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

folgt

$$P[S_{2n} = a] \sim \frac{\sqrt{4\pi n}}{2\pi n} \cdot \frac{\left(\frac{2n}{e}\right)^{2n}}{\left(\frac{n}{e}\right)^{2n}} \cdot p^n \cdot (1-p)^n = \frac{1}{\sqrt{\pi n}} (4p(1-p))^n \quad \text{für } n \rightarrow \infty.$$

Für  $p \neq \frac{1}{2}$  fallen die Wahrscheinlichkeiten also exponentiell schnell ab. Insbesondere gilt dann

$$\sum_{m=0}^{\infty} P[S_m = a] = \sum_{n=0}^{\infty} P[S_{2n} = a] < \infty,$$

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

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

Für  $p = \frac{1}{2}$  gilt dagegen  $P[S_{2n} = a] \sim 1/\sqrt{\pi n}$ , und damit

$$\sum_{m=0}^{\infty} P[S_m = a] = \sum_{n=0}^{\infty} P[S_{2n} = a] = \infty.$$

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 5.14 (Rekurrenz und unbeschränkte Oszillationen des symmetrischen Random Walks).**

Für  $p = \frac{1}{2}$  gilt

$$P[\overline{\lim} S_n = +\infty \text{ und } \underline{\lim} S_n = -\infty] = 1.$$

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 5.1 („Affe tippt Shakespeare“) gibt es  $P$ -fast sicher unendlich viele Blöcke der Länge  $k$  mit  $X_{n+1} = X_{n+2} = \dots = X_{n+k} = 1$ . Es folgt

$$P[\overline{\lim} S_n - \underline{\lim} S_n = \infty] \geq P\left[\bigcap_k \bigcup_n \{S_{n+k} - S_n = k\}\right] = 1,$$

und damit

$$1 = P[\overline{\lim} S_n = +\infty \text{ oder } \underline{\lim} S_n = -\infty] \leq P[\overline{\lim} S_n = +\infty] + P[\underline{\lim} S_n = -\infty].$$

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

$$P[\underline{\lim} S_n = -\infty] = P[\overline{\lim} S_n = +\infty] = 1.$$

□

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 Perkulationsmodell:

**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

$$P[X_i = 1] = p, \quad P[X_i = 0] = 1 - p.$$

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:

$$P[A] = 1 \iff p > \frac{1}{2},$$

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.

**Definition.** Eine Zufallsvariable  $Y : \Omega \rightarrow [-\infty, \infty]$  heißt **asymptotisch**, wenn die bzgl. der asymptotischen  $\sigma$ -Algebra  $\tau$  messbar ist.

Das Perkulationsmodell 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 Perkulationsmodelle werden in den Wahrscheinlichkeitstheorie-Lehrbüchern von *Y. Sinai* und *A. Klenke* hergeleitet.

**Korollar 5.15.** Sind  $X_i$  ( $i \in I$ ) unabhängige Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ , dann ist jede asymptotische Zufallsvariable  $Y : \Omega \rightarrow [-\infty, \infty]$   $P$ -fast sicher konstant, d.h.

$$\exists c_0 \in [-\infty, \infty] : P[Y = c_0] = 1.$$

*Beweis.* Ist  $Y$   $\tau$ -messbar, dann sind die Ereignisse  $\{Y \leq c\}$ ,  $c \in \mathbb{R}$ , in  $\tau$  enthalten. Aus dem Kolmogorovschen 0-1-Gesetz folgt:

$$F_Y(c) = P[Y \leq c] \in \{0, 1\} \quad \forall c \in \mathbb{R}.$$

Da die Verteilungsfunktion monoton wachsend ist, existiert ein  $c_0 \in [-\infty, \infty]$  mit

$$P[Y \leq c] = \begin{cases} 0 & \text{für } c < c_0 \\ 1 & \text{für } c > c_0 \end{cases},$$

$$\text{und damit } P[Y = c_0] = \lim_{\varepsilon \downarrow 0} (F_Y(c_0) - F_Y(c_0 - \varepsilon)) = 1. = 1. \quad \square$$

Beispiele für asymptotische Zufallsvariablen im Fall  $I = \mathbb{N}$  sind etwa

$$\underline{\lim}_{n \rightarrow \infty} X_n, \quad \overline{\lim}_{n \rightarrow \infty} X_n, \quad \underline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i, \quad \text{sowie} \quad \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i.$$

Insbesondere sind für unabhängige Zufallsvariablen  $X_1, X_2, \dots : \Omega \rightarrow \mathbb{R}$  sowohl

$$\underline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i \quad \text{als auch} \quad \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i \quad P\text{-f.s. konstant.}$$



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.

$$\frac{1}{n} \sum_{i=1}^n X_i \text{ konvergiert } P\text{-f.s., und der Limes ist } P\text{-f.s. konstant}$$

(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!

# Kapitel 6

## Erwartungswert und Varianz

In diesem Kapitel definieren wir den Erwartungswert, die Varianz und die Kovarianz allgemeiner reellwertiger Zufallsvariablen, und beweisen grundlegende Eigenschaften und Abschätzungen. 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  $\sigma$ -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  $X : \Omega \rightarrow \overline{\mathbb{R}}$ , wenn wir die Verteilungsfunktion  $F_X : \mathbb{R} \rightarrow [0, 1]$  definieren durch

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

### 6.1 Erwartungswert

Sei  $(\Omega, \mathcal{A}, P)$  ein Wahrscheinlichkeitsraum und  $X : \Omega \rightarrow \overline{\mathbb{R}}$  eine Zufallsvariable. Wir wollen den Erwartungswert (Mittelwert, Prognosewert) von  $X$  bezüglich der Wahrscheinlichkeitsverteilung  $P$  in sinnvoller Weise definieren. Dazu gehen wir schrittweise vor:

#### Definition des Erwartungswerts

##### Elementare Zufallsvariablen

Nimmt  $X$  nur endlich viele Werte  $c_1, \dots, c_n \in \mathbb{R}$  an, dann soll gelten:

$$E[X] = \sum_{i=1}^n c_i \cdot P[X = c_i],$$

d.h. der Erwartungswert ist das Mittel der Werte  $c_i$  gewichtet mit den Wahrscheinlichkeiten der Ereignisse  $A_i := \{X = c_i\}$ .

**Definition.** Eine Zufallsvariable von der Form

$$X = \sum_{i=1}^n c_i I_{A_i} \quad (n \in \mathbb{N}, c_i \in \mathbb{R}, A_i \in \mathcal{A})$$

heißt *elementar*. Ihr Erwartungswert bzgl.  $P$  ist

$$E[X] := \sum_{i=1}^n c_i \cdot P[A_i].$$

Diese Definition ist ein Spezialfall der Definition des Erwartungswerts diskreter Zufallsvariablen aus Kapitel 1. Insbesondere ist der Erwartungswert  $E[X]$  *wohldefiniert*, d.h. unabhängig von der gewählten Darstellung der Zufallsvariable  $X$  als Linearkombination von Indikatorfunktionen, und die Abbildung  $X \mapsto E[X]$  ist *linear* und *monoton*:

$$E[aX + bY] = a \cdot E[X] + b \cdot E[Y] \quad \text{für alle } a, b \in \mathbb{R},$$

$$X \leq Y \implies E[X] \leq E[Y].$$

Die Definition des Erwartungswerts einer elementaren Zufallsvariable stimmt genau mit der des Lebesgueintegrals der Elementarfunktion  $X$  bzgl. des Maßes  $P$  überein:

$$E[X] = \int X \, dP = \int X(\omega) P(d\omega)$$

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 Lebesgueintegrals bzw. des Erwartungswerts einer allgemeinen Zufallsvariable, siehe auch die Analysisvorlesung.

### Nichtnegative Zufallsvariablen

Die Definition des Erwartungswerts einer nichtnegativen Zufallsvariable beruht auf der monotonen Approximation durch elementare Zufallsvariablen:

**Lemma 6.1.** Sei  $X : \Omega \rightarrow [0, \infty]$  eine nichtnegative Zufallsvariable auf  $(\Omega, \mathcal{A}, P)$ . Dann existiert eine monoton wachsende Folge elementarer Zufallsvariablen  $0 \leq X_1 \leq X_2 \leq \dots$  mit

$$X = \lim_{n \rightarrow \infty} X_n = \sup_{n \in \mathbb{N}} X_n.$$

*Beweis.* Für  $n \in \mathbb{N}$  sei

$$X_n(\omega) := \begin{cases} k \cdot 2^{-n} & \text{falls } k \cdot 2^{-n} \leq X(\omega) < (k+1) \cdot 2^{-n} \text{ für ein } k = 0, 1, \dots, n \cdot 2^{-n} \\ (n+1) \cdot 2^{-n} & \text{falls } X(\omega) \geq (n+1) \cdot 2^{-n} \end{cases}$$

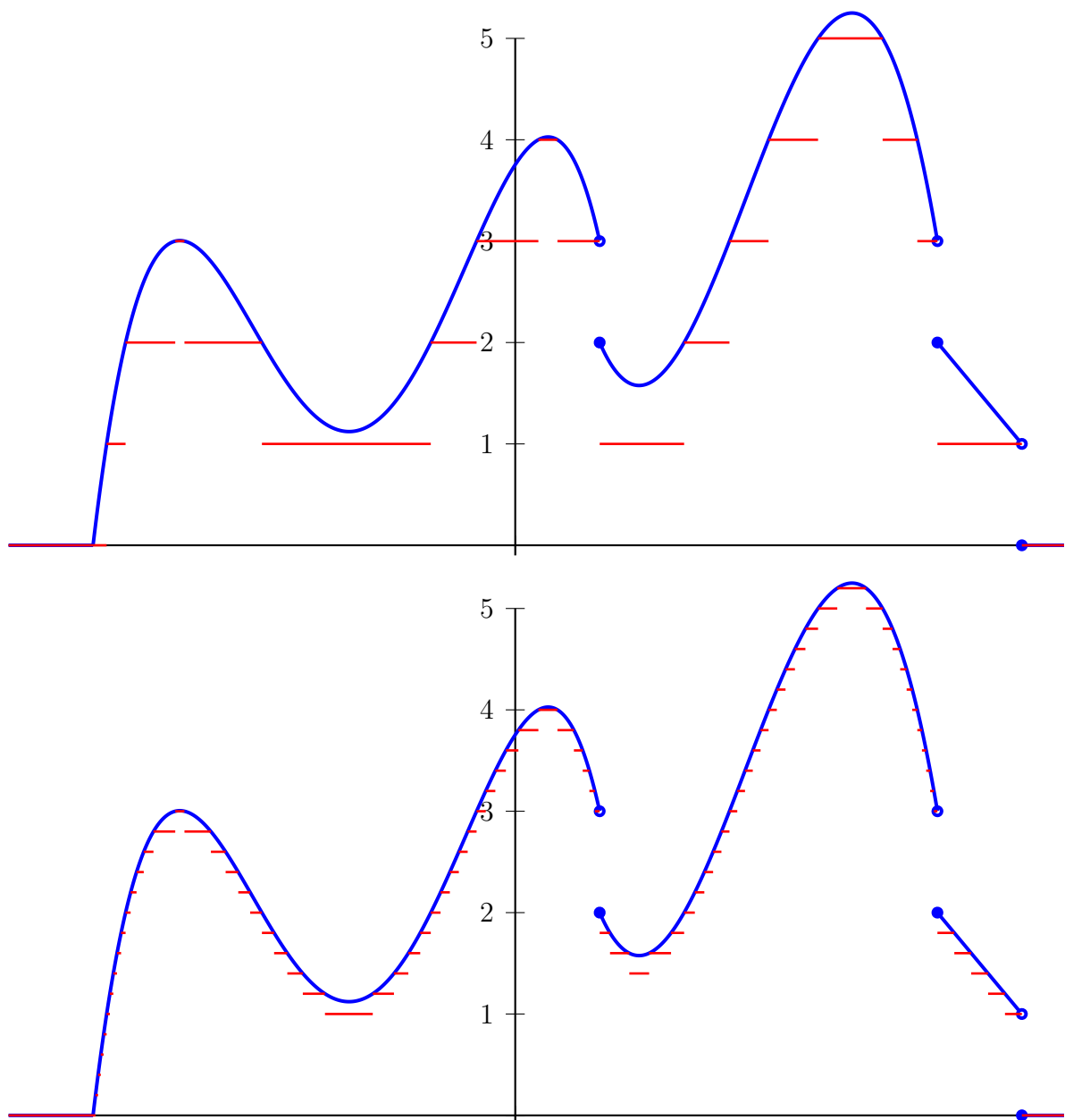


Abbildung 6.1: Approximation durch Elementarfunktionen. Hier ist die Annäherung in rot in zwei verschiedenen Feinheiten dargestellt.

Dann ist  $X_n$  eine elementare Zufallsvariable, denn es gilt

$$X_n = \sum_{k=0}^{n2^n-1} \frac{k}{2^n} I_{\{\frac{k}{2^n} \leq X < \frac{k+1}{2^n}\}} + n I_{\{X \geq n\}}.$$

Die Folge  $X_n(\omega)$  ist für jedes  $\omega$  monoton wachsend, da die Unterteilung immer feiner wird, und

$$\sup_{n \in \mathbb{N}} X_n(\omega) = \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega) \quad \text{für alle } \omega \in \Omega.$$

□

**Definition.** Sei  $X : \Omega \rightarrow [0, \infty]$  eine nicht-negative Zufallsvariable.

Der **Erwartungswert** (bzw. das **Lebesgueintegral**) von  $X$  bzgl.  $P$  ist definiert als

$$E[X] := \lim_{n \rightarrow \infty} E[X_n] = \sup_{n \rightarrow \infty} E[X_n] \in [0, \infty], \quad (6.1.1)$$

wobei  $X_n$  eine 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 6.2.** Die Definition ist unabhängig von der Wahl einer monoton wachsenden Folge  $X_n$  von nichtnegativen 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.** Sind  $X_n = I_{A_n}$  und  $X = I_A$  Indikatorfunktionen, dann folgt (6.1.1) aus der monotonen Stetigkeit von  $P$ . In diesem Fall gilt nämlich:

$$X_n \nearrow X \iff A_n \nearrow A \quad (\text{d.h. } A_n \text{ monoton wachsend und } A = \bigcup A_n).$$

Aus der monotonen Stetigkeit von  $P$  folgt dann

$$E[X] = P[A] = \lim P[A_n] = \lim E[X_n].$$

Aus der Definition des Erwartungswerts folgt unmittelbar:

**Lemma 6.3.** Für nichtnegative Zufallsvariablen  $X, Y$  mit  $X \leq Y$  gilt  $E[X] \leq E[Y]$ .

*Beweis.* Ist  $X \leq Y$ , dann gilt auch  $X_n \leq Y_n$  für die approximierenden elementaren Zufallsvariablen aus Lemma 6.1, also

$$E[X] = \sup_{n \in \mathbb{N}} E[X_n] \leq \sup_{n \in \mathbb{N}} E[Y_n] = E[Y].$$

□

### Allgemeine Zufallsvariablen

Eine allgemeine Zufallsvariable  $X : \Omega \rightarrow \overline{\mathbb{R}}$  können wir in ihren positiven und negativen Anteil zerlegen:

$$X = X^+ - X^- \quad \text{mit} \quad X^+ := \max(X, 0), \quad X^- := -\min(X, 0).$$

$X^+$  und  $X^-$  sind nichtnegative Zufallsvariablen. Ist mindestens einer der beiden Erwartungswerte  $E[X^+]$  bzw.  $E[X^-]$  endlich, dann können wir (ähnlich wie in Kapitel 1 für diskrete Zufallsvariablen) definieren:

**Definition.** Der Erwartungswert einer Zufallsvariable  $X : \Omega \rightarrow \overline{\mathbb{R}}$  mit  $E[X^+] < \infty$  oder  $E[X^-] < \infty$  ist

$$E[X] := E[X^+] - E[X^-] \in [-\infty, \infty].$$

**Notation:** Der Erwartungswert  $E[X]$  ist das Lebesgueintegral der messbaren Funktion  $X : \Omega \rightarrow \overline{\mathbb{R}}$  bzgl. des Maßes  $P$ . Daher verwenden wir auch folgende Notation:

$$E[X] = \int X \, dP = \int X(\omega) P(d\omega).$$

### Eigenschaften des Erwartungswerts

Nachdem wir den Erwartungswert einer allgemeinen Zufallsvariable  $X : \Omega \rightarrow \overline{\mathbb{R}}$  definiert haben, fassen wir nun einige grundlegende Eigenschaften des Erwartungswerts zusammen. Dazu bezeichnen wir mit

$$\mathcal{L}^1 = \mathcal{L}^1(P) = \mathcal{L}^1(\Omega, \mathcal{A}, P) := \{X : \Omega \rightarrow \overline{\mathbb{R}} \text{ Zufallsvariable} \mid E[|X|] < \infty\}$$

die Menge aller bzgl.  $P$  integrierbaren Zufallsvariablen. Für Zufallsvariablen  $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$  ist nach Lemma 6.3 sowohl  $E[X^+]$  als auch  $E[X^-]$  endlich. Also ist der Erwartungswert  $E[X]$  definiert und endlich.

**Satz 6.4.** 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  $\implies E[X] \geq 0$

(2). Die Zufallsvariable  $aX + bY$  ist bzgl.  $P$  integrierbar, und

$$E[aX + bY] = a \cdot E[X] + b \cdot E[Y].$$

Insbesondere ist der Erwartungswert monoton:

$$(3). \quad X \leq Y \text{ } P\text{-fast sicher} \implies E[X] \leq E[Y].$$

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

$$X \sim Y \quad : \iff \quad P[X = Y] = 1$$

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:

$$X \sim Y \quad \implies \quad E[X] = E[Y].$$

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 6.4 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 6.5.** *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  $X : \Omega \rightarrow \overline{\mathbb{R}}$ :

$$E[|X|] = 0 \quad \implies \quad X = 0 \quad P\text{-fast sicher.}$$

*Beweis.* Für eine Zufallsvariable  $X : \Omega \rightarrow \overline{\mathbb{R}}$  mit  $E[|X|] = 0$  und  $\varepsilon > 0$  gilt wegen der Monotonie und Linearität des Erwartungswerts:

$$P[|X| \geq \varepsilon] = E[I_{\{|X| \geq \varepsilon\}}] \leq E\left[\frac{|X|}{\varepsilon}\right] = \frac{1}{\varepsilon} E[|X|] = 0.$$

Für  $\varepsilon \searrow 0$  folgt

$$P[|X| > 0] = \lim_{\varepsilon \searrow 0} P[|X| \geq \varepsilon] = 0,$$

also  $X = 0$   $P$ -fast sicher.

Zudem folgt aus der Monotonie und Linearität des Erwartungswerts die Dreiecksungleichung:

$$E[|X + Y|] \leq E[|X| + |Y|] = E[|X|] + E[|Y|].$$

□

In der Analysis wird gezeigt, dass der Raum  $L^1(\Omega, \mathcal{A}, P)$  bzgl. der im Korollar definierten Norm ein Banachraum ist.

## Konvergenzsätze

Ein Vorteil des Lebesgueintegrals gegenüber anderen Integrationsbegriffen ist die Gültigkeit von sehr allgemeinen Konvergenzsätzen. Diese lassen sich zurückführen auf den folgenden fundamentalen Konvergenzsatz, der sich aus der oben skizzierten Konstruktion des Lebesgueintegrals ergibt:

**Satz 6.6 (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:*

$$E[\sup_{n \in \mathbb{N}} X_n] = E[\lim_{n \rightarrow \infty} X_n] = \lim_{n \rightarrow \infty} E[X_n] = \sup_{n \in \mathbb{N}} E[X_n].$$

Der Beweis findet sich in zahlreichen Lehrbüchern der Integrations- oder Wahrscheinlichkeitstheorie, siehe z.B. WILLIAMS: PROBABILITY WITH MARTINGALES, APPENDIX 5.

Eine erste wichtige Konsequenz des Satzes von der monotonen Konvergenz ist:

**Korollar 6.7.** *Für nichtnegative Zufallsvariablen  $X_i, i \in \mathbb{N}$ , gilt:*

$$E\left[\sum_{i=1}^{\infty} X_i\right] = \sum_{i=1}^{\infty} E[X_i].$$

*Beweis.*

$$\begin{aligned} E\left[\sum_{i=1}^{\infty} X_i\right] &= E\left[\lim_{n \rightarrow \infty} \sum_{i=1}^n X_i\right] \\ &= \lim_{n \rightarrow \infty} E\left[\sum_{i=1}^n X_i\right] && \text{(wegen monotoner Konvergenz)} \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n E[X_i] && \text{(wegen Linearität)} \\ &= \sum_{i=1}^{\infty} E[X_i]. \end{aligned}$$

□

**Bemerkung (Abzählbare Wahrscheinlichkeitsräume, Summation als Spezialfall von Integration).** Falls  $\Omega$  abzählbar ist, können wir jede Zufallsvariable  $X : \Omega \rightarrow \mathbb{R}$  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 6.7:

$$E[X] = \sum_{\omega \in \Omega} X(\omega) \cdot P[\{\omega\}].$$

Dieselbe Darstellung des Erwartungswerts gilt auch für allgemeine reellwertige Zufallsvariablen auf  $\Omega$ , falls der Erwartungswert definiert ist, d.h.  $E[X^+]$  oder  $E[X^-]$  endlich ist.

Insbesondere sehen wir, dass *Summation ein Spezialfall von Integration* ist: Ist  $\Omega$  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 also

$$\sum_{\omega \in \Omega} X(\omega) = \int X \, d\nu,$$

wobei  $\nu$  das durch  $\nu[A] = |A|$ ,  $A \subseteq \Omega$ , definierte Zählmaß ist.

*Konvergenzsätze wie der Satz von der monotonen Konvergenz lassen sich also auch auf Summen anwenden!*

**Beispiel.** Ist  $P$  die Gleichverteilung auf einer endlichen Menge  $\Omega$ , dann ist

$$E[X] = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} X(\omega)$$

das **arithmetische Mittel von  $X$** .

Wir beweisen nun noch zwei wichtige Konvergenzsätze, die sich aus dem Satz von der monotonen Konvergenz ergeben:

**Korollar 6.8 (Lemma von Fatou).** Seien  $X_1, X_2, \dots : \Omega \rightarrow \overline{\mathbb{R}}$  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:

$$\begin{aligned} E[\liminf X_n] &= E\left[\lim_{n \rightarrow \infty} \inf_{k \geq n} X_k\right] = \lim_{n \rightarrow \infty} E\left[\inf_{k \geq n} X_k\right] \\ &\leq \lim_{n \rightarrow \infty} \inf_{k \geq n} E[X_k] = \liminf_{n \rightarrow \infty} E[X_n], \end{aligned}$$

da die Folge der Infima monoton wachsend ist und durch die integrierbare Zufallsvariable  $Y$  nach unten beschränkt ist. Die zweite Aussage zeigt man analog.  $\square$

**Korollar 6.9 (Satz von der majorisierten Konvergenz, Lebesgue).** Sei  $X_n : \Omega \rightarrow \overline{\mathbb{R}}, n \in \mathbb{N}$ , eine  $P$ -fast sicher konvergente Folge von Zufallsvariablen. Existiert eine Majorante  $Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$  mit  $|X_n| \leq Y$  für alle  $n \in \mathbb{N}$ , dann gilt

$$E[\lim X_n] = \lim E[X_n]. \quad (6.1.2)$$

*Beweis.* Nach dem Lemma von Fatou gilt:

$$E[\liminf X_n] \leq \liminf E[X_n] \leq \limsup E[X_n] \leq E[\limsup X_n],$$

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.  $\square$

**Beispiel.** Wir betrachten Setzen mit Verdoppeln auf »Null« für eine Folge von fairen Münzwürfen. Bei Anfangseinsatz 1 beträgt das Kapital des Spielers nach  $n$  Münzwürfen

$$X_n = 2^n \cdot I_{\{n < T\}},$$

wobei  $T$  die Wartezeit auf die erste »Eins« 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 eine »Eins«, d.h. es gilt:

$$\lim_{n \rightarrow \infty} X_n = 0 \quad P\text{-fast sicher.}$$

Die Aussage (6.1.2) des Satzes von Lebesgue ist in dieser Situation nicht erfüllt!

## 6.2 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  $X : \Omega \rightarrow [0, \infty]$  aus der Verteilung von  $X$  berechnen kann.

## Diskrete Zufallsvariablen

Falls  $X$  nur abzählbar viele Werte annimmt, können wir die Zufallsvariable  $X$  auf folgende Weise als abzählbare Linearkombination von Indikatorfunktionen darstellen:

$$X = \sum_{a \in X(\Omega)} a \cdot I_{\{X=a\}}.$$

Es 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

$$E[X^+] < \infty \quad \text{oder} \quad E[X^-] < \infty.$$

Für Zufallsvariablen  $X : \Omega \rightarrow S$ , mit Werten in einer beliebigen abzählbaren Menge  $S$ , und eine Borel-messbare Funktion  $h : S \rightarrow \overline{\mathbb{R}}$  erhalten wir entsprechend

$$E[h(X)] = \sum_{a \in X(\Omega)} h(a) \cdot P[X = a], \quad (6.2.1)$$

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 1 gegebenen Definition überein.

## Allgemeine Zufallsvariablen

Die Berechnungsmethode (6.2.1) 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,  $X : \Omega \rightarrow S$  eine Zufallsvariable, und  $h : S \rightarrow [0, \infty]$  eine messbare Abbildung.

**Satz 6.10 (Transformationssatz).** *Unter den obigen Voraussetzungen gilt:*

$$E_P[h(X)] = \int h(X(\omega))P(d\omega) = \int h(x) \mu(dx) = E_\mu[h],$$

wobei  $\mu = P \circ X^{-1}$  die Verteilung von  $X$  unter  $P$  ist, und  $E_P$  bzw.  $E_\mu$  den Erwartungswert unter  $P$  bzw.  $\mu$  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 gilt:

$$E[h(X)] = \int I_B(X(\omega))P(d\omega) = P[X^{-1}(B)] = \mu[B] = \int I_B d\mu,$$

da  $I_B(X(\omega)) = I_{X^{-1}(B)}(\omega)$  gilt.

(2). Für Linearkombinationen  $h = \sum_{i=1}^n a_i I_{B_i}$  von Indikatorfunktionen mit  $n \in \mathbb{N}$ ,  $a_i \in \mathbb{R}$ , 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:

$$E[h(X)] = E[\lim h_n(X)] = \lim E[h_n(X)] = \lim \int h_n d\mu = \int h d\mu.$$

□

Das hier verwendete **Beweisverfahren der »maßtheoretischen Induktion«** wird noch sehr häufig auftreten: Wir zeigen eine Aussage

- (1). für Indikatorfunktionen,
- (2). für Elementarfunktionen,
- (3). für nichtnegative messbare Funktionen,
- (4). für allgemeine integrierbare Funktionen.

Mit maßtheoretischer Induktion zeigt man auch:

**Übung:** Jede  $\sigma(X)$ -messbare Zufallsvariable  $Y : \Omega \rightarrow \overline{\mathbb{R}}$  ist vom Typ  $Y = h(X)$  mit einer meßbaren Funktion  $h : S \rightarrow \mathbb{R}$ .

Nach Satz 6.10 ist der Erwartungswert  $E[T]$  einer reellwertigen Zufallsvariable  $T : \Omega \rightarrow [0, \infty]$  eindeutig bestimmt durch die Verteilung  $\mu_T = P \circ T^{-1}$ :

$$E[T] = \int t \mu_T(dt),$$

also auch durch die Verteilungsfunktion

$$F_T(t) = P[T \leq t] = \mu_T([0, t]), \quad t \in \mathbb{R}.$$

Der folgende Satz zeigt, wie man den Erwartungswert konkret aus  $F_T$  berechnet:

**Satz 6.11.** Für eine Zufallsvariable  $T : \Omega \rightarrow [0, \infty]$  gilt

$$E[T] = \int_0^\infty P[T > t] dt = \int_0^\infty (1 - F_T(t)) dt.$$

*Beweis.* Wegen

$$T(\omega) = \int_0^{T(\omega)} dt = \int_0^\infty I_{\{T > t\}}(\omega) dt$$

erhalten wir

$$E[T] = E \left[ \int_0^\infty I_{\{T > t\}} dt \right] = \int_0^\infty E [I_{\{T > t\}}] dt = \int_0^\infty P[T > t] dt.$$

Hierbei haben wir im Vorgriff auf Kapitel 9 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 9.1.  $\square$

**Bemerkung** (Stieltjesintegral). Das Lebesgue-Stieltjes-Integral  $\int h dF$  einer messbaren Funktion  $h : \mathbb{R} \rightarrow [0, \infty]$  bzgl. der Verteilungsfunktion  $F$  einer Wahrscheinlichkeitsverteilung  $\mu$  auf  $\mathbb{R}$  ist definiert als das Lebesgueintegral

$$\int h(t) dF(t) := \int h(t) \mu(dt).$$

Ist  $h$  stetig, dann läßt sich das Integral als Limes von Riemannsummen darstellen. Nach dem Transformationssatz gilt für eine Zufallsvariable  $T : \Omega \rightarrow [0, \infty]$ :

$$E[T] = \int t \mu_T(dt) = \int t dF_T(t).$$

Die Aussage von Satz 6.11 folgt hieraus formal durch partielle Integration.

**Beispiel (Exponentialverteilung).** Für eine exponentialverteilte Zufallsvariable  $T$  mit Parameter  $\lambda > 0$  erhalten wir:

$$E[T] = \int_0^\infty P[T > t] dt = \int_0^\infty e^{-\lambda t} dt = \frac{1}{\lambda}.$$

Es gilt also

$$\text{Mittlere Wartezeit} = \frac{1}{\text{Mittlere relative Häufigkeit pro Zeiteinheit}}$$

.

**Beispiel (Heavy tails).** Sei  $\alpha > 0$ . Für eine Zufallsvariable  $T : \Omega \rightarrow [0, \infty)$  mit

$$P[T > t] \sim t^{-\alpha} \quad \text{für } t \rightarrow \infty$$

gilt

$$E[T] = \int_0^{\infty} P[T > t] dt < \infty$$

genau dann, wenn  $\alpha > 1$ . Allgemeiner ist das  **$p$ -te Moment**

$$E[T^p] = \int_0^{\infty} P[T^p > t] dt = \int_0^{\infty} \underbrace{P[T > t^{1/p}]}_{\sim t^{\alpha/p}} dt$$

nur für  $p < \alpha$  endlich.

## Zufallsvariablen mit Dichten

Die Verteilungen vieler Zufallsvariablen haben eine Dichte bzgl. des Lebesguemaßes, oder bzgl. eines anderen geeigneten Referenzmaßes. Wir wollen uns nun überlegen, wie man in diesem Fall den Erwartungswert berechnet.

Sei  $(S, \mathcal{S})$  ein messbarer Raum und  $\nu$  ein Maß auf  $(S, \mathcal{S})$  (z.B. das Lebesguemaß oder eine Wahrscheinlichkeitsverteilung).

**Definition.** Eine **Wahrscheinlichkeitsdichte** auf  $(S, \mathcal{S}, \nu)$  ist eine messbare Funktion  $\varrho : S \rightarrow [0, \infty]$  mit

$$\int_S \varrho(x) \nu(dx) = 1.$$

**Satz 6.12.** (1). Ist  $\varrho$  eine Wahrscheinlichkeitsdichte auf  $(S, \mathcal{S}, \nu)$ , dann wird durch

$$\mu[B] := \int_B \varrho(x) \nu(dx) = \int I_B(x) \varrho(x) \nu(dx) \quad (6.2.2)$$

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

(2). Für eine messbare Funktion  $h : S \rightarrow [0, \infty]$  gilt

$$\int h(x) \mu(dx) = \int h(x) \varrho(x) \nu(dx). \quad (6.2.3)$$

Insbesondere folgt nach dem Transformationssatz:

$$E[h(X)] = \int h(x) \varrho(x) \nu(dx)$$

für jede Zufallsvariable  $X$  mit Verteilung  $\mu$ .

*Beweis.* Wir zeigen zunächst, dass  $\mu$  eine Wahrscheinlichkeitsverteilung ist: Sind  $B_1, B_2, \dots \in \mathcal{S}$  disjunkt, so folgt

$$\begin{aligned} \mu \left[ \bigcup_{i=1}^{\infty} B_i \right] &= \int I_{\bigcup_{i=1}^{\infty} B_i}(x) \cdot \varrho(x) \nu(dx) \\ &= \lim_{n \rightarrow \infty} \int I_{\bigcup_{i=1}^n B_i}(x) \cdot \varrho(x) \nu(dx) \quad (\text{wegen } \varrho \geq 0 \text{ und monotoner Konvergenz}) \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \int_{B_i} \varrho(x) \nu(dx) = \lim_{n \rightarrow \infty} \sum_{i=1}^n \mu[B_i] \\ &= \sum_{i=1}^{\infty} \mu[B_i]. \end{aligned}$$

Zudem gilt:

$$\mu[S] = \int \varrho d\nu = 1.$$

Die Aussage (6.2.3) 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 (6.2.3) in  $h$ .
- (3). Für allgemeine  $h \geq 0$  existiert eine Teilfolge  $h_n$  aus Elementarfunktionen mit  $h_n \nearrow h$ . Mit monotoner Konvergenz folgt

$$\int h d\mu = \lim \int h_n d\mu = \lim \int h_n \varrho d\nu = \int h \varrho d\nu.$$

□

**Bemerkung.** Durch (6.2.2) wird die Dichte  $\varrho(x)$  der Wahrscheinlichkeitsverteilung  $\mu$  bzgl. des Maßes  $\nu$  für  $\nu$ -fast alle  $x$  eindeutig festgelegt: Existiert  $\tilde{\varrho} \in \mathcal{L}^1(S, \mathcal{S}, \nu)$  mit

$$\int_B \varrho d\nu = \mu[B] = \int_B \tilde{\varrho} d\nu \quad \text{für alle } B \in \mathcal{S},$$

dann folgt:

$$\begin{aligned} \int_{\{\varrho > \tilde{\varrho}\}} (\varrho - \tilde{\varrho}) d\nu &= \int_{\{\varrho < \tilde{\varrho}\}} (\varrho - \tilde{\varrho}) d\nu = 0, \quad \text{also} \\ \int (\varrho - \tilde{\varrho})^+ d\nu &= \int (\varrho - \tilde{\varrho})^- d\nu = 0. \end{aligned}$$

Somit erhalten wir:

$$(\varrho - \tilde{\varrho})^+ = (\varrho - \tilde{\varrho})^- = 0 \quad \nu\text{-fast überall,}$$

und damit  $\varrho = \tilde{\varrho}$   $\nu$ -fast überall.

**Notation:** Die Aussage (6.2.3) rechtfertigt die folgende Notation für eine Wahrscheinlichkeitsverteilung  $\mu$  mit Dichte  $\varrho$  bzgl.  $\nu$ :

$$\mu(dx) = \varrho(x) \nu(dx) \quad \text{bzw.} \quad d\mu = \varrho d\nu \quad \text{bzw.} \quad \mu = \varrho \cdot \nu.$$

Für die nach der Bemerkung  $\nu$ -fast überall eindeutig bestimmte Dichte von  $\mu$  bzgl.  $\nu$  verwenden wir dementsprechend auch die folgende Notation:

$$\varrho(x) = \frac{d\mu}{d\nu}(x).$$

### Wichtige Spezialfälle:

(1). MASSENFUNKTION ALS DICHTE BZGL. DES ZÄHLMASSES.

Das Zählmaß auf einer abzählbaren Menge  $S$  ist das durch

$$\nu[B] = |B|, \quad B \subseteq S,$$

definierte Maß auf  $S$ . Die Gewichtsfunktion  $x \mapsto \mu[\{x\}]$  einer Wahrscheinlichkeitsverteilung  $\mu$  auf  $S$  ist die Dichte von  $\mu$  bzgl. des Zählmaßes  $\nu$ . Insbesondere ist die Massenfunktion einer diskreten Zufallsvariable  $X : \Omega \rightarrow S$  die Dichte der Verteilung von  $X$  bzgl.  $\nu$ :

$$\mu_X[B] = P[X \in B] = \sum_{a \in B} p_X(a) = \int_B p_X(a) \nu(da), \quad \text{für alle } B \subseteq S.$$

Die Berechnungsformel für den Erwartungswert diskreter Zufallsvariablen ergibt sich damit als Spezialfall von Satz 6.12:

$$E[h(X)] \stackrel{6.12}{=} \int h(a) p_X(a) \nu(da) = \sum_{a \in S} h(a) p_X(a) \quad \text{für alle } h : S \rightarrow [0, \infty].$$

(2). DICHTEN BZGL. DES LEBESGUEMASSES

Eine Wahrscheinlichkeitsverteilung  $\mu$  auf  $\mathbb{R}^d$  mit Borelscher  $\sigma$ -Algebra hat genau dann eine Dichte  $\varrho$  bzgl. des Lebesguemaßes  $\lambda$ , wenn

$$\mu[(-\infty, c_1] \times \dots \times (-\infty, c_d]] = \int_{-\infty}^{c_1} \dots \int_{-\infty}^{c_d} \varrho(x_1, \dots, x_d) dx_d \dots dx_1$$

für alle  $(c_1, \dots, c_d) \in \mathbb{R}^d$  gilt. Insbesondere hat die Verteilung einer reellwertigen Zufallsvariable  $X$  genau dann die Dichte  $f_X$  bzgl.  $\lambda$ , wenn

$$F_X(c) = \mu_X[(-\infty, c]] = \int_{-\infty}^c f_X(x) dx \quad \text{für alle } c \in \mathbb{R}$$



gilt. Die Verteilungsfunktion ist in diesem Fall eine Stammfunktion der Dichte, und damit  $\lambda$ -fast überall differenzierbar mit Ableitung

$$F'_X(x) = f_X(x) \quad \text{für fast alle } x \in \mathbb{R}.$$

Für den Erwartungswert ergibt sich:

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

für alle messbaren Funktionen  $h: \mathbb{R} \rightarrow \mathbb{R}$  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 Lebesguemaßes ist  $\varrho(x) = (2\pi)^{-1/2} \cdot e^{-x^2/2}$ . Damit ergibt sich für den Erwartungswert und die Varianz einer Zufallsvariable  $Z \sim N(0, 1)$ :

$$\begin{aligned} E[Z] &= \int_{-\infty}^{\infty} x \cdot (2\pi)^{-1/2} \cdot e^{-x^2/2} \, dx = 0, \quad \text{und} \\ \text{Var}[Z] &= E[(Z - E[Z])^2] = E[Z^2] \\ &= \int_{-\infty}^{\infty} x^2 \cdot (2\pi)^{-1/2} \cdot e^{-x^2/2} \, dx \\ &= \int_{-\infty}^{\infty} 1 \cdot (2\pi)^{-1/2} \cdot e^{-x^2/2} \, dx = 1. \end{aligned}$$

Hierbei haben wir im letzten Schritt partielle Integration benutzt.

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

$$E[X] = m + \sigma E[Z] = m,$$

und

$$\text{Var}[X] = \text{Var}[\sigma Z] = \sigma^2 \text{Var}[Z] = \sigma^2.$$

Die Parameter  $m$  und  $\sigma$  geben also den Erwartungswert und die Standardabweichung der Normalverteilung an.

- (3). **RELATIVE DICHTEN:** Seien  $\mu$  und  $\nu$  zwei Wahrscheinlichkeitsverteilungen auf einem messbaren Raum  $(S, \mathcal{S})$  mit Dichten  $f$  bzw.  $g$  bezüglich eines Referenzmaßes  $\lambda$  (z.B. Zählmaß oder Lebesguemaß). Gilt  $g > 0$   $\lambda$ -fast überall, dann hat  $\mu$  bzgl.  $\nu$  die Dichte

$$\frac{d\mu}{d\nu} = \frac{f}{g} = \frac{d\mu/d\lambda}{d\nu/d\lambda},$$

denn nach Satz 6.12 gilt:

$$\begin{aligned}\mu[B] &= \int_B f \, d\lambda = \int_B \frac{f}{g} g \, d\lambda \\ &= \int_B \frac{f}{g} \, d\nu \quad \text{für alle } B \in \mathcal{S}.\end{aligned}$$

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 ??.

## Existenz von Dichten

Wir geben abschließend ohne Beweis den Satz von Radon-Nikodym an. Dieser Satz besagt, dass eine Wahrscheinlichkeitsverteilung (oder allgemeiner ein  $\sigma$ -endliches Maß)  $\mu$  genau dann eine Dichte bzgl. eines anderen ( $\sigma$ -endlichen) Maßes  $\nu$  hat, wenn alle  $\nu$ -Nullmengen auch  $\mu$ -Nullmengen sind. Ein Maß  $\mu$  auf einem messbaren Raum  $(S, \mathcal{S})$  heißt  **$\sigma$ -endlich**, wenn eine Folge von messbaren Mengen  $B_n \in \mathcal{S}$  mit  $\mu[B_n] < \infty$  und  $S = \bigcup_{n \in \mathbb{N}} B_n$  existiert.

**Definition.** (1). Ein Maß  $\mu$  auf  $(S, \mathcal{S})$  heißt **absolutstetig** bzgl. eines anderen Maßes  $\nu$  auf demselben messbaren Raum ( $\mu \ll \nu$ ) falls für alle  $B \in \mathcal{S}$  gilt:

$$\nu[B] = 0 \quad \implies \quad \mu[B] = 0$$

(2). Die Maße  $\mu$  und  $\nu$  heißen **äquivalent** ( $\mu \sim \nu$ ), falls  $\mu \ll \nu$  und  $\nu \ll \mu$ .

**Beispiel.** Ein Diracmaß  $\delta_x, x \in \mathbb{R}$ , ist nicht absolutstetig bzgl. das Lebesguemaßes  $\lambda$  auf  $\mathbb{R}$ , denn es gilt  $\lambda[\{x\}] = 0$ , aber  $\delta_x[\{x\}] > 0$ . Umgekehrt ist auch das Lebesguemaß nicht absolutstetig bzgl. des Diracmaßes.

**Satz 6.13 (Radon-Nikodym).** Für  $\sigma$ -endliche Maße  $\mu$  und  $\nu$  gilt  $\mu \ll \nu$  genau dann, wenn eine Dichte  $\varrho \in \mathcal{L}^1(S, \mathcal{S}, \nu)$  existiert mit

$$\mu[B] = \int_B \varrho \, d\nu \quad \text{für alle } B \in \mathcal{S}.$$

Die eine Richtung des Satzes zeigt man leicht: Hat  $\mu$  eine Dichte bzgl.  $\nu$ , und gilt  $\nu[B] = 0$ , so folgt

$$\mu[B] = \int_B \varrho \, d\nu = \int \varrho \cdot I_B \, d\nu = 0,$$

da  $\varrho \cdot I_B = 0$   $\nu$ -fast überall. Der Beweis der Umkehrung ist nicht so einfach, und kann funktionalanalytisch erfolgen, siehe z.B. Klenke: „Wahrscheinlichkeitstheorie“. Einen stochastischen Beweis über Martingaltheorie werden wir in der Vorlesung „Stochastische Prozesse“ führen.

**Beispiel (Absolutstetigkeit von diskreten Wahrscheinlichkeitsverteilungen).** Sind  $\mu$  und  $\nu$  Wahrscheinlichkeitsverteilungen (oder  $\sigma$ -endliche 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  $\mu$  bzgl.  $\nu$  durch

$$\frac{d\mu}{d\nu}(x) = \begin{cases} \frac{\mu(x)}{\nu(x)} & \text{falls } \nu(x) \neq 0 \\ \text{beliebig} & \text{sonst} \end{cases}$$

gegeben. Man beachte, dass die Dichte nur für  $\nu$ -fast alle  $x$ , also für alle  $x$  mit  $\nu(x) \neq 0$ , eindeutig bestimmt ist.

## 6.3 Varianz, Kovarianz und lineare Regression

### Varianz und Standardabweichung

Sei  $X : \Omega \rightarrow \overline{\mathbb{R}}$  eine integrierbare Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ . Wie zuvor für diskrete Zufallsvariablen (s. Abschnitt 3.1) definieren wir auch im allgemeinen Fall die *Varianz*  $\text{Var}[X]$  und die *Standardabweichung*  $\sigma[X]$  durch

$$\text{Var}[X] := E[(X - E[X])^2], \quad \sigma[X] := \sqrt{\text{Var}[X]}.$$

Auch in diesem Fall folgen aus der Linearität des Erwartungswerts die Rechenregeln

$$\text{Var}[X] = E[X^2] - E[X]^2, \quad \text{und} \quad (6.3.1)$$

$$\text{Var}[aX + b] = \text{Var}[aX] = a^2 \cdot \text{Var}[X] \quad \text{für alle } a, b \in \mathbb{R}. \quad (6.3.2)$$

Insbesondere ist die Varianz genau dann endlich, wenn  $E[X^2]$  endlich ist. Nach Korollar 6.5 gilt zudem genau dann  $\text{Var}[X] = 0$ , wenn  $X$   $P$ -f.s. konstant gleich  $E[X]$  ist. Aufgrund des Transformationssatzes für den Erwartungswert können wir die Varianz auch allgemein aus der Verteilung  $\mu_X = P \circ X^{-1}$  berechnen:

**Korollar 6.14.** Die Varianz  $\text{Var}[X]$  hängt nur von der Verteilung  $\mu_X = P \circ X^{-1}$  ab:

$$\text{Var}[X] = \int (x - \bar{x})^2 \mu_X(dx) \quad \text{mit} \quad \bar{x} = E[X] = \int x \mu(dx).$$

*Beweis.* Nach Satz 6.12 gilt

$$\text{Var}[X] = E[(X - E[X])^2] = \int (x - E[X])^2 \mu_X(dx)$$

mit  $E[X] = \int x \mu_X(dx)$ . □

**Beispiel (Empirisches Mittel und empirische Varianz).** Ist die zugrundeliegende Wahrscheinlichkeitsverteilung auf  $\Omega$  eine empirische Verteilung

$$P = \frac{1}{n} \sum_{i=1}^n \delta_{\omega_i}$$

von  $n$  Elementen  $\omega_1, \dots, \omega_n$  aus einer Grundgesamtheit (z.B. alle Einwohner von Bonn, oder eine Stichprobe daraus), dann ist die Verteilung einer Abbildung  $X: \Omega \rightarrow S$  (statistisches Merkmal, z.B. Alter der Einwohner von Bonn) gerade die empirische Verteilung der auftretenden Werte  $x_i = X(\omega_i)$ :

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

Die Gewichte der empirischen Verteilung sind die relativen Häufigkeiten

$$\mu_X(a) = \frac{h(a)}{n}, \quad h(a) = |\{1 \leq i \leq n : x_i = a\}|.$$

Für den Erwartungswert einer Funktion  $g(X)$ ,  $g: S \rightarrow \mathbb{R}$ , ergibt sich

$$E[g(X)] = \sum_{a \in \{x_1, \dots, x_n\}} g(a) \cdot \frac{h(a)}{n} = \frac{1}{n} \sum_{i=1}^n g(x_i),$$

d.h. der Erwartungswert bzgl. der empirischen Verteilung ist das arithmetische Mittel der Werte  $g(x_i)$ .

Ist  $X$  reellwertig, so erhalten wir als Erwartungswert und Varianz das **empirische Mittel**

$$E[X] = \sum_{a \in \{x_1, \dots, x_n\}} a \cdot \frac{h(a)}{n} = \frac{1}{n} \sum_{i=1}^n x_i =: \bar{x}_n,$$

und die **empirische Varianz**

$$\begin{aligned} \text{Var}[X] &= E[(X - E[X])^2] = \sum_{a \in \{x_1, \dots, x_n\}} (a - \bar{x}_n)^2 \cdot \frac{h(a)}{n} \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \overline{(x^2)}_n - (\bar{x}_n)^2 =: \sigma_n^2. \end{aligned}$$

Sind die  $x_i$  selbst unabhängige Stichproben von einer Wahrscheinlichkeitsverteilung  $\mu$ , dann ist die empirische Verteilung  $n^{-1} \sum_{i=1}^n \delta_{x_i}$  nach dem Gesetz der großen Zahlen eine Approximation von  $\mu$ , siehe Abschnitt 7.2 unten. Daher verwendet man das Stichprobenmittel  $\bar{x}_n$  und die Stichprobenvarianz  $\sigma_n^2$  bzw. die renormierte Stichprobenvarianz

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

in der Statistik, um den Erwartungswert und die Varianz einer zugrundeliegenden (unbekannten) Verteilung zu schätzen.

**Beispiel (Exponentialverteilung).** Für eine zum Parameter  $\lambda > 0$  exponentialverteilte Zufallsvariable  $T$  gilt  $E[T] = \frac{1}{\lambda}$ . Mit partieller Integration folgt zudem:

$$\begin{aligned} E[T^2] &= \int_0^{\infty} t^2 f_T(t) dt = \int_0^{\infty} t^2 \lambda e^{-\lambda t} dt \\ &= \int_0^{\infty} 2te^{-\lambda t} dt = \frac{2}{\lambda} \int_0^{\infty} t f_T(t) dt \\ &= \frac{2}{\lambda} E[T] = \frac{2}{\lambda^2}, \end{aligned}$$

also

$$\sigma(T) = \sqrt{\text{Var}[T]} = (E[T^2] - E[T]^2)^{1/2} = \frac{1}{\lambda}.$$

Die Standardabweichung ist also genauso groß wie der Erwartungswert!

**Beispiel (Heavy Tails).** Eine Zufallsvariable  $X : \Omega \rightarrow \mathbb{R}$  mit Verteilungsdichte

$$f_X(x) \sim |x|^{-p} \quad \text{für } |x| \rightarrow \infty$$

ist integrierbar für  $p > 2$ . Für  $p \in (2, 3]$  gilt jedoch

$$\text{Var}[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x) dx = \infty.$$

## Quadratintegrierbare Zufallsvariablen

Für einen gegebenen Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  bezeichnen wir mit  $\mathcal{L}^2(\Omega, \mathcal{A}, P)$  den Raum aller bezüglich  $P$  quadratintegrierbaren Zufallsvariablen:

$$\mathcal{L}^2(\Omega, \mathcal{A}, P) = \{X : \Omega \rightarrow \overline{\mathbb{R}} \text{ messbar} \mid E[X^2] < \infty\}.$$

Der Raum ist ein Unterraum des Vektorraums aller  $\mathcal{A}/\mathcal{B}(\overline{\mathbb{R}})$  messbaren Abbildungen, denn für  $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$  und  $a \in \mathbb{R}$  gilt:

$$E[(aX + Y)^2] \leq E[2(aX)^2 + 2Y^2] = 2a^2E[X^2] + 2E[Y^2] < \infty.$$

Zudem gilt

$$\mathcal{L}^2(\Omega, \mathcal{A}, P) \subseteq \mathcal{L}^1(\Omega, \mathcal{A}, P),$$

denn aus  $|X| \leq (X^2 + 1)/2$  folgt

$$E[|X|] \leq E\left[\frac{1}{2}(X^2 + 1)\right] = \frac{1}{2}(E[X^2] + 1) < \infty$$

für alle  $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ . Hierbei 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 (6.3.1) ist umgekehrt eine Zufallsvariable aus  $\mathcal{L}^1$  genau dann in  $\mathcal{L}^2$  enthalten, wenn sie endliche Varianz hat.

Auf dem Vektorraum

$$L^2(\Omega, \mathcal{A}, P) = \mathcal{L}^2(\Omega, \mathcal{A}, P) / \sim$$

der Äquivalenzklassen von  $P$ -fast sicher gleichen quadratintegrierbaren Zufallsvariablen wird durch

$$(X, Y)_{L^2} := E[XY] \quad \text{und} \quad \|X\|_{L^2} := (X, X)_{L^2}^{1/2}$$

ein Skalarprodukt und eine Norm definiert. Hierbei ist der Erwartungswert  $E[XY]$  wegen  $|XY| \leq (X^2 + Y^2)/2$  definiert. Insbesondere gilt die **Cauchy-Schwarz-Ungleichung**

$$|E[XY]| \leq E[X^2]^{1/2} \cdot E[Y^2]^{1/2} \quad \text{für alle } X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P).$$

In der Analysis wird gezeigt, dass  $L^2(\Omega, \mathcal{A}, P)$  bzgl. des  $L^2$ -Skalarprodukts ein Hilbertraum, also vollständig bzgl. der  $L^2$ -Norm ist.

## Beste Prognosen

Angenommen wir wollen den Ausgang eines Zufallsexperiments vorhersagen, dass durch eine reellwertige Zufallsvariable  $X : \Omega \rightarrow \mathbb{R}$  beschrieben wird. Welches ist der beste Prognosewert  $a$  für  $X(\omega)$ , 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 (**mean square error**)

$$\text{MSE} = E[(X - a)^2]$$

bzw. die Wurzel (root mean square error)

$$\text{RMSE} = \text{MSE}^{1/2} = \|X - a\|_{L^2(\Omega, \mathcal{A}, P)}.$$

**Satz 6.15 (Erwartungswert als bester  $L^2$ -Prognosewert).** *Ist  $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ , dann gilt für alle  $a \in \mathbb{R}$ :*

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

Der mittlere quadratische Fehler des Prognosewertes  $a$  ist also die Summe der Varianz von  $X$  und des Quadrats des **Bias** (systematischer bzw. mittlerer Prognosefehler)  $a - E[X]$ :

$$\text{MSE} = \text{Varianz} + \text{Bias}^2.$$

Insbesondere ist der mittlere quadratische Fehler genau für  $a = E[X]$  minimal.

*Beweis.* Für  $a \in \mathbb{R}$  gilt wegen der Linearität des Erwartungswertes:

$$\begin{aligned} E[(X - a)^2] &= E[(X - E[X] + E[X] - a)^2] \\ &= E[(X - E[X])^2] + 2E[(X - E[X]) \cdot (E[X] - a)] + E[(E[X] - a)^2] \\ &\quad = \underbrace{(E[X] - E[X]) \cdot (E[X] - a)}_{=0} \\ &= \text{Var}[X] + (E[X] - a)^2. \end{aligned}$$

□

Verwendet man eine andere Norm, um den Prognosefehler zu messen, dann ergeben sich im Allgemeinen andere beste Prognosewerte. Beispielsweise gilt:

**Satz 6.16 (Median als bester  $L^1$ -Prognosewert).** *Ist  $X \in L^1(\Omega, \mathcal{A}, P)$  und  $m$  ein Median der Verteilung von  $X$ , dann gilt für alle  $a \in \mathbb{R}$ :*

$$E[|X - a|] \geq E[|X - m|]$$

.

*Beweis.* Für  $m \geq a$  folgt die Behauptung aus der Identität

$$|X - m| - |X - a| \leq (m - a)(I_{(-\infty, m)}(X) - I_{[m, \infty)}(X))$$

durch Bilden des Erwartungswertes. Der Beweis für  $m \leq a$  verläuft analog. □

Insbesondere minimieren Stichprobenmittel und Stichprobenmedian einer Stichprobe  $x_1, \dots, x_n \in \mathbb{R}$  also die Summe der quadratischen bzw. absoluten Abweichungen  $\sum (x_i - a)^2$  bzw.  $\sum |x_i - a|$ .

## Kovarianz und Korrelation

Seien  $X$  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**

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

und den **Korrelationskoeffizienten**

$$\rho[X, Y] := \frac{\text{Cov}[X, Y]}{\sigma[X]\sigma[Y]},$$

falls  $\sigma[X] \cdot \sigma[Y] \neq 0$ . Die Zufallsvariablen  $X$  und  $Y$  heißen unkorreliert, falls  $\text{Cov}[X, Y] = 0$  gilt, d.h. falls

$$E[XY] = E[X] \cdot E[Y].$$

Um die Kovarianz zu berechnen, benötigen wir die gemeinsame Verteilung der Zufallsvariablen  $X$  und  $Y$ . Aus dem Transformationssatz für den Erwartungswert folgt:

**Korollar 6.17.** Die Kovarianz  $\text{Cov}[X, Y]$  hängt nur von der gemeinsamen Verteilung

$$\mu_{X,Y} = P \circ (X, Y)^{-1}$$

der Zufallsvariablen  $X$  und  $Y$  ab:

$$\text{Cov}[X, Y] = \int \left( x - \int z \mu_X(dz) \right) \left( y - \int z \mu_Y(dz) \right) \mu_{X,Y}(dx dy).$$

*Beweis.* Nach dem Transformationssatz gilt

$$\begin{aligned} \text{Cov}[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= \int \left( x - \int z \mu_X(dz) \right) \left( y - \int z \mu_Y(dz) \right) \mu_{X,Y}(dx dy). \end{aligned}$$

□

Aus der Linearität des Erwartungswertes folgt, dass die Abbildung  $\text{Cov} : \mathcal{L}^2 \times \mathcal{L}^2 \rightarrow \mathbb{R}$  symmetrisch und bilinear ist. Die Varianz  $\text{Var}[X] = \text{Cov}[X, X]$  ist die zugehörige quadratische Form. Insbesondere gilt wie im diskreten Fall:

$$\text{Var} \left[ \sum_{i=1}^n X_i \right] = \sum_{i=1}^n \text{Var}[X_i] + 2 \cdot \sum_{\substack{i,j=1 \\ i < j}}^n \text{Cov}[X_i, X_j].$$



Sind die Zufallsvariablen  $X_1, \dots, X_n$  unkorreliert, dann folgt:

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

Die folgende Aussage ist ein Spezialfall der Cauchy-Schwarz-Ungleichung. Wir geben trotzdem einen vollständigen Beweis, da dieser auch in Zusammenhang mit linearer Regression von Interesse ist.

**Satz 6.18** (Cauchy-Schwarz). (1). Für  $X, Y \in \mathcal{L}^2$  gilt:

$$|\text{Cov}[X, Y]| \leq \text{Var}[X]^{1/2} \cdot \text{Var}[Y]^{1/2} = \sigma[X] \cdot \sigma[Y]. \quad (6.3.3)$$

(2). Im Fall  $\sigma[X] \cdot \sigma[Y] \neq 0$  gilt für den Korrelationskoeffizienten

$$|\varrho[X, Y]| \leq 1. \quad (6.3.4)$$

Gleichheit in (6.3.3) bzw. (6.3.4) gilt genau dann, wenn ein  $a \neq 0$  und ein  $b \in \mathbb{R}$  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 (6.3.3) ist trivialerweise erfüllt. Wir nehmen nun an, dass  $\sigma[X] \neq 0$  gilt.

(1). Für  $a \in \mathbb{R}$  gilt:

$$\begin{aligned} 0 &\leq \text{Var}[Y - aX] = \text{Var}[Y] - 2a \text{Cov}[X, Y] + a^2 \text{Var}[X] \\ &= \left( a \cdot \sigma[X] - \frac{\text{Cov}[X, Y]}{\sigma[X]} \right)^2 - \frac{\text{Cov}[X, Y]^2}{\text{Var}[X]} + \text{Var}[Y]. \end{aligned} \quad (6.3.5)$$

Da der erste Term für  $a := \frac{\text{Cov}[X, Y]}{\sigma[X]^2}$  verschwindet, folgt:

$$\text{Var}[Y] - \frac{\text{Cov}[X, Y]^2}{\text{Var}[X]} \geq 0.$$

(2). Die Ungleichung  $|\varrho[X, Y]| \leq 1$  folgt unmittelbar aus (6.3.3). Zudem gilt genau dann Gleichheit in (6.3.5) bzw. (6.3.3), wenn  $\text{Var}[Y - aX] = 0$  gilt, also  $Y - aX$   $P$ -fast sicher konstant ist. In diesem Fall folgt

$$\text{Cov}[X, Y] = \text{Cov}[X, aX] = a \text{Var}[X],$$

also hat  $\varrho[X, Y]$  dasselbe Vorzeichen wie  $a$ .

□

**Beispiel (Empirischer Korrelationskoeffizient).** Ist die zugrundeliegende Wahrscheinlichkeitsverteilung eine empirische Verteilung  $P = \frac{1}{n} \sum_{i=1}^n \delta_{\omega_i}$ , und sind  $X, Y : \Omega \rightarrow \mathbb{R}$  reellwertige Abbildungen (statistische Merkmale), dann gilt

$$\mu_{X,Y} = \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)} \quad \text{mit} \quad x_i = X(\omega_i) \quad \text{und} \quad y_i = Y(\omega_i).$$

Als Kovarianz ergibt sich

$$\text{Cov}[X, Y] = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) = \frac{1}{n} \left( \sum_{i=1}^n x_i y_i \right) - \bar{x}_n \bar{y}_n.$$

Der entsprechende **empirische Korrelationskoeffizient** der Daten  $(x_i, y_i)$ ,  $1 \leq i \leq n$ , ist

$$\varrho[X, Y] = \frac{\text{Cov}[X, Y]}{\sigma[X]\sigma[Y]} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\left( \sum_{i=1}^n (x_i - \bar{x}_n)^2 \right)^{1/2} \left( \sum_{i=1}^n (y_i - \bar{y}_n)^2 \right)^{1/2}} =: r_n.$$

Den empirischen Korrelationskoeffizienten verwendet man als Schätzer für die Korrelation von Zufallsgrößen mit unbekanntem Verteilungen.

Die Grafiken 6.3 und 6.3 zeigen Stichproben mit verschiedenen Korrelationskoeffizienten  $\varrho$ .

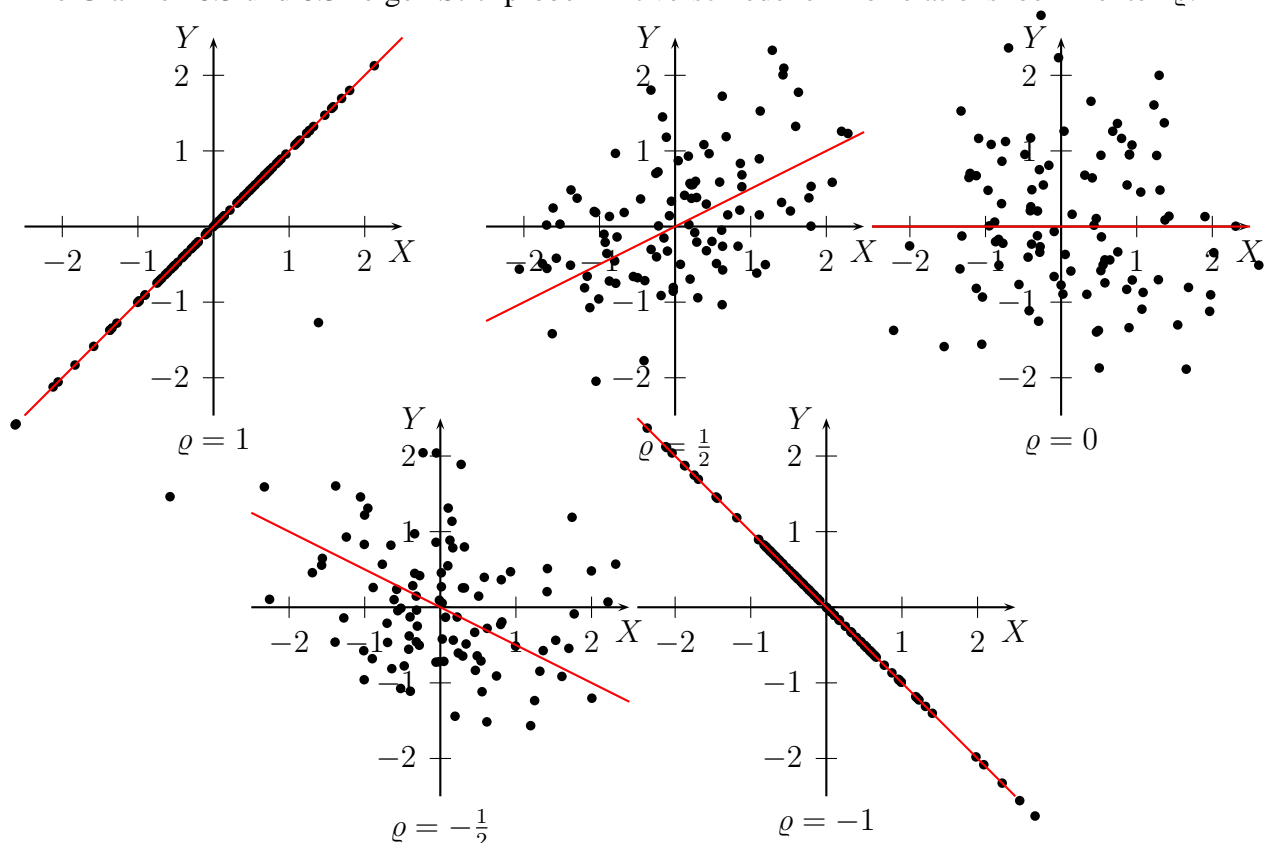


Abbildung 6.2: Stichprobe von 100 Punkten von korrelierten Standardnormalverteilungen

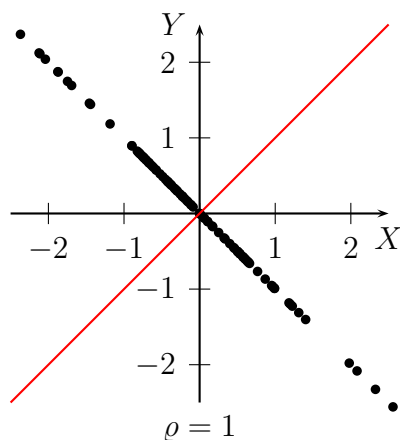


Abbildung 6.3: Stichprobe von 100 Punkten von korrelierten Standardnormalverteilungen

### Anwendung auf lineare Prognose (Regression)

Seien  $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$  Zufallsvariablen mit  $\sigma[X] \neq 0$ . Angenommen, wir kennen den Wert  $X(\omega)$  in einem Zufallsexperiment und suchen die beste *lineare* Vorhersage

$$\hat{Y}(\omega) = aX(\omega) + b, \quad (a, b \in \mathbb{R}) \quad (6.3.6)$$

für  $Y(\omega)$  im quadratischen Mittel, d.h. den Minimierer des mittleren quadratischen Fehlers,

$$\text{MSE} := E[(\hat{Y} - Y)^2],$$

unter alle Zufallsvariablen  $\hat{Y}$ , die affine Funktionen von  $X$  sind.

**Korollar 6.19.** *Der mittlere quadratische Fehler ist minimal unter allen Zufallsvariablen  $\hat{Y} = aX + b$  ( $a, b \in \mathbb{R}$ ) für*

$$\hat{Y}(\omega) = E[Y] + \frac{\text{Cov}[X, Y]}{\text{Var}[X]} \cdot (X(\omega) - E[X]).$$

*Beweis.* Es gilt

$$\begin{aligned} \text{MSE} &= \text{Var}[Y - \hat{Y}] + E[Y - \hat{Y}]^2 \\ &= \text{Var}[Y - aX] + (E[Y] - aE[X] - b)^2. \end{aligned}$$

Der zweite Term ist minimal für

$$b = E[Y] - aE[X],$$

und der erste Term für

$$a = \frac{\text{Cov}[X, Y]}{\sigma[X]^2},$$

siehe den Beweis der Cauchy-Schwarz-Ungleichung, Satz 6.18. Die bzgl. des mittleren quadratischen Fehlers optimale Prognose für  $Y$  gestützt auf  $X$  ist also

$$\hat{Y}_{\text{opt}} = aX + b = E[Y] + a(X - E[X]).$$

□

**Beispiel (Regressionsgerade, Methode der kleinsten Quadrate).** Im Beispiel der empirischen Verteilung von oben erhalten wir die Regressionsgerade  $y = ax + b$ , die die Quadratsumme

$$\sum_{i=1}^n (ax_i + b - y_i)^2 = n \cdot \text{MSE}$$

der Abweichungen minimiert. Es gilt

$$a = \frac{\text{Cov}[X, Y]}{\sigma[X]^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

und

$$b = E[Y] - a \cdot E[X] = \bar{y}_n - a \cdot \bar{x}_n.$$

Die Regressionsgeraden sind in Grafik 6.3 eingezeichnet.

**Beispiel (Zweidimensionale Normalverteilung).** Die zweidimensionale Normalverteilung  $N(m, C)$  ist die Verteilung im  $\mathbb{R}^2$  mit Dichte

$$f_{m,C}(x) = \frac{1}{2\pi \cdot \sqrt{\det C}} \cdot \exp\left(-\frac{1}{2}(x - m) \cdot C^{-1}(x - m)\right), \quad x \in \mathbb{R}^2.$$

Hierbei ist  $m \in \mathbb{R}^2$  und  $C = \begin{pmatrix} v_1 & c \\ c & v_2 \end{pmatrix}$  eine symmetrische positiv-definite Matrix mit Koeffizienten  $c \in \mathbb{R}$  und  $v_1, v_2 > 0$ . Mit  $\sigma_i := \sqrt{v_i}, i = 1, 2$ , und  $\varrho := \frac{c}{\sigma_1 \sigma_2}$  gilt:

$$\det C = v_1 v_2 - c^2 = \sigma_1^2 \sigma_2^2 \cdot (1 - \varrho^2), \quad \text{und}$$

$$C^{-1} = \frac{1}{\det C} \begin{pmatrix} v_2 & -c \\ -c & v_1 \end{pmatrix} = \frac{1}{1 - \varrho^2} \cdot \begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\varrho}{\sigma_1 \sigma_2} \\ -\frac{\varrho}{\sigma_1 \sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix},$$

also

$$f_{m,C}(x) = \frac{\exp\left(-\frac{1}{2(1-\varrho^2)} \left[ \left(\frac{x_1 - m_1}{\sigma_1}\right)^2 - 2\varrho \frac{x_1 - m_1}{\sigma_1} \cdot \frac{x_2 - m_2}{\sigma_2} + \left(\frac{x_2 - m_2}{\sigma_2}\right)^2 \right]\right)}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \varrho^2}}.$$

Die folgende Aussage zeigt, dass die Koeffizienten  $m_i, \sigma_i$  und  $\varrho$  tatsächlich der Mittelwert, die Standardabweichung und die Korrelation der Koordinaten  $x_1$  und  $x_2$  sind:

*Behauptung:*

- (1).  $f_{m,C}$  ist eine Wahrscheinlichkeitsdichte bzgl. des Lebesguemaßes im  $\mathbb{R}^2$ .
- (2). Für reellwertige Zufallsvariablen  $X_1, X_2$  mit gemeinsamer Verteilung  $\mu_{X_1, X_2} = N(m, C)$  und  $i = 1, 2$  gilt

$$E[X_i] = m_i, \quad \text{Var}[X_i] = v_i, \quad \text{und} \quad \text{Cov}[X_1, X_2] = c, \quad (6.3.7)$$

d.h.  $m$  ist der Mittelwertvektor und  $C = (\text{Cov}[X_i, X_j])_{i,j}$  die Kovarianzmatrix der Normalverteilung  $N(m, C)$ .

Der Beweis der Behauptung wird der Leserin/dem Leser als Übung überlassen - wir zeigen nur exemplarisch die Berechnung der Kovarianz im Fall  $m = 0$ . Mit quadratischer Ergänzung können wir den Exponenten in der Dichte  $f_{0,C}(x)$  schreiben als

$$-\frac{1}{2(1-\varrho^2)} \left( \frac{x_1}{\sigma_1} - \varrho \frac{x_2}{\sigma_2} \right)^2 - \frac{1}{2} \left( \frac{x_2}{\sigma_2} \right)^2.$$

Mit  $\tilde{m}(x_2) = \frac{x_2 \varrho \sigma_1}{\sigma_2}$  erhalten wir dann nach dem Satz von Fubini:

$$\begin{aligned} & \int_{\mathbb{R}^2} x_1 x_2 f_{0,C}(x) dx \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}} \int \int x_1 x_2 \exp\left[-\frac{1}{2(1-\varrho^2)\sigma_1^2} (x_1 - \tilde{m}(x_2))^2\right] dx_1 \exp\left(-\frac{x_2^2}{2\sigma_2^2}\right) dx_2 \\ &= \frac{1}{\sqrt{2\pi\sigma_2^2}} \int \underbrace{x_2 \cdot \tilde{m}(x_2)}_{x_2^2 \varrho \sigma_1 / \sigma_2} \cdot \exp\left(-\frac{x_2^2}{2\sigma_2^2}\right) dx_2 = \varrho \sigma_1 \sigma_2 = c, \end{aligned}$$

wobei wir im zweiten und dritten Schritt die Formeln für den Erwartungswert und die Varianz von eindimensionalen Normalverteilungen verwendet haben. Nach dem Transformationssatz ergibt sich:

$$E[X_1 X_2] = \int x_1 x_2 \mu_{X_1, X_2}(dx) = c.$$

Da auf ähnliche Weise  $E[X_1] = E[X_2] = 0$  folgt, ist  $c$  die Kovarianz von  $X_1$  und  $X_2$ .

**Bemerkung.** Ist  $X = (X_1, X_2)$  ein  $N(m, C)$ -verteilter Zufallsvektor, dann ist jede Linearkombination  $Y = \alpha_1 X_1 + \alpha_2 X_2$ ,  $\alpha \in \mathbb{R}^2$ , normalverteilt mit Mittelwert  $\alpha \cdot m$  und Varianz  $\alpha \cdot C \alpha$ . Auch dies kann man durch eine explizite Berechnung der Verteilungsfunktion aus der gemeinsamen Dichte von  $X_1$  und  $X_2$  zeigen. Wir werden multivariate Normalverteilungen systematischer in Abschnitt 9.3 untersuchen, und dort auch einen eleganteren Beweis der letzten Aussage mithilfe von charakteristischen Funktionen geben.

**Beispiel (Autoregressiver Prozess).** 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“

$$X_n = \underbrace{\alpha X_{n-1}}_{\substack{\text{lineares} \\ \text{Bewegungsgesetz}}} + \underbrace{\varepsilon Z_n}_{\substack{\text{zufällige Störung,} \\ \text{Rauschen}}}, \quad n \in \mathbb{N}, \quad (6.3.8)$$

definierte stochastische Prozess  $(X_n)_{n=0,1,2,\dots}$  heißt **autoregressiver Prozess AR(1)** mit Parametern  $\varepsilon, \alpha \in \mathbb{R}$ . Autoregressive Prozesse werden zur Modellierung von Zeitreihen eingesetzt. Im allgemeineren autoregressiven Modell AR( $p$ ),  $p \in \mathbb{N}$ , mit Parametern  $\varepsilon, \alpha_1, \dots, \alpha_p \in \mathbb{R}$  lautet das Bewegungsgesetz

$$X_n = \sum_{i=1}^p \alpha_i X_{n-i} + \varepsilon Z_n, \quad n \geq p.$$

Grafik 6.3 zeigt simulierte Trajektorien von AR(1)- und AR(2)-Prozessen:

Das folgende Lemma fasst einige grundlegende Eigenschaften des AR(1) Modells zusammen.

**Lemma 6.20.** Für den AR(1)-Prozess mit Parametern  $\varepsilon, \alpha$  und  $m \in \mathbb{R}, \sigma > 0$  gilt:

$$(1). X_{n-1} \sim N(m, \sigma^2) \implies X_n \sim N(\alpha m, \alpha^2 \sigma^2 + \varepsilon^2).$$

(2). Für  $|\alpha| < 1$  ist die Verteilung  $\mu = N(0, \frac{\varepsilon^2}{1-\alpha^2})$  ein Gleichgewicht, d.h.

$$X_0 \sim \mu \implies X_n \sim \mu \quad \forall n \in \mathbb{N}.$$

Bei Startverteilung  $P \circ X_0^{-1} = \mu$  gilt:

$$\text{Cov}[X_n, X_{n-k}] = \alpha^k \cdot \frac{\varepsilon^2}{1-\alpha^2} \quad \text{für alle } 0 \leq k \leq n.$$

### Exponentieller Abfall der Korrelationen

*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 (6.3.7). Der Beweis der übrigen Aussagen wird dem Leser als Übungsaufgabe überlassen.  $\square$

**Bemerkung.** (1). Der AR(1)-Prozess ist eine *Markovkette* mit Übergangswahrscheinlichkeiten  $p(x, \cdot) = N(\alpha x, \varepsilon^2)$ , s. Abschnitt 9.1 unten.

(2). Ist die gemeinsame Verteilung der Startwerte  $X_0, X_1, \dots, X_{p-1}$  eine multivariate Normalverteilung, dann ist der AR( $p$ )-Prozess ein *Gaussprozess*, d.h. die gemeinsame Verteilung von  $X_0, X_1, \dots, X_n$  ist für jedes  $n \in \mathbb{N}$  eine multivariate Normalverteilung.

## Unabhängigkeit und Unkorreliertheit

Wir zeigen abschließend, dass auch für allgemeine Zufallsvariablen  $X$  und  $Y$  aus Unabhängigkeit die Unkorreliertheit von beliebigen Funktionen  $f(X)$  und  $g(Y)$  folgt. Seien  $X : \Omega \rightarrow S$  und  $Y : \Omega \rightarrow T$  Zufallsvariablen mit Werten in messbaren Räumen  $(S, \mathcal{S})$  und  $(T, \mathcal{T})$ .

**Satz 6.21.** *Es sind äquivalent:*

(1). *Die Zufallsvariablen  $X$  und  $Y$  sind unabhängig, d.h.*

$$P[X \in A, Y \in B] = P[X \in A] \cdot P[Y \in B] \quad \text{für alle } A \in \mathcal{S} \text{ und } B \in \mathcal{T}$$

(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.*

$$E[f(X) \cdot g(Y)] = E[f(X)] \cdot E[g(Y)]. \quad (6.3.9)$$

*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 (6.3.9) für Indikatorfunktionen  $f$  und  $g$  erfüllt. Wegen der Linearität beider Seiten dieser Gleichung in  $f$  und  $g$  gilt (6.3.9) 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 (6.3.9) folgt durch monotone Konvergenz. Allgemeine Funktionen zerlegen wir in ihren Positiv- und Negativanteil, und wenden die Aussage auf diese an. Also gilt  $\text{Cov}[f(X), g(Y)] = 0$  für alle messbaren  $f, g$  mit  $f, g \geq 0$  bzw.  $f(X), g(Y) \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ .  $\square$

**Korollar 6.22.** Sind  $X, Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$  unabhängig, so gilt:

$$X \cdot Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P) \quad \text{und} \quad E[XY] = E[X] \cdot E[Y].$$

*Beweis.* Nach Satz 6.21 gilt:

$$E[|XY|] = E[|X|] \cdot E[|Y|] < \infty.$$

Die Formel für  $E[XY]$  folgt durch die Zerlegungen  $X = X^+ - X^-$  und  $Y = Y^+ - Y^-$ .  $\square$



# Kapitel 7

## 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.

### 7.1 Grundlegende Ungleichungen und Konvergenz von Zufallsvariablen

#### 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:

$$P \left[ \lim_{n \rightarrow \infty} Y_n = Y \right] = P[\{\omega \in \Omega | Y_n(\omega) \rightarrow Y(\omega)\}] = 1.$$

(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

$$\lim_{n \rightarrow \infty} P[|Y_n - Y| > \varepsilon] = 0 \quad \text{für alle } \varepsilon > 0 \text{ gilt.}$$

(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

$$\lim_{n \rightarrow \infty} E[|Y_n - Y|^p] = 0.$$

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 7.1.** (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  $\varepsilon > 0$ :

$$\begin{aligned} 1 &= P[|Y_n - Y| < \varepsilon \text{ schließlich}] \\ &= P\left[\bigcup_m \bigcap_{n \geq m} \{|Y_n - Y| < \varepsilon\}\right] \\ &= \lim_{m \rightarrow \infty} P\left[\bigcap_{n \geq m} \{|Y_n - Y| < \varepsilon\}\right] \\ &\leq \lim_{m \rightarrow \infty} \inf_{n \geq m} P[|Y_n - Y| < \varepsilon] \\ &= \liminf_{n \rightarrow \infty} P[|Y_n - Y| < \varepsilon]. \end{aligned}$$

Es folgt  $\lim_{n \rightarrow \infty} P[|Y_n - Y| < \varepsilon] = 1$  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  $\sigma$ -Algebra. Wir betrachten die Zufallsvariablen

$$Y_1 = I_{(0,1]}, Y_2 = I_{(0, \frac{1}{2}]}, Y_3 = I_{(\frac{1}{2}, 1]}, Y_4 = I_{(0, \frac{1}{4}]}, Y_5 = I_{(\frac{1}{4}, \frac{1}{2}]}, Y_6 = I_{(\frac{1}{2}, \frac{3}{4}]}, Y_6 = I_{(\frac{3}{4}, 1]}, \dots$$

Dann gilt

$$P[|Y_n| > \varepsilon] = P[Y_n = 1] \rightarrow 0 \quad \text{für alle } \varepsilon > 0,$$

also konvergiert  $Y_n$  stochastisch gegen 0, obwohl

$$\limsup Y_n(\omega) = 1 \quad \text{für alle } \omega \in \Omega \text{ gilt.}$$

□

Hier ist ein weiteres Beispiel, das den Unterschied zwischen stochastischer und fast sicherer Konvergenz zeigt:

**Beispiel.** Sind  $T_1, T_2, \dots$  unter  $P$  unabhängige  $\text{Exp}(1)$ -verteilte Zufallsvariablen, dann konvergiert  $T_n/\log n$   $P$ -stochastisch gegen 0, denn

$$P \left[ \left| \frac{T_n}{\log n} \right| \geq \varepsilon \right] = P[T_n \geq \varepsilon \cdot \log n] = n^{-\varepsilon} \xrightarrow{n \rightarrow \infty} 0$$

für alle  $\varepsilon > 0$ . Andererseits gilt nach (5.1.6) aber

$$\limsup_{n \rightarrow \infty} \frac{T_n}{\log n} = 1 \quad P\text{-fast sicher,}$$

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

$$\sum_{n=1}^{\infty} P[|Y_n - Y| \geq \varepsilon] < \infty \quad \text{für alle } \varepsilon > 0.$$

**Lemma 7.2.** *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:

$$P[\limsup |Y_n| \leq \varepsilon] \geq P[|Y_n| \geq \varepsilon \text{ nur endlich oft}] = 1.$$

Es folgt

$$P[\limsup |Y_n| \neq 0] = P \left[ \bigcup_{\varepsilon \in \mathbb{Q}_+} \{\limsup |Y_n| > \varepsilon\} \right] = 0.$$

□

Ähnlich zeigt man:

**Lemma 7.3.** *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

$$P \left[ |Y_{n_k}| \geq \frac{1}{k} \right] \leq \frac{1}{k^2}.$$

Nach dem Lemma von Borel-Cantelli folgt

$$P \left[ |Y_{n_k}| \geq \frac{1}{k} \text{ nur endlich oft} \right] = 1,$$

also  $Y_{n_k} \rightarrow 0$   $P$ -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.

### Die Markov-Čebyšev-Ungleichung

Sei  $X : \Omega \rightarrow \overline{\mathbb{R}}$  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 7.4 (Allgemeine Markov-Ungleichung).** Sei  $h : [0, \infty] \rightarrow [0, \infty]$  monoton wachsend und Borel-messbar. Dann gilt

$$P[|X| \geq c] \leq \frac{E[h(|X|) ; |X| \geq c]}{h(c)} \leq \frac{E[h(|X|)]}{h(c)} \quad \text{für alle } c > 0 \text{ mit } h(c) \neq 0.$$

*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:

$$P[|X| \geq c] \leq \frac{E[|X|]}{c} \quad \text{für alle } c > 0.$$

Insbesondere gilt für eine Zufallsvariable  $X$  mit  $E[|X|] = 0$ :

$$P[|X| \geq c] = 0 \quad \text{für alle } c > 0,$$

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:

$$P[|Y - E[Y]| \geq c] \leq \frac{E[(Y - E[Y])^2]}{c^2} = \frac{\text{Var}[Y]}{c^2} \quad \text{für alle } c > 0.$$

Diese Ungleichung haben wir bereits in Abschnitt 3.2 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 8.2 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 7.5 ( $\mathcal{L}^p$ -Konvergenz impliziert stochastische Konvergenz).** Für  $1 \leq p < \infty$  gilt:

$$E[|X_n - X|^p] \rightarrow 0 \quad \Rightarrow \quad P[|X_n - X| > \varepsilon] \rightarrow 0 \quad \text{für alle } \varepsilon > 0.$$

*Beweis.* Nach der Markovungleichung mit  $h(x) = x^p$  gilt:

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

□

**Bemerkung.** Aus stochastischer Konvergenz folgt im Allgemeinen nicht  $\mathcal{L}^p$ -Konvergenz (Übung). Es gilt aber: Konvergiert  $X_n \rightarrow X$  stochastisch, und ist die Folge der Zufallsvariablen  $|X_n|^p$  ( $n \in \mathbb{N}$ ) **gleichmäßig integrierbar**, d.h.

$$\sup_{n \in \mathbb{N}} E[|X_n|^p; |X_n| \geq c] \rightarrow 0 \quad \text{für } c \rightarrow \infty,$$

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:

## 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:

$$E[\ell(X)] = E[aX + b] = aE[X] + b = \ell(E[X]) \quad (7.1.1)$$

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 7.6 (Jensensche Ungleichung).** *Ist  $P$  eine Wahrscheinlichkeitsverteilung,  $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$  eine reellwertige Zufallsvariable, und  $h : \mathbb{R} \rightarrow \mathbb{R}$  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 (7.1.1)) 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  $h : \mathbb{R} \rightarrow \mathbb{R}$  ist genau dann konvex, wenn

$$h(\lambda x + (1 - \lambda)y) \leq \lambda h(x) + (1 - \lambda)h(y) \quad \text{für alle } \lambda \in [0, 1] \text{ und } x, y \in \mathbb{R}$$

gilt, d.h. wenn alle Sekanten oberhalb des Funktionsgraphen liegen.

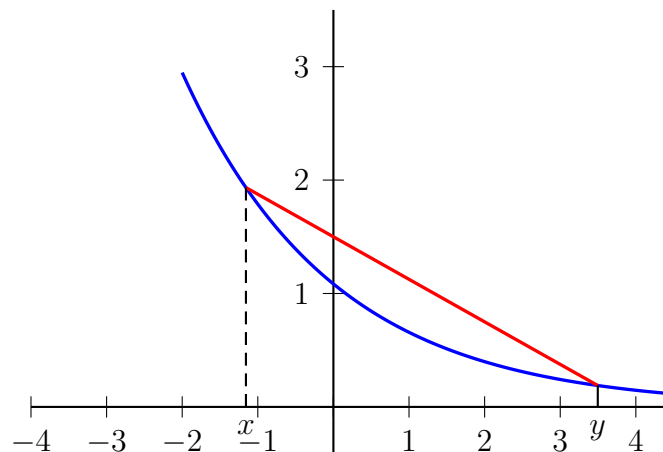


Abbildung 7.1: Sekante an konvexer Funktion

Hieraus folgt, daß jede konvexe Funktion stetig ist: Für  $a < b < x < y < c < d$  gilt nämlich

$$\frac{h(b) - h(a)}{b - a} \leq \frac{h(y) - h(x)}{y - x} \leq \frac{h(d) - h(c)}{d - c}.$$

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  $\ell$  (Stützgerade) mit  $\ell(x_0) = h(x_0)$  und  $\ell \leq h$ , siehe die Analysis Vorlesung oder [A. KLENKE: „WAHRSCHEINLICHKEITSTHEORIE“, Abschnitt 7.2].

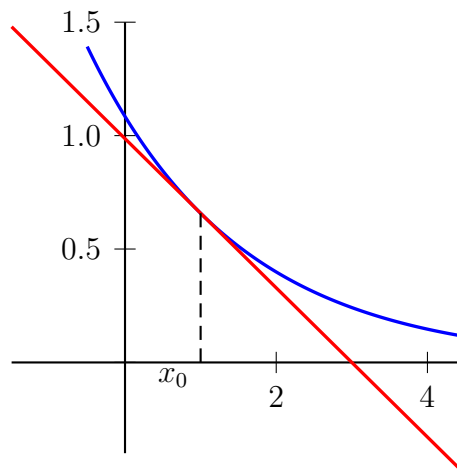


Abbildung 7.2: Darstellung von  $\ell(x)$  und  $h(x)$

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  $E[h(X)^-] \leq E[\ell(X)^-] < \infty$ .  $\square$

**Korollar 7.7 ( $\mathcal{L}^q$ -Konvergenz impliziert  $\mathcal{L}^p$ -Konvergenz).** Für  $1 < p \leq q$  gilt:

$$\|X\|_p := E[|X|^p]^{\frac{1}{p}} \leq \|X\|_q.$$

Insbesondere folgt  $\mathcal{L}^p$ -Konvergenz aus  $\mathcal{L}^q$ -Konvergenz.

*Beweis.* Nach der Jensenschen Ungleichung gilt

$$E[|X|^p]^{q/p} \leq E[|X|^q],$$

da die Funktion  $h(x) = |x|^{q/p}$  für  $q \geq p$  konvex ist.  $\square$

Nach dem Korollar gilt für  $p \leq q$ :

$$\mathcal{L}^p(\Omega, \mathcal{A}, P) \supseteq \mathcal{L}^q(\Omega, \mathcal{A}, P),$$

und

$$X_n \rightarrow X \text{ in } \mathcal{L}^q \Rightarrow X_n \rightarrow X \text{ in } \mathcal{L}^p.$$

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**:

$$E[|XY|] \leq \|X\|_p \cdot \|Y\|_q \quad \text{für } p, q \in [1, \infty] \text{ mit } \frac{1}{p} + \frac{1}{q} = 1.$$

## 7.2 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  $X_1, X_2, \dots \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$  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

$$\text{Cov}[X_i, X_j] \leq c_{|i-j|} \quad \text{für alle } i, j \in \mathbb{N}. \quad (7.2.1)$$

Die Bedingung (A) ist insbesondere erfüllt, wenn die *Korrelationen exponentiell abfallen*, d.h. wenn

$$|\text{Cov}[X_i, X_j]| \leq c \cdot \alpha^{|i-j|}$$

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)  $\text{Cov}[X_i, X_j] = 0$  für alle  $i, j \in \mathbb{N}$ , und

(A2)  $v := \sup_i \text{Var}[X_i] < \infty$ ,

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 3.2 ein schwaches Gesetz der großen Zahlen bewiesen.

**Wichtig:** Es wird **keine Unabhängigkeit vorausgesetzt!**

Sei nun

$$S_n = X_1 + \dots + X_n$$

die Summe der ersten  $n$  Zufallsvariablen.

### Das schwache Gesetz der großen Zahlen

Den Beweis des schwachen Gesetzes der großen Zahlen aus Abschnitt 3.2 können wir auf den hier betrachteten allgemeinen Fall erweitern:

**Satz 7.8** (Schwaches Gesetz der großen Zahlen,  $\mathcal{L}^2$ -Version). *Unter der Voraussetzung (A) gilt für alle  $n \in \mathbb{N}$  und  $\varepsilon > 0$ :*

$$E \left[ \left( \frac{S_n}{n} - \frac{E[S_n]}{n} \right)^2 \right] \leq \frac{v}{n}, \quad \text{und} \quad (7.2.2)$$

$$P \left[ \left| \frac{S_n}{n} - \frac{E[S_n]}{n} \right| \geq \varepsilon \right] \leq \frac{v}{\varepsilon^2 n} \quad (7.2.3)$$

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

$$\frac{S_n}{n} \rightarrow m \quad \text{in } \mathcal{L}^2(\Omega, \mathcal{A}, P) \text{ und } P\text{-stochastisch.}$$

*Beweis.* Unter Verwendung der Voraussetzung an die Kovarianzen erhalten wir

$$\begin{aligned} E \left[ \left( \frac{S_n}{n} - \frac{E[S_n]}{n} \right)^2 \right] &= \text{Var} \left[ \frac{S_n}{n} \right] = \frac{1}{n^2} \text{Var}[S_n] \\ &= \frac{1}{n^2} \sum_{i,j=1}^n \text{Cov}[X_i, X_j] \leq \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n c_{|i-j|} \\ &\leq \frac{1}{n^2} \sum_{i=1}^n \sum_{k=-\infty}^{\infty} c_{|k|} = \frac{v}{n} \end{aligned}$$

Die zweite Behauptung folgt daraus durch Anwenden der Čebyšev-Ungleichung.  $\square$

**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*

$$L^2(\Omega, \mathcal{A}, P) = \mathcal{L}^2(\Omega, \mathcal{A}, P) / \sim \quad \text{konvergiert gegen } 0.$$

Unkorreliertheit der  $X_i$  bedeutet gerade, dass die Zufallsvariablen

$$Y_i := X_i - E[X_i]$$

orthogonal in  $L^2$  sind - beschränkte Varianzen der  $X_i$  ist gleichbedeutend mit der Beschränktheit der  $L^2$  Normen der  $Y_i$ . Es gilt

$$S_n - E[S_n] = \sum_{i=1}^n Y_i,$$

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 (7.2.1) 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 (7.2.2) bzw. (7.2.3).
- (3). Eine für große  $n$  deutlich bessere Abschätzung des Fehlers in (7.2.3) (mit exponentiellem Abfall in  $n$ ) erhält man bei Unabhängigkeit und exponentieller Integrierbarkeit der  $X_i$  mithilfe der *exponentiellen Ungleichung*, siehe Satz 8.3 unten.

## Das starke Gesetz für quadratintegrierbare Zufallsvariablen

Unter derselben Voraussetzung wie in Satz 7.8 gilt sogar  $P$ -fast sichere Konvergenz:

**Satz 7.9 (Starkes Gesetz großer Zahlen,  $\mathcal{L}^2$ -Version).** *Unter der Voraussetzung (A) konvergiert*

$$\frac{S_n(\omega)}{n} - \frac{E[S_n]}{n} \longrightarrow 0$$

für  $P$ -fast alle  $\omega \in \Omega$ . Insbesondere gilt

$$\frac{S_n}{n} \longrightarrow m \quad P\text{-fast sicher,}$$

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  $\tilde{X}_i := X_i - E[X_i]$ ; diese sind wieder unkorreliert mit beschränkten Varianzen. Zu zeigen ist dann:

$$\frac{S_n}{n} \rightarrow 0 \quad P\text{-fast sicher.}$$

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:

$$P \left[ \left| \frac{S_{k^2}}{k^2} \right| \geq \varepsilon \right] \leq \frac{1}{\varepsilon^2} \text{Var} \left[ \frac{S_{k^2}}{k^2} \right] \leq \frac{1}{\varepsilon^2 k^2} \sup_i \text{Var}[X_i].$$

Da die Varianzen beschränkt sind, ist der gesamte Ausdruck durch die Summanden einer summierbaren Reihe beschränkt. Somit ergibt sich nach Borel-Cantelli:

$$\frac{S_{k^2}(\omega)}{k^2} \rightarrow 0$$

für alle  $\omega$  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

$$D_k := \max_{k^2 \leq l < (k+1)^2} |S_l - S_{k^2}|.$$

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  $k^2 \leq l \leq k^2 + 2k$ :

$$\begin{aligned} P[|S_l - S_{k^2}| > \varepsilon k^2] &\leq \frac{1}{\varepsilon^2 k^4} \operatorname{Var}[S_l - S_{k^2}] \leq \frac{1}{\varepsilon^2 k^4} \operatorname{Var} \left[ \sum_{i=k^2+1}^l X_i \right] \\ &\leq \frac{l - k^2}{\varepsilon^2 k^4} \sup_i \operatorname{Var}[X_i] \leq \operatorname{const} \cdot \frac{k}{k^4}. \end{aligned}$$

Nach Lemma 7.2 folgt daher

$$\frac{D_k(\omega)}{k^2} \rightarrow 0$$

für alle  $\omega$  außerhalb einer Nullmenge  $N_2$ .

- (3). Zu gegebenem  $n$  wählen wir nun  $k = k(n)$  mit  $k^2 \leq n < (k+1)^2$ . Durch Kombination der ersten beiden Schritte erhalten wir:

$$\left| \frac{S_n(\omega)}{n} \right| \leq \frac{|S_{k^2}(\omega)| + D_k(\omega)}{n} \leq \left| \frac{S_{k^2}(\omega)}{k^2} \right| + \frac{D_k(\omega)}{k^2} \rightarrow 0 \quad \text{für } n \rightarrow \infty$$

für alle  $\omega$  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  $S_n = X_1 + \dots + X_n$  ein Random Walk im  $\mathbb{R}^d$  mit unabhängigen identisch verteilten Inkrementen  $X_i$  mit Verteilung  $\mu$ . 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  $S_n$ ):

$$\frac{S_n(\omega)}{n} \rightarrow m \quad \text{für } P\text{-fast alle } \omega,$$

wobei  $m = \int_{\mathbb{R}^d} x \mu(dx)$  der Schwerpunkt der Inkrementverteilung ist. Insbesondere gilt für  $m \neq 0$ :

$$S_n \sim m \cdot n \quad \text{für } n \rightarrow \infty \quad P\text{-fast sicher,}$$

d.h.  $S_n$  wächst linear mit Geschwindigkeit  $m$ . Im Fall  $m = 0$  gilt dagegen

$$\frac{S_n(\omega)}{n} \rightarrow 0 \quad P\text{-fast sicher,}$$

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*:

$$\begin{aligned}\limsup_{n \rightarrow \infty} \frac{S_n(\omega)}{\sqrt{n \log \log n}} &= +1 && P\text{-fast sicher,} \\ \liminf_{n \rightarrow \infty} \frac{S_n(\omega)}{\sqrt{n \log \log n}} &= -1 && P\text{-fast sicher,}\end{aligned}$$

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:

$$\frac{1}{n} \log X_n = \frac{1}{n} \sum_{i=1}^n \log Y_i \rightarrow E[\log Y_1] =: \alpha \quad P\text{-f.s.}$$

Also existiert für  $\varepsilon > 0$  ein  $N(\omega)$  mit  $N(\omega) < \infty$   $P$ -fast sicher,

$$X_n(\omega) \leq e^{(\alpha+\varepsilon)n} \quad \text{und} \quad X_n(\omega) \geq e^{(\alpha-\varepsilon)n} \quad \text{für alle } n \geq N(\omega).$$

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  $\alpha$  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  $\alpha$  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:

$$E[X_n] \nearrow \infty, \quad \text{obwohl } X_n \rightarrow 0 \quad P\text{-fast sicher.}$$

## Von $\mathcal{L}^2$ nach $\mathcal{L}^1$ mit Unabhängigkeit

Sind Zufallsvariablen  $X, Y : \Omega \rightarrow S$  unabhängig, so sind  $f(X)$  und  $g(Y)$  für beliebige beschränkte oder nichtnegative Funktionen  $f, g : S \rightarrow \mathbb{R}$  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 7.10 (Kolmogorovs Gesetz der großen Zahlen).** Seien  $X_1, X_2, \dots \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$  paarweise unabhängig und identisch verteilt mit  $E[X_i] = m$ . Dann gilt:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i = m \quad P\text{-fast sicher.}$$

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

$$\frac{1}{n} \sum_{i=1}^n X_i(\omega) \longrightarrow m = \int x \mu_{X_i}(dx) \quad P\text{-fast sicher,}$$

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.

von Satz 7.10. 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

$$\begin{aligned} \sum_{i=1}^{\infty} P[Y_i \neq X_i] &= \sum_{i=1}^{\infty} P[X_i > i] \\ &= \sum_{i=1}^{\infty} P[X_1 > i] \quad (X_i \text{ identisch verteilt}) \\ &\leq \int_0^{\infty} P[X_1 > x] dx \quad (P[X_1 > x] \text{ monoton fallend}) \\ &= E[X_1] < \infty. \end{aligned}$$

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

$$E[Y_i] = E[X_i; X_i \leq i] = E[X_1 \cdot I_{\{X_1 \leq i\}}] \longrightarrow E[X_1] = m, \quad \text{für } i \rightarrow \infty,$$

also auch

$$E\left[\frac{S_n}{n}\right] = \frac{1}{n} \sum_{i=1}^n E[Y_i] \longrightarrow m \quad \text{für } n \rightarrow \infty.$$

(4). *P-fast sichere Konvergenz von  $\frac{S_n}{n}$  entlang der Teilfolgen  $k_n = \lfloor \alpha^n \rfloor$ ,  $\alpha > 1$ .*

*Vorbemerkung:* Es gilt

$$\sum_{n \geq m} \frac{1}{k_n^2} = \frac{1}{\lfloor \alpha^m \rfloor^2} + \frac{1}{\lfloor \alpha^{m+1} \rfloor^2} + \dots \leq \frac{\text{const.}}{\lfloor \alpha^m \rfloor^2} = \frac{\text{const.}}{k_m^2}$$

mit einer von  $m$  unabhängigen Konstanten.

*Behauptung:*

$$\frac{S_{k_n}}{k_n} \longrightarrow \lim_{n \rightarrow \infty} E\left[\frac{S_{k_n}}{k_n}\right] = m \quad \text{P-fast sicher.}$$

*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

$$\sum_{n=1}^{\infty} \text{Var}\left[\frac{S_{k_n}}{k_n}\right] < \infty$$

gilt. Wegen

$$\text{Var}[Y_i] \leq E[Y_i^2] = E[X_i^2; X_i \leq i] = E[X_1^2; X_1 \leq i]$$



erhalten wir mithilfe der Vorbemerkung

$$\begin{aligned}
 \sum_{n=1}^{\infty} \operatorname{Var} \left[ \frac{S_{k_n}}{k_n} \right] &= \sum_{n=1}^{\infty} \frac{1}{k_n^2} \cdot \sum_{i=1}^{k_n} \operatorname{Var}[Y_i] \\
 &\leq \sum_{i=1}^{\infty} E[X_1^2; X_1 \leq i] \cdot \sum_{n: k_n \geq i} \frac{1}{k_n^2} \\
 &\leq \operatorname{const.} \cdot \sum_{i=1}^{\infty} E[X_1^2; X_1 \leq i] \cdot \frac{1}{i^2} \\
 &\leq \operatorname{const.} \cdot \sum_{i=1}^{\infty} \sum_{j=1}^i j^2 \cdot P[X_1 \in (j-1, j]] \cdot \frac{1}{i^2} \\
 &= \operatorname{const.} \cdot \sum_{j=1}^{\infty} j^2 \cdot P[X_1 \in (j-1, j]] \cdot \sum_{i=j}^{\infty} \frac{1}{i^2} \\
 &\leq \operatorname{const.} \cdot \sum_{j=1}^{\infty} j \cdot P[X_1 \in (j-1, j]] \\
 &= \operatorname{const.} \cdot E \left[ \sum_{j=1}^{\infty} j \cdot I_{\{X_1 \in (j-1, j]\}} \right] \\
 &\leq \operatorname{const.} \cdot E[X_1 + 1] < \infty.
 \end{aligned}$$

(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  $n \rightarrow \infty$  erhalten wir wegen  $\frac{k_{n+1}}{k_n} \rightarrow \alpha$  und  $\frac{S_{k_n}(\omega)}{k_n} \rightarrow m$ :

$$\frac{m}{\alpha} \leq \liminf \frac{S_l(\omega)}{l} \leq \limsup \frac{S_l(\omega)}{l} \leq \alpha m$$

für alle  $\omega$  außerhalb einer von  $\alpha$  abhängenden Nullmenge  $N_\alpha$ . Für  $\omega$  außerhalb der Nullmenge  $\bigcup_{\substack{\alpha > 1 \\ \alpha \in \mathbb{Q}}} N_\alpha$  folgt somit:

$$\lim_{l \rightarrow \infty} \frac{S_l(\omega)}{l} = m.$$

□

**Korollar 7.11 (Gesetz der großen Zahlen ohne Integrierbarkeit).** Seien  $X_1, X_2, \dots$  paarweise unabhängige, identisch verteilte, nicht-negative Zufallsvariablen. Dann gilt:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \cdot \sum_{i=1}^n X_i(\omega) = E[X_1] \in [0, \infty] \quad P\text{-fast sicher.}$$

*Beweis.* Nach Satz 7.10 gilt die Aussage im Fall  $E[X_1] < \infty$ . Für  $E[X_1] = \infty$  erhalten wir für  $k \in \mathbb{N}$ :

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i \geq \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (X_i \wedge k) = E[X_1 \wedge k] \quad P\text{-fast sicher.}$$

Für  $k \rightarrow \infty$  folgt dann mit monotoner Konvergenz

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i \geq E[X_1] = \infty,$$

und damit die Behauptung. □

## 7.3 Empirische Verteilungen

### Schätzen von Kenngrößen einer unbekanntem Verteilung

Angenommen, wir haben eine Stichprobe aus reellen Beobachtungswerten  $X_1, X_2, \dots, X_n$  gegeben, und möchten die zugrundeliegende Wahrscheinlichkeitsverteilung  $\mu$  auf  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  möglichst weitgehend rekonstruieren. Im einfachsten Modell interpretieren wir die Beobachtungswerte als Realisierungen unabhängiger Zufallsvariablen  $X_1, X_2, \dots$  mit Verteilung  $\mu$ .

(1). SCHÄTZEN DES ERWARTUNGSWERTES: Sei  $\int |x| \mu(dx) < \infty$ . Um den Erwartungswert

$$m = \int x \mu(dx)$$

zu schätzen, verwenden wir das **empirische Mittel**

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i.$$

Das empirische Mittel ist ein *erwartungstreuer Schätzer* für  $m$ , d.h.  $\bar{X}_n$  ist eine Funktion von den Beobachtungswerten  $X_1, \dots, X_n$  mit  $E[\bar{X}_n] = m$ . Obere Schranken für den Schätzfehler  $P[|\bar{X}_n - m| > \varepsilon], \varepsilon > 0$ , erhält man z.B. mithilfe der Čebyšev- oder der exponentiellen Markov-Ungleichung. Für  $n \rightarrow \infty$  gilt nach dem Gesetz der großen Zahlen

$$\bar{X}_n \longrightarrow m \quad P\text{-fast sicher,}$$

d.h.  $\bar{X}_n$  ist eine *konsistente* Folge von Schätzern für  $m$ .

(2). SCHÄTZEN DER VARIANZ: Um die Varianz

$$v = \int (x - m)^2 \mu(dx)$$

der zugrundeliegenden Verteilung zu schätzen, verwendet man meistens die **renormierte Stichprobenvarianz**

$$\tilde{V}_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Der Vorfaktor  $\frac{1}{n-1}$  (statt  $\frac{1}{n}$ ) gewährleistet unter anderem, dass  $\tilde{V}_n$  ein *erwartungstreuer* Schätzer für  $v$  ist, denn aus

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - m)^2 - (\bar{X}_n - m)^2 & (7.3.1) \\ \text{Stichprobenvarianz} &= \text{MSE} - \text{Stichprobenbias}^2 \end{aligned}$$

folgt

$$E \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right] = \frac{1}{n} \sum_{i=1}^n \text{Var}[X_i] - \text{Var}[\bar{X}_n] = \frac{n-1}{n} v,$$

also  $E[\tilde{V}_n] = v$ .

Um zu zeigen, dass  $\tilde{V}_n$  eine konsistente Folge von Schätzern für  $v$  ist, können wir erneut das Gesetz der großen Zahlen anwenden. Da die Zufallsvariablen  $X_i - \bar{X}_n$ ,  $1 \leq i \leq n$ , selbst nicht unabhängig sind, verwenden wir dazu die Zerlegung (7.3.1). Nach dem starken Gesetz der großen Zahlen für nichtnegative Zufallsvariablen erhalten wir

$$\frac{n-1}{n} \tilde{V}_n = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2 - (\bar{X}_n - m)^2 \longrightarrow v \quad P\text{-fast sicher,}$$

also auch  $\tilde{V}_n \rightarrow v$   $P$ -fast sicher.

(3). SCHÄTZEN VON INTEGRALEN: Allgemeiner können wir für jede Funktion  $f \in \mathcal{L}^1(S, \mathcal{S}, \mu)$  das Integral

$$\theta = \int f d\mu$$

erwartungstreu durch die **empirischen Mittelwerte**

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n f(X_i)$$

schätzen. Dies haben wir schon in Kapitel 3 für Monte Carlo Verfahren verwendet. Da die Zufallsvariablen  $f(X_i)$  wieder unabhängig und identisch verteilt sind mit Erwartungswert  $\theta$ , gilt nach dem starken Gesetz der großen Zahlen:

$$\hat{\theta}_n \longrightarrow \theta \quad P\text{-fast sicher.} \quad (7.3.2)$$

(4). **SCHÄTZEN DER VERTEILUNG:** Die gesamte Verteilung  $\mu$  können wir durch die **empirische Verteilung**

$$\hat{\mu}_n(\omega) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i(\omega)}$$

der Zufallsstichprobe schätzen.  $\hat{\mu}_n$  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 (7.3.2) ergibt sich die folgende Approximationseigenschaft der empirischen Verteilungen:

$$\int f d\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n f(X_i) \xrightarrow{n \rightarrow \infty} \int f d\mu \quad (7.3.3)$$

$P$ -fast sicher für alle  $f \in \mathcal{L}^1(S, \mathcal{S}, \mu)$ .

## Konvergenz der empirischen Verteilungsfunktionen

Für die *empirischen Verteilungsfunktionen*

$$F_n(c) = \hat{\mu}_n[(-\infty, c]] = \frac{1}{n} |\{1 \leq i \leq n : X_i \leq c\}|$$

von unabhängigen, identisch verteilten, reellwertigen Zufallsvariablen  $X_1, X_2, \dots$  mit Verteilungsfunktion  $F$  ergibt sich wegen  $F_n(c) = \int I_{(-\infty, c]} d\hat{\mu}_n$ :

$$\lim_{n \rightarrow \infty} F_n(c) = F(c) \quad P\text{-fast sicher für alle } c \in \mathbb{R}. \quad (7.3.4)$$

Diese Aussage kann man noch etwas verschärfen:

**Satz 7.12 (Glivenko-Cantelli).** *Sind  $X_1, X_2, \dots$  unabhängig und identisch verteilt mit Verteilungsfunktion  $F$ , dann gilt für die empirischen Verteilungsfunktionen  $F_n$ :*

$$\sup_{c \in \mathbb{R}} |F_n(c) - F(c)| \longrightarrow 0 \quad P\text{-fast sicher.} \quad (7.3.5)$$

*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

$$-\infty = c_0 < c_1 < c_2 < \dots < c_k = \infty \quad \text{mit } F(c_i) - F(c_{i-1}) \leq \frac{\varepsilon}{2}$$

für alle  $1 \leq i \leq k$ . Da  $F_n$  nach 7.3.4 mit Wahrscheinlichkeit 1 punktweise gegen  $F$  konvergiert, existiert zudem ein  $n_0 \in \mathbb{N}$  mit

$$\max_{0 \leq i \leq n} |F_n(c_i) - F(c_i)| < \frac{\varepsilon}{2} \quad \text{für alle } n \geq n_0.$$

Wegen der Monotonie der Verteilungsfunktionen folgt dann

$$F_n(c) - F(c) \leq F_n(c_i) - F(c_{i-1}) \leq \frac{\varepsilon}{2} + F_n(c_i) - F(c_i) < \varepsilon,$$

und entsprechend

$$F(c) - F_n(c) \leq F(c_i) - F_n(c_{i-1}) \leq \frac{\varepsilon}{2} + F(c_i) - F_n(c_i) < \varepsilon,$$

für alle  $n \geq n_0, c \in \mathbb{R}$ , und  $1 \leq i \leq k$  mit  $c_{i-1} \leq c \leq c_i$ . Also gilt auch

$$\sup_{c \in \mathbb{R}} |F_n(c) - F(c)| < \varepsilon \quad \text{für alle } n \geq n_0.$$

□

**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  $x_1, \dots, x_n$  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

$$\left( \Phi^{-1} \left( \frac{k - \frac{1}{2}}{n} \right), x_{(k)} \right), \quad k = 1, 2, \dots, n,$$

wobei  $\Phi$  die Verteilungsfunktion der Standardnormalverteilung ist, und

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

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  $\sigma$ , dann liegen die Punkte für große  $n$  näherungsweise auf einer Geraden mit Steigung  $\sigma$  und Achsenabschnitt  $m$ , da für die Verteilungsfunktion und die Quantile der theoretischen Verteilung dann

$$F(c) = P[X \leq c] = P[\sigma Z + m \leq c] = P\left[Z \leq \frac{c - m}{\sigma}\right] = \Phi\left(\frac{c - m}{\sigma}\right),$$

bzw.

$$F^{-1}(u) = m + \sigma\Phi^{-1}(u)$$

gilt. Die folgende Grafik zeigt QQ-Plots bzgl. der Normalverteilung für verschiedene Datensätze.

## Histogramme und Multinomialverteilung

Die empirische Verteilung  $\hat{\mu}_n(\omega) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i(\omega)}$  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  $x_1, \dots, x_n$ , die in einer endlichen Menge  $S$  liegen, ist der Vektor

$$\vec{h} = (h_a)_{a \in S}, \quad h_a = |\{1 \leq i \leq n | x_i = a\}|,$$

der Häufigkeiten der möglichen Werte  $a \in S$  unter  $x_1, \dots, x_n$ . Graphisch stellt man ein Histogramm durch ein Balkendiagramm dar:

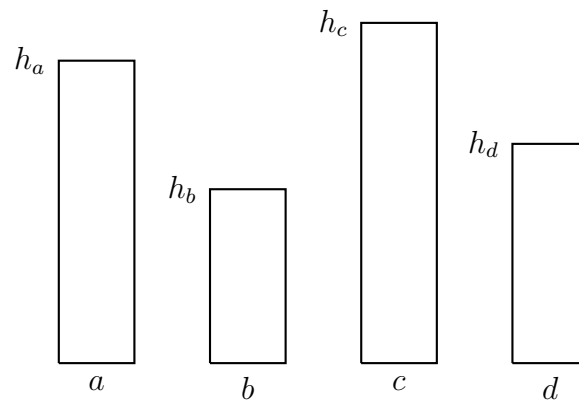


Abbildung 7.3: Histogramm der Klassen  $a, b, c$  und  $d$  mit den jeweiligen Häufigkeiten  $h_a, h_b, h_c$  und  $h_d$

Der Raum  $\text{Hist}(n)$  aller möglichen Histogramme von  $n$  Beobachtungswerten ist eine Teilmenge von  $\{0, 1, \dots, n\}^S$ :

$$\text{Hist}(n) = \left\{ \vec{h} = (h_a)_{a \in S} \mid h_a \in \mathbb{Z}_+, \sum_{a \in S} h_a = n \right\} \subseteq \{0, 1, \dots, n\}^S.$$

Sie nun  $\mu$  eine Wahrscheinlichkeitsverteilung auf der endlichen Menge  $S$ . Wir wollen die Verteilung des Histogrammvektors bestimmen, wenn die Beobachtungswerte unabhängige Stichproben von der Verteilung  $\mu$  sind. Wir betrachten also unabhängige Zufallsvariablen  $X_1, \dots, X_n$  auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  mit Verteilung  $\mu$  und die Häufigkeiten

$$H_a(\omega) := |\{1 \leq i \leq n : X_i(\omega) = a\}|$$

der möglichen Werte  $a \in S$ . Die Zufallsvariable  $H_a$  ist  $\text{Bin}(n, p)$ -verteilt mit  $p = \mu[\{a\}]$ . Wir berechnen nun die **gemeinsame Verteilung** aller dieser Häufigkeiten, d.h. die Verteilung  $\mu_H$  des Zufallsvektors

$$H = (H_a)_{a \in S} : \Omega \longrightarrow \text{Hist}(n)$$

mit Werten im Raum der Histogramme. Dazu verwenden wir die Unabhängigkeit der  $X_i$ . Mit  $I = \{1, \dots, n\}$  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  $I = \{0, 1, \dots, n\}$  in Teilmengen  $i_a, a \in S$ , mit jeweils  $k_a$  Elementen, und der **Multinomialkoeffizient**

$$\binom{n}{\vec{k}} := \frac{n!}{\prod_{a \in S} k_a!}, \quad k_a \in \{0, 1, \dots, n\} \text{ mit } \sum_{a \in S} k_a = n,$$

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.

## 7.4 Entropie

Wir definieren nun die Entropie einer diskreten Wahrscheinlichkeitsverteilung. Mithilfe des Gesetzes der großen Zahlen können wir eine statistische Interpretation dieser Größe geben, aus der sich insbesondere der Quellenkodierungssatz von Shannon ergibt.



## Definition und Eigenschaften

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

$$u(x) \leq 0 \quad \text{für alle } x \in [0, 1], \quad (7.4.1)$$

$$u(x) \geq x - 1 \quad \text{für alle } x \geq 0, \quad (7.4.2)$$

und absolutem Minimum  $u(1/e) = -1/e$ .

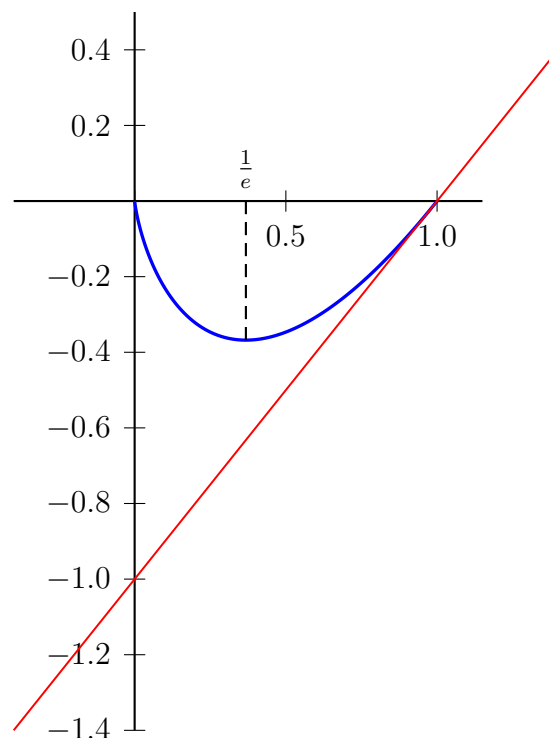


Abbildung 7.4: Graph der Funktion  $u(x)$  (blau) und ihrer unteren Schranke  $x - 1$  (rot)

Sei nun  $S$  eine abzählbare Menge, und  $\mu = (\mu(x))_{x \in S}$  eine Wahrscheinlichkeitsverteilung auf  $S$ .

**Definition.** Die Größe

$$H(\mu) := - \sum_{\substack{x \in S \\ \mu(x) \neq 0}} \mu(x) \log \mu(x) = - \sum_{x \in S} u(\mu(x)) \in [0, \infty]$$

heißt **Entropie** der Wahrscheinlichkeitsverteilung  $\mu$ .

Anschaulich können wir  $-\log \mu(x)$  interpretieren als Maß für die »Überraschung« bzw. den »Informationsgewinn«, falls eine Stichprobe von der Verteilung  $\mu$  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  $\mu$ . Eine wichtige Eigenschaft der Entropie, die auch die Wahl des Logarithmus erklärt, ist:

**Satz 7.13 (Faktorisierungseigenschaft).** Für beliebige diskrete Wahrscheinlichkeitsverteilungen  $\mu$  und  $\nu$  gilt:

$$H(\mu \otimes \nu) = H(\mu) + H(\nu).$$

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:

$$\begin{aligned} H(\mu \otimes \nu) &= \sum_{\substack{x,y \\ \mu(x)\nu(y) \neq 0}} \mu(x)\nu(y) \log(\mu(x)\nu(y)) \\ &= - \sum_{x:\mu(x) \neq 0} \mu(x) \log(\mu(x)) - \sum_{y:\nu(y) \neq 0} \nu(y) \log(\nu(y)) \\ &= H(\mu) + H(\nu). \end{aligned}$$

□

Wir bestimmen nun auf einer gegebenen abzählbaren Menge  $S$  die Wahrscheinlichkeitsverteilungen mit minimaler bzw. maximaler Entropie.

### Extrema der Entropie:

- (1). **Entropieminima:** Nach (7.4.1) ist die Entropie stets nicht-negativ, und es gilt:

$$H(\mu) = 0 \iff \mu(x) \in \{0, 1\} \quad \forall x \in S \iff \mu \text{ ist ein Diracmaß.}$$

Die Diracmaße sind also die Entropieminima. Ist das Zufallsexperiment deterministisch, d.h.  $\mu$  ein Diracmaß, dann tritt bei Ziehen einer Stichprobe von  $\mu$  keine Überraschung bzw. kein Informationszuwachs auf.

- (2). **Entropiemaximum:** Ist  $S$  endlich, dann gilt für alle Wahrscheinlichkeitsverteilungen  $\mu$  auf  $S$ :

$$H(\mu) \leq -\log \left( \frac{1}{|S|} \right) = H(\mathcal{U}_S),$$

wobei  $\mathcal{U}_S$  die Gleichverteilung auf  $S$  ist. Nach der Jensenschen Ungleichung gilt nämlich

$$\begin{aligned} -\sum_{x \in S} u(\mu(x)) &= -|S| \cdot \int u(\mu(x)) \mathcal{U}_S(dx) \\ &\leq -|S| \cdot u\left(\int \mu(x) \mathcal{U}_S(dx)\right) \\ &= -|S| \cdot u\left(\frac{1}{|S|}\right) = -\log \frac{1}{|S|} \end{aligned}$$

mit Gleichheit genau dann, wenn  $\mu$  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  $\mu$ .

Auf einer abzählbar unendlichen Menge existiert keine Wahrscheinlichkeitsverteilung mit maximaler Entropie.

**Beispiel (Entropie von Markovketten).** Sei  $p(x, y)$  ( $x, y \in S$ ) eine stochastische Matrix auf einer endlichen Menge  $S$ , die die Gleichverteilung  $\mathcal{U}_S$  als Gleichgewicht hat, d.h. für alle  $y \in S$  gilt:

$$\sum_{x \in S} p(x, y) = |S| \cdot \sum_{x \in S} \mathcal{U}_S(x) p(x, y) = |S| \cdot \mathcal{U}_S(y) = 1. \quad (7.4.3)$$

Beispielsweise ist  $p$  die Übergangsmatrix eines 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  $\{0, 1\}^n$  („Ehrenfestmodell“).

Der folgende Satz zeigt, dass die Entropie  $H(\mu p^n)$  der Verteilung zur Zeit  $n$  einer Markovkette mit Startverteilung  $\mu$  und Übergangsmatrix  $p$  monoton wächst:

**Satz 7.14 (Zunahme der Entropie).** *Ist  $p$  eine stochastische Matrix auf  $S$  mit (7.4.3), dann gilt:*

$$H(\mu p) \geq H(\mu)$$

für jede Wahrscheinlichkeitsverteilung  $\mu$  auf  $S$ . Insbesondere ist  $n \mapsto H(\mu p^n)$  monoton wachsend.

*Beweis.* Aus der Jensenschen Ungleichung folgt:

$$\begin{aligned} -H(\mu p) &= \sum_{y \in S} u \left( \sum_{x \in S} \mu(x) p(x, y) \right) \\ &\leq \sum_{y \in S} \sum_{x \in S} u(\mu(x)) p(x, y) \\ &= \sum_{x \in S} u(\mu(x)) = -H(\mu). \end{aligned}$$

Hierbei haben wir im zweiten Schritt benutzt, dass die Funktion  $u$  konvex ist, und dass  $x \mapsto p(x, y)$  nach (7.4.3) für jedes  $y \in S$  die Gewichtsfunktion einer Wahrscheinlichkeitsverteilung ist.  $\square$

In der Interpretation der statistischen Physik geht die zeitliche Entwicklung auf makroskopischer Ebene (Thermodynamik) von einem geordneten hin zu einem ungeordneten Zustand 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 Markovmodells 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.

## Statistische Interpretation der Entropie

Sei  $\mu$  eine Wahrscheinlichkeitsverteilung auf einer abzählbaren Menge  $S$ . Die Wahrscheinlichkeit einer Folge von Ausgängen  $x_1, \dots, x_n$  bei Entnehmen einer Stichprobe aus  $n$  unabhängigen Zufallsgrößen mit Verteilung  $\mu$  beträgt

$$p_n(x_1, \dots, x_n) = \prod_{i=1}^n \mu(x_i).$$

Der gemittelte Informationszuwachs durch Auswertung der Werte  $x_1, \dots, x_n$  ist also

$$-\frac{1}{n} \log p_n(x_1, \dots, x_n).$$

Mithilfe des Gesetzes der großen Zahlen können wir die Asymptotik dieser Größen für  $n \rightarrow \infty$  untersuchen:

**Satz 7.15 (Shannon - Mc Millan).** *Seien  $X_1, X_2, \dots : \Omega \rightarrow S$  unter  $P$  unabhängige Zufallsvariablen mit Verteilung  $\mu$ . Dann gilt  $P$ -fast sicher*

$$-\frac{1}{n} \log p_n(X_1, \dots, X_n) \longrightarrow H(\mu) \quad \text{für } n \rightarrow \infty.$$

*Beweis.* Mit Wahrscheinlichkeit 1 gilt  $\mu(X_i) > 0$  für alle  $i$ , also nach Korollar 7.11:

$$-\frac{1}{n} \log p_n(X_1, \dots, X_n) = -\frac{1}{n} \sum_{i=1}^n \log \mu(X_i) \xrightarrow{n \rightarrow \infty} -\int \log \mu d\mu = H(\mu).$$

□

**Bemerkung (Exponentielle Skala).** Die Aussage des Satzes besagt, dass auf der „exponentiellen Skala“ fast sicher

$$p_n(X_1, \dots, X_n) \simeq e^{-nH(\mu)}$$

gilt, d.h. beide Ausdrücke sind asymptotisch äquivalent bis auf subexponentielle (also z.B. polynomiell) wachsende Faktoren. Eine asymptotische Beschreibung von Wahrscheinlichkeiten auf der exponentiellen Skala ist Gegenstand der Theorie großer Abweichungen, siehe Abschnitt Satz 8.3 und Kapitel ?? unten.

## Entropie und Kodierung

Wir betrachten nun eine Anwendung der Entropie auf 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  $X_1, X_2, \dots$  mit Verteilung  $\mu$ , 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  $n \cdot \log |S|$  Bits benötigt. Wir betrachten stattdessen „effiziente“ Kodierungen, die nur den „meisten“ Wörtern mit  $n$  Buchstaben eindeutig eine Binärfolge zuordnen.

**Definition.** Eine Folge von Mengen  $B_n \subseteq S^n$  ( $n \in \mathbb{N}$ ) heißt **wesentlich** bzgl.  $\mu$ , falls

$$P[(X_1, \dots, X_n) \in B_n] = \mu^n[B_n] \rightarrow 1 \quad \text{für } n \rightarrow \infty.$$

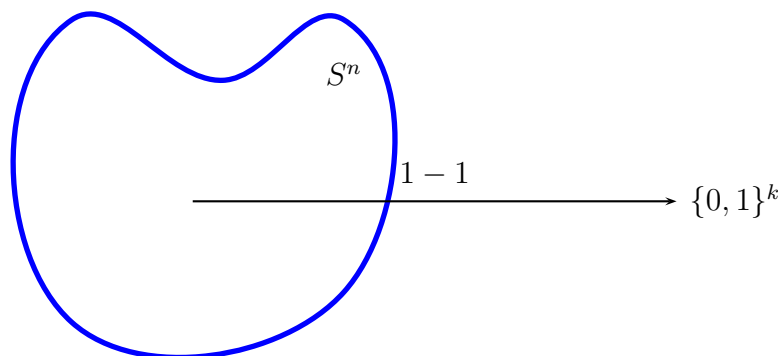


Abbildung 7.5: Perfekte Kodierung

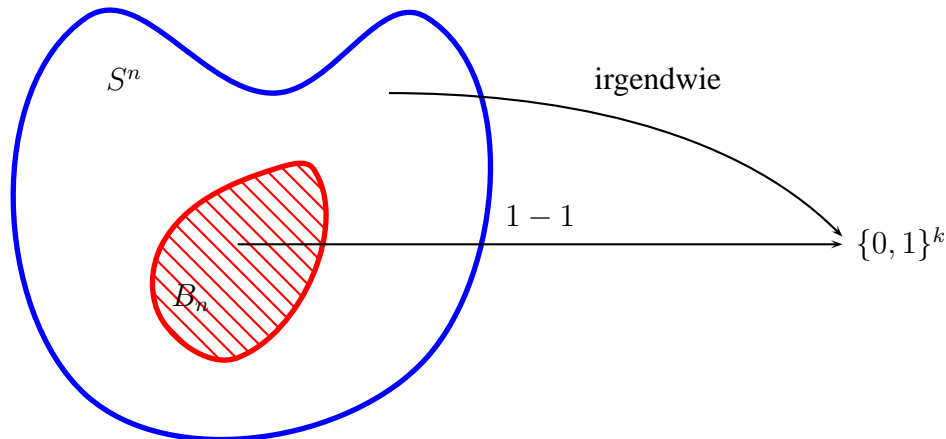


Abbildung 7.6: Effiziente Kodierung bzgl. einer Folge von wesentlichen Mengen  $B_n$ .

**Korollar 7.16 (Maßkonzentrationsatz von McMillan).** Für jedes  $\varepsilon > 0$  ist die Folge

$$B_n := \{(x_1, \dots, x_n) \in S^n \mid e^{-n(H(\mu)+\varepsilon)} \leq p_n(x_1, \dots, x_n) \leq e^{-n(H(\mu)-\varepsilon)}\}, \quad n \in \mathbb{N},$$

wesentlich bzgl.  $\mu$ , und es gilt

$$|B_n| \leq e^{n(H(\mu)+\varepsilon)} \quad \text{für alle } n \in \mathbb{N}.$$

*Beweis.* Es gilt

$$B_n = \left\{ (x_1, \dots, x_n) \in S^n \mid H(\mu) - \varepsilon \leq -\frac{1}{n} \log p_n(x_1, \dots, x_n) \leq H(\mu) + \varepsilon \right\}. \quad (7.4.4)$$

Da aus der fast sicheren Konvergenz von  $-\frac{1}{n} \log p_n(X_1, \dots, X_n)$  gegen die Entropie  $H(\mu)$  die stochastische Konvergenz folgt, ist die Folge  $B_n$  ( $n \in \mathbb{N}$ ) nach Satz 7.15 wesentlich bzgl.  $\mu$ . Zudem gilt wegen  $p_n(x_1, \dots, x_n) \geq e^{-n(H(\mu)+\varepsilon)}$  für  $(x_1, \dots, x_n) \in B_n$ :

$$1 \geq P[(X_1, \dots, X_n) \in B_n] = \sum_{x \in B_n} p_n(x_1, \dots, x_n) \geq |B_n| \cdot e^{-n(H(\mu)+\varepsilon)},$$

also  $|B_n| \leq e^{n(H(\mu)+\varepsilon)}$  □

Der Maßkonzentrationsatz zeigt, dass Folgen von wesentlichen Mengen existieren, die auf der exponentiellen Skala nicht viel schneller als  $\exp(n \cdot H(\mu))$  wachsen.

Wie groß sind wesentliche Mengen mindestens? Für  $p \in (0, 1)$  sei

$$K(n, p) = \inf \{|A_n| \mid A_n \subseteq S^n \text{ mit } P[(X_1, \dots, X_n) \in A_n] \geq p\}$$

die mindestens benötigte Anzahl von Wörtern, um den Text  $(X_1, \dots, X_n)$  mit Wahrscheinlichkeit  $\geq p$  korrekt zu erfassen. Dann ist  $\log_2 K(n, p)$  die für eine korrekte binäre Kodierung von  $(X_1, \dots, X_n)$  mit Wahrscheinlichkeit  $\geq p$  mindestens benötigte Anzahl von Bits.

**Satz 7.17 (Quellenkodierungssatz von Shannon).** Für alle  $p \in (0, 1)$  gilt:

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \log K(n, p) &= H(\mu), \quad \text{bzw.} \\ \lim_{n \rightarrow \infty} \frac{1}{n} \log_2 K(n, p) &= H_2(\mu) := - \sum_{x: \mu(x) \neq 0} \mu(x) \log_2 \mu(x). \end{aligned}$$

Insbesondere gilt: Ist  $A_n$  ( $n \in \mathbb{N}$ ) wesentlich bzgl.  $\mu$ , so ist

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log |A_n| \geq H(\mu).$$

**Bemerkung.** (1). 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.

(2). Der Quellenkodierungssatz zeigt, dass es keine Folge von wesentlichen Mengen gibt, die auf der exponentiellen Skala deutlich langsamer wächst als die im Maßkonzentrationsatz konstruierten Folgen.

*Beweis.* Wir zeigen separat eine obere und eine untere Schranke für  $\frac{1}{n} \log K(n, p)$ :

*Obere Schranke:*  $\limsup_{n \rightarrow \infty} \frac{1}{n} \log K(n, p) \leq H(\mu)$ :

Zum Beweis sei  $\varepsilon > 0$  gegeben. Nach Korollar 7.16 ist die Folge

$$B_n = \{x \in S^n \mid e^{-n(H(\mu)+\varepsilon)} \leq p_n(x_1, \dots, x_n) \leq e^{-n(H(\mu)-\varepsilon)}\}$$

wesentlich bzgl.  $\mu$ , und  $\frac{1}{n} \log |B_n| \leq H(\mu) + \varepsilon$ . Wegen

$$\lim_{n \rightarrow \infty} P[(X_1, \dots, X_n) \in B_n] = 1 > p, \quad (7.4.5)$$

folgt

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log K(n, p) \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log |B_n| \leq H(\mu) + \varepsilon.$$

Die Behauptung ergibt sich für  $\varepsilon \rightarrow 0$ .

*Untere Schranke:*  $\liminf_{n \rightarrow \infty} \frac{1}{n} \log K(n, p) \geq H(\mu)$ :

Seien  $A_n \subseteq S^n$  mit  $P[(X_1, \dots, X_n) \in A_n] \geq p$ . Dann gilt wegen (7.4.5) und (7.4.4) auch

$$p \leq \liminf_{n \rightarrow \infty} P[(X_1, \dots, X_n) \in A_n \cap B_n] \leq \liminf_{n \rightarrow \infty} (|A_n \cap B_n| \cdot e^{-n(H(\mu)-\varepsilon)}),$$

also für alle  $\varepsilon > 0$

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log |A_n| \geq \liminf_{n \rightarrow \infty} \frac{1}{n} \log |A_n \cap B_n| \geq H(\mu) - \varepsilon.$$

Für  $\varepsilon \rightarrow 0$  folgt

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log |A_n| \geq H(\mu).$$

□

# Kapitel 8

## Grenzwertsätze

Sind  $X_i : \Omega \rightarrow \mathbb{R}, i \in \mathbb{N}$ , unabhängige identisch verteilte (i.i.d.) Zufallsvariablen mit Erwartungswert  $m$ , dann konvergieren die Mittelwerte  $\frac{S_n}{n}$  der Summen  $S_n = \sum_{i=1}^n X_i$  nach dem Gesetz der großen Zahlen für  $n \rightarrow \infty$  fast sicher gegen  $m$ . Wir wollen nun die Verteilung von  $S_n$  für große  $n$  genauer untersuchen. Dabei unterscheidet man zwei unterschiedliche Arten von Aussagen:

- *Zentrale Grenzwertsätze* beschreiben „typische“ Fluktuationen um den Grenzwert aus dem 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 Abschnitt 8.4.
- 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 Abschnitt 8.2.

Mit dem Satz von de Moivre/Laplace bzw. der Bernsteinungleichung haben wir bereits entsprechende Aussagen kennengelernt, falls die  $X_i$  Bernoulli-verteilte Zufallsvariablen sind. In diesem Kapitel werden wir sehen, dass keine spezifische Form der Verteilung vorausgesetzt werden muss, sondern die Aussagen ganz allgemein unter geeigneten Integrierbarkeitsbedingungen gelten.

Ein wichtiges Hilfsmittel zum Beweis allgemeiner Grenzwertsätze sind momentenerzeugende und charakteristische Funktionen:



## 8.1 Charakteristische und Momentenerzeugende Funktionen

In diesem Abschnitt führen wir charakteristische und momentenerzeugende Funktionen von reellen Zufallsvariablen ein und beweisen einige grundlegende Aussagen über diese Funktionen. Insbesondere zeigen wir, dass sich die Verteilung einer reellen Zufallsvariable eindeutig aus ihrer charakteristischen Funktion rekonstruieren lässt.

### Definition und Eigenschaften

Sei  $(\Omega, \mathcal{A}, P)$  ein Wahrscheinlichkeitsraum und  $X : \Omega \rightarrow \mathbb{R}$  eine reellwertige Zufallsvariable mit Verteilung  $\mu$ .

**Definition.** (1). Die Funktion  $M : \mathbb{R} \rightarrow (0, \infty]$ ,

$$M(t) := E[e^{tX}] = \int_{\mathbb{R}} e^{tx} \mu(dx),$$

heißt **momentenerzeugende Funktionen** der Zufallsvariable  $X$  bzw. der Verteilung  $\mu$ .

(2). Die Funktion  $\phi : \mathbb{R} \rightarrow \mathbb{C}$ ,

$$\phi(t) := E[e^{itX}] = \int_{\mathbb{R}} e^{itx} \mu(dx),$$

heißt **charakteristische Funktion** von  $X$  bzw.  $\mu$ .

Da die Funktionen  $t \mapsto e^{tx}$  und  $t \mapsto e^{itx}$  für  $t \in \mathbb{R}$  nichtnegativ bzw. beschränkt sind, sind die Erwartungswerte definiert. Dabei wird der Erwartungswert einer komplexwertigen Zufallsvariable separat für Real- und Imaginärteil berechnet.

**Rechenregeln** Die folgenden Rechenregeln ergeben sich unmittelbar aus der Definition:

(1). Sind  $X$  und  $Y$  unabhängige reellwertige Zufallsvariablen auf  $(\Omega, \mathcal{A}, P)$ , dann gilt

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) \quad \text{und} \quad \phi_{X+Y}(t) = \phi_X(t) \cdot \phi_Y(t)$$

für alle  $t \in \mathbb{R}$ .

(2). Für  $a, b \in \mathbb{R}$  gilt

$$M_{aX+b}(t) = e^{bt} \cdot M_X(at) \quad \text{und} \quad \phi_{aX+b}(t) = e^{ibt} \cdot \phi_X(at)$$

für alle  $t \in \mathbb{R}$ .

(3). Für momentenerzeugende bzw. charakteristische Funktionen gilt stets

$$\begin{aligned} M(0) &= \phi(0) = 1, & \text{und} \\ \phi(-t) &= \overline{\phi(t)} & \text{für alle } t \in \mathbb{R}. \end{aligned}$$

Die Funktion  $\phi(-t) = \int e^{-itx} \mu(dx)$  ist die *Fouriertransformation* des Maßes  $\mu$ . Ist  $\mu$  absolutstetig bzgl. des Lebesguemaßes mit Dichte  $f$ , dann ist  $\phi(-t)$  die Fouriertransformation der Funktion  $f$ :

$$\phi(t) = \int_{\mathbb{R}} e^{-itx} f(x) dx = \widehat{f}(t).$$

Entsprechend ist

$$M(-t) = \int_{\mathbb{R}} e^{-tx} \mu(dx) \quad (t > 0)$$

die *Laplacestransformation* des Maßes  $\mu$  bzw. der Dichte  $f$ .

**Bemerkung (Zusammenhang von  $M$  und  $\phi$ ).** (1). Gilt  $M(s) < \infty$  für ein  $s > 0$  (bzw. analog für ein  $s < 0$ ), dann ist  $M$  auf dem Intervall  $[0, s]$  (bzw.  $[s, 0]$ ) endlich, denn nach der Jensenschen Ungleichung folgt:

$$M(t) = E[e^{tX}] \leq E[e^{sX}]^{t/s} < \infty \quad \text{für alle } t \in [0, s] \text{ bzw. } t \in [s, 0].$$

(2). Gilt  $M(t) < \infty$  auf  $(-\delta, \delta)$  für ein  $\delta > 0$ , dann ist  $M$  analytisch fortsetzbar auf den Streifen  $\{z \in \mathbb{C} : |\operatorname{Re}(z)| < \delta\}$  in der komplexen Zahlenebene, und es gilt

$$\phi(t) = M(it) \quad \text{für alle } t \in \mathbb{R}.$$

Die letzte Bemerkung ermöglicht manchmal eine vereinfachte Berechnung der charakteristischen Funktion.

**Beispiel.** (1). Für eine standardnormalverteilte Zufallsvariable  $Z$  gilt:

$$M_Z(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx - x^2/2} dx = e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx = e^{t^2/2} < \infty \quad \text{für alle } t \in \mathbb{R}.$$

Also ist die charakteristische Funktion gegeben durch

$$\phi_Z(t) = M_Z(it) = e^{-t^2/2} \quad \text{für alle } t \in \mathbb{R}.$$

- (2). Eine normalverteilte Zufallsvariable  $X$  mit Mittel  $m$  und Varianz  $\sigma^2$  können wir darstellen als  $X = \sigma Z + m$  mit  $Z \sim N(0, 1)$ . Also gilt:

$$\begin{aligned} M_X(t) &= e^{mt} M_Z(\sigma t) = \exp\left(mt + \frac{\sigma^2 t^2}{2}\right), \\ \phi_X(t) &= \exp\left(imt - \frac{\sigma^2 t^2}{2}\right). \end{aligned}$$

Sind  $X_1, \dots, X_n$  unabhängige,  $N(m, \sigma^2)$ -verteilte Zufallsvariablen, dann erhalten wir:

$$\phi_{X_1 + \dots + X_n}(t) = \prod_{i=1}^n \phi_{X_i}(t) = \exp\left(inmt - \frac{n\sigma^2 t^2}{2}\right).$$

Da die rechte Seite die charakteristische Funktion von  $N(nm, n\sigma^2)$  ist, folgt nach dem Fourierinversionssatz (s.u., Satz 8.2):

$$X_1 + \dots + X_n \sim N(nm, n\sigma^2).$$

- (3). Die Binomialverteilung mit Parametern  $n$  und  $p$  ist die Verteilung der Summe  $\sum_{i=1}^n Y_i$  von unabhängigen Bernoulli( $p$ )-verteilten Zufallsvariablen  $Y_1, \dots, Y_n$ . Also sind

$$\begin{aligned} \phi(t) &= \prod_{i=1}^n \phi_{Y_i}(t) = (1 - p + pe^{it})^n, \quad \text{und} \\ M(t) &= (1 - p + pe^t)^n \end{aligned}$$

die charakteristische und momentenerzeugende Funktion von  $\text{Bin}(n, p)$ .

- (4). Die *Cauchyverteilung* ist die absolutstetige Wahrscheinlichkeitsverteilung auf  $\mathbb{R}$  mit Dichte

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

Für eine Cauchyverteilte Zufallsvariable  $X$  gilt  $M_X(t) = \infty$  für alle  $t \neq 0$  (und sogar  $E[|X|^n] = \infty \quad \forall n \in \mathbb{N}$ ). Trotzdem existiert

$$\phi_X(t) = e^{-|t|} \quad \text{für alle } t \in \mathbb{R}.$$

Die charakteristische Funktion ist allerdings bei 0 nicht differenzierbar.

Wir zeigen nun, daß sich die Momente  $E[X^n]$  einer Zufallsvariable  $X : \Omega \rightarrow \mathbb{R}$  unter geeigneten Voraussetzungen aus der momentenerzeugenden bzw. charakteristischen Funktion berechnen lassen. Die nötigen Voraussetzungen sind allerdings im Fall der momentenerzeugenden Funktion viel stärker:

**Satz 8.1.** (1). Ist  $M$  endlich auf  $(-\delta, \delta)$ ,  $\delta > 0$ , dann gilt

$$E[e^{zX}] = \sum_{n=0}^{\infty} \frac{z^n}{n!} E[X^n] \quad \text{für alle } z \in \mathbb{C} \text{ mit } |z| < \delta.$$

Insbesondere folgt

$$M(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} E[X^n] \quad \text{für alle } t \in (-\delta, \delta),$$

und somit

$$M^{(n)}(0) = E[X^n] \quad \text{für alle } n \geq 0.$$

(2). Ist  $E[|X|^n] < \infty$  für ein  $n \in \mathbb{N}$ , dann gilt  $\phi \in C^n(\mathbb{R})$  und

$$\phi^{(n)}(t) = i^n \cdot E[X^n e^{itX}] \quad \text{für alle } t \in \mathbb{R}. \quad (8.1.1)$$

*Beweis.* (1). Aus der Voraussetzung und dem Satz von der monotonen Konvergenz folgt für  $s \in (0, \delta)$ :

$$\sum_{n=0}^{\infty} \frac{s^n}{n!} E[|X|^n] = E[e^{s|X|}] \leq E[e^{sX}] + E[e^{-sX}] < \infty.$$

Insbesondere existieren alle Momente  $E[X^n]$ ,  $n \in \mathbb{N}$ , sowie die exponentiellen Momente  $E[e^{zX}]$  für  $z \in \mathbb{C}$  mit  $|\operatorname{Re}(z)| < \delta$ . Nach dem Satz von Lebesgue erhalten wir für diese  $z$  zudem

$$\sum_{n=0}^{\infty} \frac{z^n}{n!} E[X^n] = \lim_{m \rightarrow \infty} E \left[ \sum_{n=0}^m \frac{(zX)^n}{n!} \right] = E \left[ \lim_{m \rightarrow \infty} \sum_{n=0}^m \frac{(zX)^n}{n!} \right] = E[e^{zX}],$$

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 (8.1.1) nach Definition von  $\phi(t)$ . Ist  $E[|X|^{n+1}] < \infty$ , dann folgt nach Induktionsvoraussetzung und mit dem Satz von Lebesgue:

$$\begin{aligned} \frac{\phi^{(n)}(t+h) - \phi^{(n)}(t)}{h} &= \frac{1}{h} E \left[ (iX)^n (e^{i(t+h)X} - e^{itX}) \right] \\ &= E \left[ (iX)^n \frac{1}{h} \int_t^{t+h} iX e^{isX} ds \right] \rightarrow E \left[ (iX)^{n+1} e^{itX} \right] \end{aligned}$$

für  $h \rightarrow 0$ , also

$$\phi^{n+1}(t) = E[(iX)^{n+1} \cdot e^{itX}].$$

Die Stetigkeit der rechten Seite in  $t$  folgt ebenfalls aus dem Satz von Lebesgue und der Voraussetzung  $E[|X|^{n+1}] < \infty$ . □

**Beispiel.** Für eine Zufallsvariable  $X$  mit Dichte  $f_X(x) = \text{const.} \cdot e^{-|x|^{1/2}}$  gilt  $E[|X|^n] < \infty$  für alle  $n \in \mathbb{N}$ . Also ist die charakteristische Funktion beliebig oft differenzierbar. Die momentenerzeugende Funktion  $M(t) = E[e^{tX}]$  ist hingegen nur für  $t = 0$  endlich.

**Bemerkung (Satz von Bochner).** Eine Funktion  $\phi : \mathbb{R} \rightarrow \mathbb{C}$  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).  $\phi$  ist gleichmäßig stetig.
- (3).  $\phi$  ist *nicht negativ definit*, d.h.

$$\sum_{i,j=1}^n \phi(t_i - t_j) z_i \bar{z}_j \geq 0 \quad \forall n \in \mathbb{N}, t_1, \dots, t_n \in \mathbb{R}, z_1, \dots, z_n \in \mathbb{C}.$$

Dass jede charakteristische Funktion einer Wahrscheinlichkeitsverteilung die Eigenschaften (1)-(3) hat, prüft man leicht nach (Übung). Der Beweis der umgekehrten Aussage findet sich z.B. in Vol. II des Lehrbuchs von Feller.

## Inversion der Fouriertransformation

Die folgende zentrale Aussage zeigt, dass eine Wahrscheinlichkeitsverteilung *eindeutig* durch ihre charakteristische Funktion  $\phi$  festgelegt ist, und liefert eine *explizite Formel* zur Rekonstruktion der Verteilung aus  $\phi$ :

**Satz 8.2 (Lévy's Inversionsformel).** Sei  $\phi$  die charakteristische Funktion einer Zufallsvariable  $X$  mit Verteilung  $\mu$ . Dann gilt:

(1).

$$\frac{1}{2} \mu[\{a\}] + \mu[(a, b)] + \frac{1}{2} \mu[\{b\}] = \frac{1}{2\pi} \lim_{T \rightarrow \infty} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \phi(t) dt \quad \forall a < b.$$

(2). Gilt  $\int_{-\infty}^{\infty} |\phi(t)| dt < \infty$ , dann ist  $\mu$  absolutstetig mit stetiger Dichte

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t) dt.$$

**Bemerkung.** (1). Die Verteilung  $\mu$  ist durch (1) eindeutig festgelegt, denn für  $c, d \in \mathbb{R}$  mit  $c < d$  gilt:

$$\frac{1}{2} \mu[\{a\}] + \mu[(a, b)] + \frac{1}{2} \mu[\{b\}] = \frac{1}{2} \left( \mu \left[ [a, b] \right] + \mu \left[ (a, b) \right] \right) \rightarrow \mu[(c, d)],$$

für  $a \searrow c$  und  $b \nearrow d$ .

- (2). Ist die Verteilung  $\mu$  absolutstetig mit quadratintegrierbarer Dichte  $f$ , dann ist auch die entsprechende charakteristische Funktion

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx$$

quadratintegrierbar. Die Aussage (2) aus Satz 8.2 ist in diesem Fall die klassische *Fourierinversionsformel der Analysis*, siehe z.B. Forster „Analysis 3“.

Im Beweis der Inversionsformel verwenden wir den Satz von Fubini, der besagt, dass wir die Integrationsreihenfolge in Doppelintegralen vertauschen dürfen, wenn der Integrand produktintegrierbar ist. Für den Beweis des Satzes von Fubini verweisen wir auf die Analysisvorlesung oder Abschnitt 9.1.

von Satz 8.2. (1). Sei  $T > 0$  und  $a < b$ . Nach dem Satz von Fubini können wir die Integrationsreihenfolge in dem folgendem Doppelintegral vertauschen, und erhalten:

$$\frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \underbrace{\phi(t)}_{= \int e^{itx} \mu(dx)} dt = \frac{1}{\pi} \int \underbrace{\int_{-T}^T \frac{e^{it(x-a)} - e^{it(x-b)}}{2it} dt}_{=: g(T,x)} \mu(dx) \quad (8.1.2)$$

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

$$\left| \frac{e^{it(x-a)} - e^{it(x-b)}}{it} \right| \leq \frac{|t \cdot (x-a) - t \cdot (x-b)|}{|t|} = |a-b| \quad \text{gilt.}$$

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:

$$\lim_{t \rightarrow \infty} S(t) = \frac{\pi}{2}, \quad \lim_{t \rightarrow -\infty} S(t) = -\frac{\pi}{2}.$$

Damit erhalten wir:

$$\lim_{T \rightarrow \infty} g(T, x) = \frac{\pi}{2} \operatorname{sgn}(x - a) - \frac{\pi}{2} \operatorname{sgn}(x - b) = \pi \cdot I_{(a,b)}(x) + \frac{\pi}{2} \cdot I_{\{a,b\}}(x),$$

wobei wir  $\operatorname{sgn}(0) := 0$  setzen. Da  $S$  beschränkt ist, ist auch  $g(T, x)$  beschränkt in  $T$  und  $x$ .

Nach dem Satz von Lebesgue folgt daher aus (8.1.2) für  $T \rightarrow \infty$

$$\begin{aligned} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \phi(t) dt &= \frac{1}{\pi} \int g(T, x) \mu(dx) \\ &\xrightarrow{T \rightarrow \infty} \mu[(a, b)] + \frac{1}{2} \mu[\{a, b\}]. \end{aligned}$$

- (2). Ist  $\phi$  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):

$$\int_a^b f(x) dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi(t) \underbrace{\int_a^b e^{-itx} dx}_{= \frac{e^{-ita} - e^{-itb}}{it}} dt \stackrel{(1)}{=} \frac{1}{2} \mu[\{a\}] + \mu[(a, b)] + \frac{1}{2} \mu[\{b\}].$$

Insbesondere folgt

$$\int_{a+\varepsilon}^{b-\varepsilon} f(x) dx \leq \mu[(a, b)] \leq \int_a^b f(x) dx \quad \forall \varepsilon > 0,$$

also für  $\varepsilon \searrow 0$ :

$$\mu[(a, b)] = \int_a^b f(x) dx.$$

□

## 8.2 Erste Anwendungen auf Grenzwertsätze

Charakteristische und momentenerzeugende Funktionen werden häufig beim Beweis von Grenzwertsätzen der Wahrscheinlichkeitstheorie verwendet. Wir skizzieren an dieser Stelle schon einmal die Anwendung charakteristischer Funktionen zum Beweis des zentralen Grenzwertsatzes und zeigen anschließend, wie obere Schranken für die Wahrscheinlichkeiten großer Abweichungen vom Gesetz der großen Zahlen mithilfe momentenerzeugender Funktionen hergeleitet werden können. Der detaillierte Beweis des zentralen Grenzwertsatzes wird dann nach weiteren Vorbereitungen in Abschnitt 8.3 ausgeführt. Die Analyse der Asymptotik der Wahrscheinlichkeiten großer Abweichungen auf der exponentiellen Skala werden wir in Kapitel ?? durch den Beweis einer unteren Schranke vervollständigen.

## Zentraler Grenzwertsatz

Seien  $X_1, X_2, \dots \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$  unabhängige und identisch verteilte Zufallsvariablen mit  $E[X_i] = 0$  für alle  $i$ , und sei  $S_n = X_1 + \dots + X_n$ . Nach dem Gesetz der großen Zahlen gilt:

$$\frac{S_n}{n} \rightarrow 0 \quad P\text{-fast sicher.}$$

### 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

$$\text{Var}[S_n] = n \cdot \text{Var}[X_1],$$

also ist

$$\text{Var} \left[ \frac{S_n}{\sqrt{n}} \right] = \frac{1}{n} \cdot \text{Var}[S_n] = \text{Var}[X_1] =: \sigma^2$$

unabhängig von  $n$ .

Um die Asymptotik der Verteilungen der entsprechend standardisierten Summen  $\frac{S_n}{\sqrt{n}}$  zu bestimmen, betrachten wir die charakteristischen Funktionen. Da die Summanden  $X_i$  unabhängig und identisch verteilt sind, erhalten wir

$$\phi_{\frac{S_n}{\sqrt{n}}}(t) = \phi_{S_n} \left( \frac{t}{\sqrt{n}} \right) \stackrel{X_i \text{ iid}}{=} \left[ \phi_{X_1} \left( \frac{t}{\sqrt{n}} \right) \right]^n.$$

Wegen  $X_1 \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$  ist  $\phi_{X_1}$  zweimal stetig differenzierbar, und die Taylorentwicklung bei  $t = 0$  ist gegeben durch

$$\phi_{X_1}(t) = 1 + i \cdot E[X_1] \cdot t - \frac{1}{2} E[X_1^2] \cdot t^2 + o(t^2) = 1 - \frac{1}{2} \sigma^2 t^2 + o(t^2).$$

Damit folgt:

$$\begin{aligned} \phi_{\frac{S_n}{\sqrt{n}}}(t) &= \left( 1 - \frac{\sigma^2 t^2}{2n} + o\left(\frac{t^2}{n}\right) \right)^n \\ &\xrightarrow{n \nearrow \infty} \exp\left(-\frac{\sigma^2 t^2}{2}\right) = \phi_{N(0, \sigma^2)}(t) \quad \forall t \in \mathbb{R}. \end{aligned}$$

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 Verteilung der standardisierten Summen  $\frac{S_n}{\sqrt{n}}$  konvergiert schwach gegen die Normalverteilung  $N(0, \sigma^2)$ .

Den detaillierten Beweis werden wir in Abschnitt 8.3 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  $X_1 \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ .

## Große Abweichungen vom Gesetz der großen Zahlen

Seien  $X_1, X_2, \dots \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$  i.i.d. Zufallsvariablen mit Erwartungswert  $m$  und momentenerzeugender Funktion

$$M(t) = E[e^{tX_1}],$$

und sei  $S_n = X_1 + \dots + X_n$ .

Der folgende Satz verschärft die *nicht-asymptotische* obere Schranke für große Abweichungen vom Gesetz der großen Zahlen aus der Bernstein-Ungleichung (Satz 2.6), und verallgemeinert diese auf nicht Bernoulliverteilte Zufallsvariablen.

**Satz 8.3 (Chernoff).** Für alle  $n \in \mathbb{N}$  und  $a \in \mathbb{R}$  gilt:

$$\begin{aligned} P \left[ \frac{S_n}{n} \geq a \right] &\leq e^{-nI(a)} \quad \text{falls } a \geq m, \text{ bzw.} \\ P \left[ \frac{S_n}{n} \leq a \right] &\leq e^{-nI(a)} \quad \text{falls } a \leq m, \end{aligned}$$

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  $\tilde{X}_i = X_i - E[X_i]$ , die wieder unabhängig und identisch verteilt sind. Man überzeugt sich leicht, dass aus der Behauptung für  $\tilde{X}_i$  die Behauptung für  $X_i$  folgt (Übung).

(2). *Exponentielle Markovungleichung:* Für alle  $t \geq 0$  gilt:

$$\begin{aligned} P \left[ \frac{S_n}{n} \geq a \right] &= P[S_n \geq na] \leq e^{-tna} E[e^{tS_n}] \\ &\stackrel{X_i \text{ iid}}{=} e^{-tna} E[e^{tX_1}]^n = e^{-(at - \log M(t)) \cdot n}. \end{aligned}$$

- (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

$$\sup_{t \geq 0} (at - \log M(t)) = \sup_{t \in \mathbb{R}} (at - \log M(t)) = I(a).$$

Dies ist in der Tat der Fall, denn für  $t < 0$  und  $a \geq 0$  gilt nach der Jensenschen Ungleichung und der Voraussetzung  $m = 0$ :

$$\begin{aligned} at - \log M(t) &\leq -\log E[e^{tX_1}] \leq -E[\log e^{tX_1}] \\ &= -tm = 0 = a \cdot 0 - \log M(0). \end{aligned}$$

□

**Bemerkung (Kumulantenerzeugende Funktion, Legendretransformation).** (1). Die Funktion  $\Lambda(t) := \log M(t)$  heißt *logarithmische momentenerzeugende* oder *kumulantenerzeugende Funktion* von  $X_1$ . Diese Funktion hat u.a. folgende Eigenschaften:

- (a)  $\Lambda$  ist konvex und *unterhalbstetig*, d.h.  $\liminf_{s \rightarrow t} \Lambda(s) \geq \Lambda(t)$  für alle  $t \in \mathbb{R}$ .  
 (b)  $\Lambda(0) = 0$ .  
 (c) Gilt  $M(t) < \infty$  auf  $(-\delta, \delta)$  für ein  $\delta > 0$ , dann ist

$$\begin{aligned} \Lambda'(0) &= \frac{M'(0)}{M(0)} = m, \quad \text{und} \\ \Lambda''(0) &= \frac{M''(0)}{M(0)} - \frac{M'(0)^2}{M(0)^2} = E[X_1^2] - E[X_1]^2 = \text{Var}[X_1]. \end{aligned}$$

Die höheren Ableitungen von  $\Lambda$  heißen *Kumulanten* von  $X_1$ .

- (2). Die Ratenfunktion  $I$  ist die *Legendre-Transformation* von  $\Lambda$ :

$$I(a) = \sup_{t \in \mathbb{R}} f_a(t) \quad \text{mit} \quad f_a(t) = at - \Lambda(t),$$

d.h.  $I(a)$  ist der negative Achsenabschnitt der (eindeutigen) Tangente an den Graphen von  $\Lambda$  mit Steigung  $a$  (wobei  $I(a) = \infty$ , falls keine solche Tangente existiert).

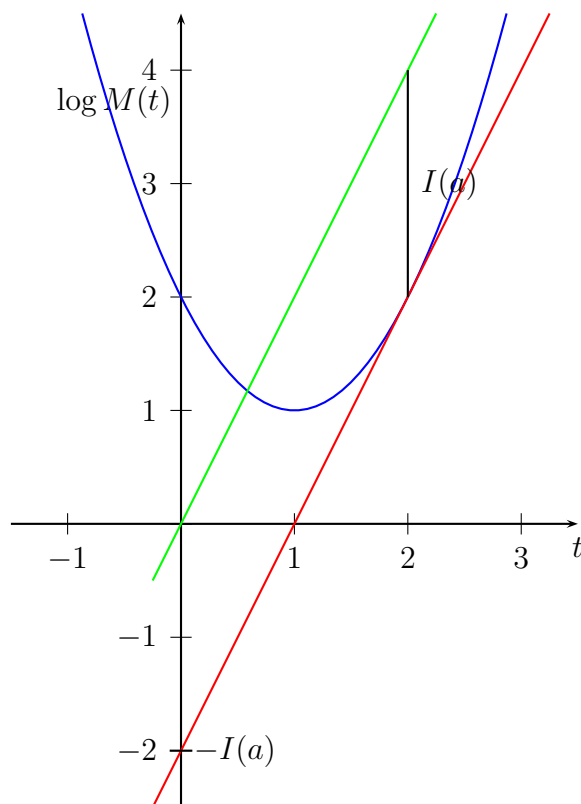


Abbildung 8.1: Geometrische Darstellung der Ratenfunktion  $I(a)$  als negativer Achsenabschnitt der eindeutigen Tangente mit Steigung  $a$  (rot) an die Kumulantenerzeugende Funktion (blau)

Wichtige Eigenschaften der Ratenfunktion sind:

- (a)  $I$  ist wieder konvex und unterhalbstetig.
- (b)  $I(a) \geq f_a(0) = 0 \quad \forall a \in \mathbb{R}$ .
- (c) Gilt  $M(t) < \infty$  auf  $(-\delta, \delta)$  für ein  $\delta > 0$ , dann ist  $f_a \in C^\infty(-\delta, \delta)$  mit  $f_a(0) = 0$  und  $f'_a(0) = a - m$ . Also folgt:

$$I(a) = \sup f_a > 0 \quad \forall a \neq m.$$

Unter der Voraussetzung der letzten Bemerkung (c) ist die exponentielle Abfallrate strikt positiv, d.h. es ergibt sich ein *exponentieller Abfall der Wahrscheinlichkeiten großer Abweichungen!* Sind die Zufallsvariablen  $X_i$  nicht exponentiell integrierbar, dann kann es auch passieren, daß  $I(a) = 0$  für  $a \neq m$ . 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 ?? unten.

**Beispiel.** Für konkrete Verteilungen der Zufallsvariablen  $X_i$  kann man die Kumulantenerzeugende Funktion  $\Lambda$  und die Ratenfunktion  $I$  häufig explizit berechnen:

(1). Für normalverteilte Zufallsvariablen  $X_i \sim N(m, \sigma^2)$  gilt  $I(a) = \frac{(a-m)^2}{2\sigma^2}$ , also

$$P\left[\frac{S_n}{n} \geq a\right] \leq e^{-\frac{(a-m)^2 n}{2\sigma^2}} \quad \text{für alle } a \geq m.$$

Die Ratenfunktion hat eine Nullstelle beim Erwartungswert  $m$ , da die Mittelwert  $S_n/n$  gegen diese konvergieren. Jenseits von  $m$  fallen die Wahrscheinlichkeiten exponentiell ab, und zwar mit einer Rate die quadratisch wächst.

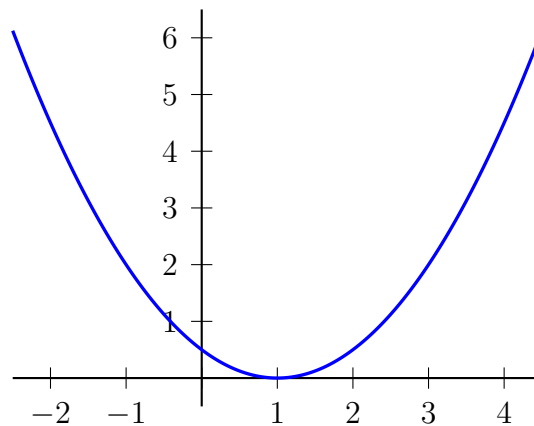


Abbildung 8.2: Legendre-Transformation der logarithmischen momentenerzeugenden Funktion einer  $\mathcal{N}(1, 1)$ -verteilten Zufallsvariable

(2). Für  $X_i \sim \text{Exp}(\lambda)$  gilt

$$I(a) = \begin{cases} \lambda a - 1 - \log(\lambda a) & \text{für } a > 0 \\ \infty & \text{für } a \leq 0 \end{cases}.$$

In diesem Fall hat die Ratenfunktion 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  $+\infty$ .

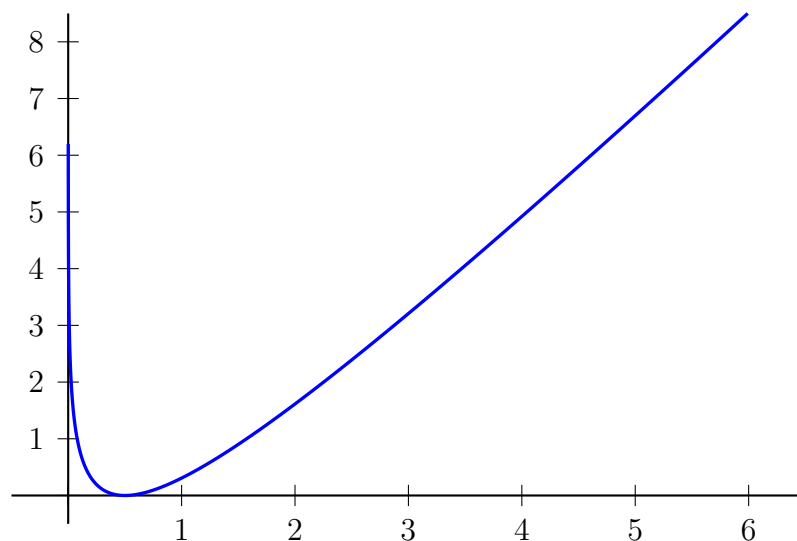


Abbildung 8.3: Legendre-Transformierte der logarithmischen momentenerzeugenden Funktion einer  $\text{Exp}(2)$ -verteilten Zufallsvariable

(3). Für  $X_i \sim \text{Bernoulli}(p)$  erhält man

$$I(a) = a \log \left( \frac{a}{p} \right) + (1 - a) \log \left( \frac{1 - a}{1 - p} \right) \quad \text{für } a \in (0, 1).$$

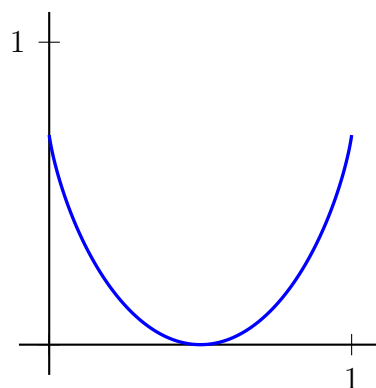


Abbildung 8.4: Legendre-Transformation der logarithmischen momentenerzeugenden Funktion einer  $\text{Bernoulli}(1/2)$ -verteilten Zufallsvariable

Wegen  $I(a) \geq 2(a - p)^2$  verschärft die Abschätzung aus dem Satz von Chernoff in diesem Fall die in Satz 2.6 hergeleitete obere Schranke

$$P \left[ \frac{S_n}{n} \geq a \right] \leq e^{-2(a-p)^2 n} \quad \text{für } a \geq p.$$

Wir werden später sehen, daß  $I(a)$  sich als relative Entropie der  $\text{Bernoulli}(a)$ -Verteilung bzgl. der  $\text{Bernoulli}(p)$ -Verteilung interpretieren läßt.

**Beispiel (Ehrenfestmodell im Gleichgewicht).** Es befinden sich  $n = 10^{23}$  Moleküle in einem Gefäß. Jedes Molekül sei mit Wahrscheinlichkeit  $\frac{1}{2}$  in der linken bzw. rechten Hälfte. Seien  $X_i$  ( $1 \leq i \leq n$ ) Bernoulli( $\frac{1}{2}$ )-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  $p := P \left[ \frac{S_n}{n} \geq \frac{1}{2} + 10^{-10} \right]$ ?

Eine Abschätzung mit der Čebyšev-Ungleichung liefert:

$$p \leq 10^{20} \cdot \text{Var} \left[ \frac{S_n}{n} \right] = \frac{1}{4} \cdot 10^{-3} = \frac{1}{4000}.$$

Durch Anwenden der exponentiellen Abschätzung erhält man dagegen die viel präzisere Aussage

$$p \leq e^{-2n(10^{-10})^2} = e^{-2000}.$$

Eine Abweichung von der Größenordnung  $10^{-10}$  vom Mittelwert ist also **praktisch unmöglich**!

Die makroskopische Größe  $S_n/n$  ist daher de facto deterministisch.

### 8.3 Verteilungskonvergenz

Sei  $S$  ein metrischer Raum mit Borelscher  $\sigma$ -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  $\mu_n$  einer Folge  $Y_n$  von Zufallsvariablen mit Werten in  $S$  einführen. Naheliegender wäre es zu definieren, dass eine Folge  $\mu_n$  von Wahrscheinlichkeitsverteilungen auf  $(S, \mathcal{B}(S))$  gegen eine Wahrscheinlichkeitsverteilung  $\mu$  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 Wahrscheinlichkeitsverteilungen:** Eine Folge  $(\mu_n)_{n \in \mathbb{N}}$  von Wahrscheinlichkeitsverteilungen auf  $S$  (mit Borelscher  $\sigma$ -Algebra) **konvergiert schwach** gegen eine Wahrscheinlichkeitsverteilung  $\mu$  auf  $S$  ( $\mu_n \xrightarrow{w} \mu$ ), falls

$$\int f d\mu_n \longrightarrow \int f d\mu \quad \text{für alle stetigen, beschränkten } f : S \rightarrow \mathbb{R} \text{ gilt.}$$

- (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

$$\text{Verteilung}(Y_n) \xrightarrow{w} \text{Verteilung}(Y),$$

d.h. falls

$$E[f(Y_n)] \longrightarrow E[f(Y)] \quad \text{für alle } f \in C_b(S) \text{ gilt.}$$

Konvergenz in Verteilung bezeichnet man auf Englisch als „convergence in distribution“ oder „convergence in law.“ Entsprechend verwendet man die Kurzschreibweisen  $Y_n \xrightarrow{D} Y$  oder  $Y_n \xrightarrow{L} Y$ , falls  $Y_n$  in Verteilung gegen  $Y$  konvergiert.

**Beachte:** Die Zufallsvariablen  $Y_n, n \in \mathbb{N}$ , und  $Y$  können bei der Verteilungskonvergenz auf verschiedenen Wahrscheinlichkeitsräumen definiert sein!

## Schwache Konvergenz von Wahrscheinlichkeitsverteilungen

Um den Begriff der schwachen Konvergenz besser zu erfassen, beginnen wir mit einigen Bemerkungen und Beispielen:

**Bemerkung.** (1). 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.

- (2). Wir werden in Satz 8.5 zeigen, dass im Fall  $S = \mathbb{R}$  die Folge  $\mu_n$  genau dann schwach gegen  $\mu$  konvergiert, wenn für die Verteilungsfunktionen

$$F_{\mu_n}(x) \longrightarrow F_{\mu}(x) \quad \text{für alle Stetigkeitsstellen } x \text{ von } F,$$

d.h. für alle  $x \in \mathbb{R}$  mit  $\mu[\{x\}] = 0$ , gilt.

Neben schwacher Konvergenz betrachtet man häufig u.a. auch die folgenden Konvergenzarten auf positiven bzw. beschränkten signierten Maßen:

- **Vage Konvergenz:**  $\mu_n$  konvergiert vage gegen  $\mu$ , falls

$$\int f d\mu_n \longrightarrow \int f d\mu$$

für alle stetigen Funktionen  $f$  mit kompaktem Träger gilt.

- **Konvergenz in Variationsdistanz:**  $\mu_n$  konvergiert  $\mu$  in Variationsdistanz, falls

$$\|\mu - \mu_n\|_{\text{TV}} := \frac{1}{2} \sup_{\substack{f: S \rightarrow \mathbb{R} \text{ messbar} \\ \text{mit } |f| \leq 1}} \left| \int f d\mu - \int f d\mu_n \right| \longrightarrow 0.$$

Die Variationsdistanz zweier Wahrscheinlichkeitsverteilungen läßt sich auch wie folgt darstellen:

$$\|\mu - \nu\|_{\text{TV}} = \sup_{A \in \mathcal{S}} |\mu[A] - \nu[A]|.$$

Im diskreten Fall gilt

$$\|\mu - \nu\|_{\text{TV}} = \frac{1}{2} \sum_{x \in S} |\mu[\{x\}] - \nu[\{x\}]|.$$

Diesen Abstands begriff haben wir bereits in Abschnitt 3.5 bei der Konvergenz ins Gleichgewicht von Markovketten verwendet.

Offensichtlich folgt aus der Konvergenz in Variationsdistanz die schwache Konvergenz, aus der wiederum die vage Konvergenz folgt:

$$\|\mu_n - \mu\|_{\text{TV}} \rightarrow 0 \implies \mu_n \xrightarrow{w} \mu \implies \mu_n \rightarrow \mu \text{ vage.}$$

Die folgenden Beispiele verdeutlichen die unterschiedlichen Konvergenzbegriffe:

**Beispiel.** (1). *Diracmaße:* Für  $x, x_n \in S$  ( $n \in \mathbb{N}$ ) mit  $x_n \rightarrow x$  gilt  $\delta_{x_n} \xrightarrow{w} \delta_x$ .

*Beweis:*

$$\int f d\delta_{x_n} = f(x_n) \rightarrow f(x) = \int f d\delta_x \quad \text{für alle } f \in C_b(\mathbb{R}).$$

*Alternativer Beweis im Fall  $S = \mathbb{R}$ :*

$$F_{\delta_{x_n}}(c) = I_{[x_n, \infty)}(c) \xrightarrow{n \rightarrow \infty} I_{[x, \infty)}(c) = F_{\delta_x}(c) \quad \text{für alle } c \neq x,$$

d.h. für alle Stetigkeitsstellen von  $F_{\delta_x}$ .

In diesem Beispiel gilt i.A. keine Konvergenz in Variationsnorm, denn  $\|\delta_{x_n} - \delta_x\|_{\text{TV}} = 1$  für  $x_n \neq x$ .



- (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 Diracmaß  $\delta_0$ , denn mit dem Satz von Lebesgue folgt für  $f \in C_b(\mathbb{R})$

$$\begin{aligned} \int f d\mu_n &= \int f(x) \frac{1}{\sqrt{2\pi/n}} e^{-\frac{x^2}{2/n}} dx \\ &\stackrel{y=\sqrt{n}x}{=} \int f\left(\frac{y}{\sqrt{n}}\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\ &\stackrel{\text{Lebesgue}}{\longrightarrow} f(0) \cdot \underbrace{\int \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy}_{=1} \\ &= \int f d\delta_0. \end{aligned}$$

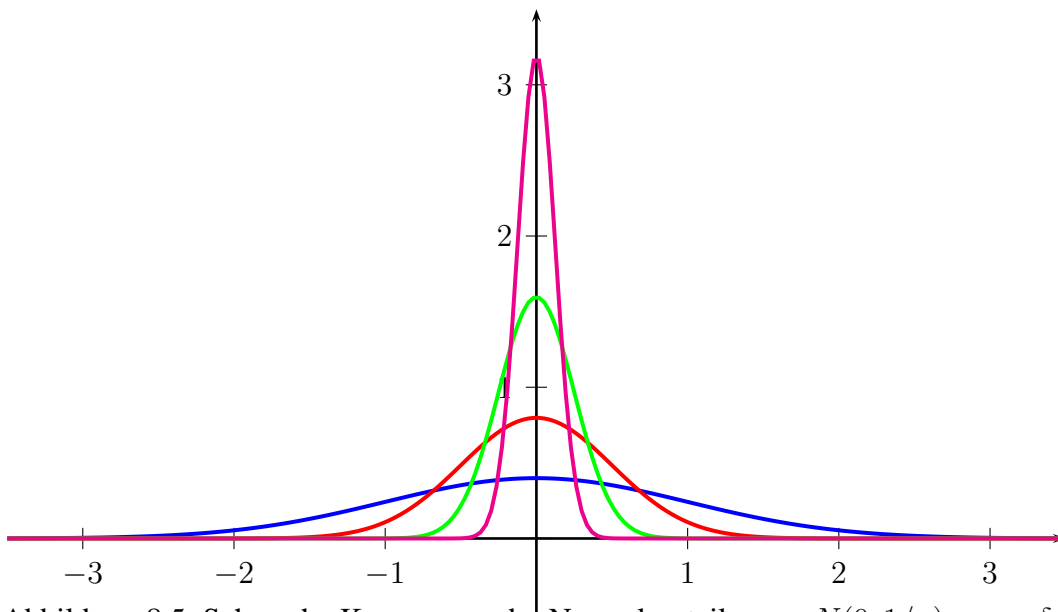


Abbildung 8.5: Schwache Konvergenz der Normalverteilungen  $N(0, 1/n)$  gegen  $\delta_0$ .

- (3). **Schwache vs. vage Konvergenz:** Die Folge  $\mu_n = N(0, n)$  konvergiert vage gegen das Nullmaß  $\mu$  mit  $\mu[A] = 0$  für alle  $A$ . In der Tat gilt für  $f \in C(\mathbb{R})$  mit  $f(x) = 0$  für  $x \notin [-K, K]$ :

$$\left| \int f d\mu_n \right| = \left| \int_{-K}^K f(x) \cdot \frac{1}{\sqrt{2\pi n}} e^{-x^2/2n} dx \right| \leq \frac{2K}{\sqrt{2\pi n}} \cdot \sup |f| \xrightarrow{n \rightarrow \infty} 0.$$

Es gilt aber keine schwache Konvergenz, da

$$\int 1 d\mu_n = \mu_n[\mathbb{R}] = 1 \not\rightarrow 0.$$

Die Masse wandert in diesem Fall ins Unendliche ab.

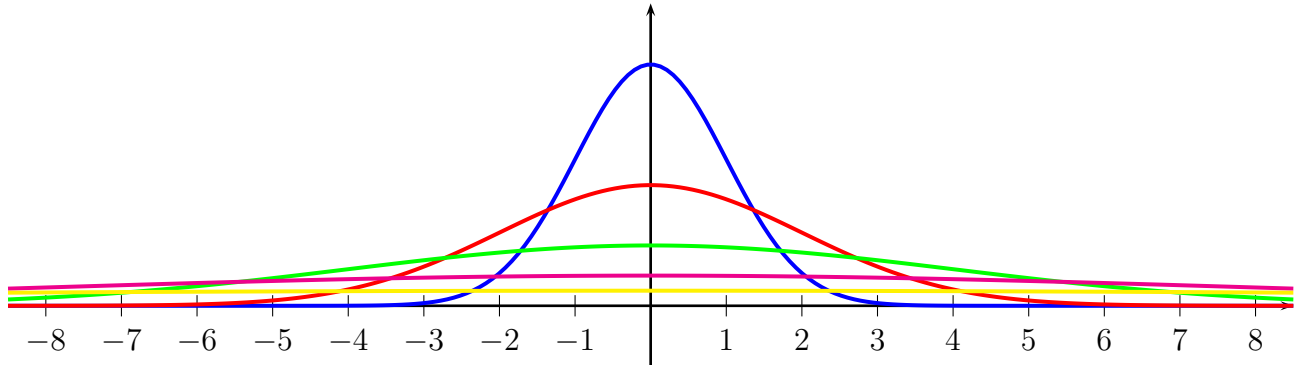


Abbildung 8.6: Konvergenz der Dichten der Normalverteilungen  $N(0, n)$  gegen die Nullfunktion.

- (4). **Wartezeiten:** Die Wartezeit  $T_p$  auf den ersten Erfolg bei unabhängigen Ereignissen mit Erfolgswahrscheinlichkeit  $p \in (0, 1)$  ist geometrisch verteilt:

$$P[T_p > k] = (1 - p)^k \quad \text{für alle } k \in \mathbb{N}.$$

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  $\lambda/n$  stattfinden. Dann ist  $\frac{1}{n}T_{\lambda/n}$  die Wartezeit bis zum ersten Eintreten eines Ereignisses. Für  $n \rightarrow \infty$  gilt:

$$P\left[\frac{1}{n}T_{\lambda/n} > x\right] = P\left[T_{\lambda/n} > nx\right] = \left(1 - \frac{\lambda}{n}\right)^{\lfloor nx \rfloor} \xrightarrow{n \nearrow \infty} e^{-\lambda x} \quad \forall x \geq 0.$$

Also konvergiert die Verteilung von  $\frac{1}{n}T_{\lambda/n}$  schwach gegen die Exponentialverteilung mit Parameter  $\lambda$ . Konvergenz in Variationsdistanz gilt nicht, da die approximierenden Verteilungen diskret, und die Grenzverteilungen stetig sind.

- (5). **Diskrete Approximation von Wahrscheinlichkeitsverteilungen:** Allgemeiner können wir eine gegebene Wahrscheinlichkeitsverteilung auf verschiedene Arten durch diskrete Wahrscheinlichkeitsverteilungen, also Konvexkombinationen von Diracmaßen approximieren:

- (a) **Klassische numerische Approximation:** Sei  $\mu$  eine absolutstetige Wahrscheinlichkeitsverteilung auf  $[0, 1]$  mit Dichtefunktion proportional zu  $g(x)$ , und sei

$$\mu_n := \sum_{i=1}^n w_n^{(i)} \delta_{\frac{i}{n}},$$

mit

$$w_n^{(i)} = \frac{g\left(\frac{i}{n}\right)}{\sum_{j=1}^n g\left(\frac{j}{n}\right)}.$$

Dann konvergiert  $\mu_n$  schwach gegen  $\mu$ , denn

$$\begin{aligned} \int f d\mu_n &= \sum_{i=1}^n w_n^{(i)} f\left(\frac{i}{n}\right) = \frac{\frac{1}{n} \sum_{i=1}^n f\left(\frac{i}{n}\right) g\left(\frac{i}{n}\right)}{\frac{1}{n} \sum_{i=1}^n g\left(\frac{i}{n}\right)} \\ &\xrightarrow{n \nearrow \infty} \frac{\int_0^1 f g dx}{\int_0^1 g dx} = \int f d\mu \quad \forall f \in C([0, 1]). \end{aligned}$$

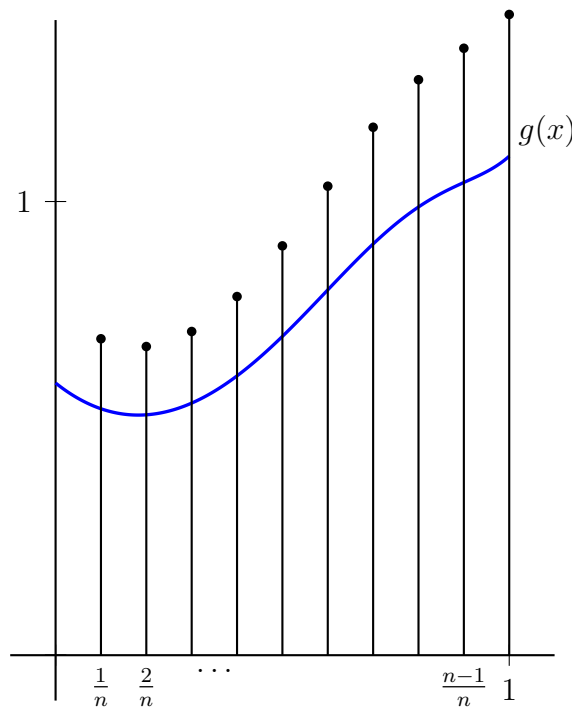


Abbildung 8.7: Stützstellen und Gewichte einer deterministischen Approximation von  $\mu$ .

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  $(S, \mathcal{S}, \mu)$  ein beliebiger Wahrscheinlichkeitsraum. Sind  $X_1, X_2, \dots : \Omega \rightarrow S$  unabhängige Zufallsvariablen auf  $(\Omega, \mathcal{A}, P)$  mit Verteilung  $\mu$ , dann konvergieren die **empirischen Verteilungen**

$$\hat{\mu}_n(\omega, \bullet) := \frac{1}{n} \sum_{i=1}^n \delta_{X_i(\omega)}$$

$P$ -f.s. schwach gegen  $\mu$ , denn für  $f \in C_b(S)$  gilt nach dem starken Gesetz großer Zahlen für  $P$ -fast alle  $\omega$

$$\int f d\hat{\mu}_n(\omega, \bullet) = \frac{1}{n} \sum_{i=1}^n \underbrace{f(X_i(\omega))}_{\substack{\text{iid,} \\ \text{beschränkt}}} \xrightarrow{\text{GdgZ}} E[f(X_1)] = \int f d\mu.$$

## 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. Wir untersuchen nun den Zusammenhang der schwachen Konvergenz der Verteilungen mit anderen Konvergenzarten in dem Fall, dass  $Y_n$  ( $n \in \mathbb{N}$ ) und  $Y$  reellwertige Zufallsvariablen sind, die auf einem gemeinsamen Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  definiert sind.

**Satz 8.4.** *Konvergiert  $Y_n$   $P$ -fast sicher oder  $P$ -stochastisch gegen  $Y$ , dann konvergiert  $Y_n$  auch in Verteilung gegen  $Y$ .*

*Beweis.* Sei  $f \in C_b(\mathbb{R})$ . Konvergiert  $Y_n$  fast sicher gegen  $Y$ , dann konvergiert auch  $f(Y_n)$  fast sicher gegen  $f(Y)$ . Nach dem Satz von Lebesgue folgt

$$E[f(Y_n)] \longrightarrow E[f(Y)].$$

Konvergiert  $Y_n$  nur stochastisch gegen  $Y$ , dann hat jede Teilfolge  $(Y_{n_k})_{k \in \mathbb{N}}$  von  $(Y_n)_{n \in \mathbb{N}}$  eine fast sicher gegen  $Y$  konvergente Teilfolge  $(Y_{n_{k_l}})_{l \in \mathbb{N}}$ . Wie zuvor folgt

$$E[f(Y_{n_{k_l}})] \longrightarrow E[f(Y)].$$

Also hat jede Teilfolge der Folge  $(E[f(Y_n)])_{n \in \mathbb{N}}$  der Erwartungswerte eine gegen  $E[f(Y)]$  konvergente Teilfolge, d.h. es gilt erneut

$$E[f(Y_n)] \longrightarrow E[f(Y)].$$

□

Wir beweisen nun eine partielle Umkehrung der Aussage aus Satz 8.4:

**Satz 8.5 (Skorokhod - Darstellung und Charakterisierung der schwachen Konvergenz).**

Seien  $\mu_n, \mu$  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  $\mu$ .
- (2).  $F_n(c) \rightarrow F(c)$  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  $\mu_n$  bzw.  $\mu$ , sodass  $G_n \rightarrow G$   $P$ -fast sicher.

*Beweis.* „(3)  $\Rightarrow$  (1)“ folgt aus Satz 8.4.

„(1)  $\Rightarrow$  (2)“: Für  $c \in \mathbb{R}$  gilt:

$$F_n(c) = \int I_{(-\infty, c]} d\mu_n \quad \text{und} \quad F(c) = \int I_{(-\infty, c]} d\mu. \quad (8.3.1)$$

Sei  $\varepsilon > 0$ . Wir definieren stetige Approximationen der Indikatorfunktion  $I_{(-\infty, c]}$  durch

$$f_\varepsilon(x) = \begin{cases} 1 & \text{für } x \leq c - \varepsilon \\ 0 & \text{für } x \geq c \\ \frac{c-x}{\varepsilon} & \text{für } x \in [(c - \varepsilon), c) \end{cases}, \quad \text{und} \quad g_\varepsilon(x) = \begin{cases} 1 & \text{für } x \leq c \\ 0 & \text{für } x \geq c + \varepsilon \\ \frac{c+\varepsilon-x}{\varepsilon} & \text{für } x \in (c, c + \varepsilon) \end{cases}.$$

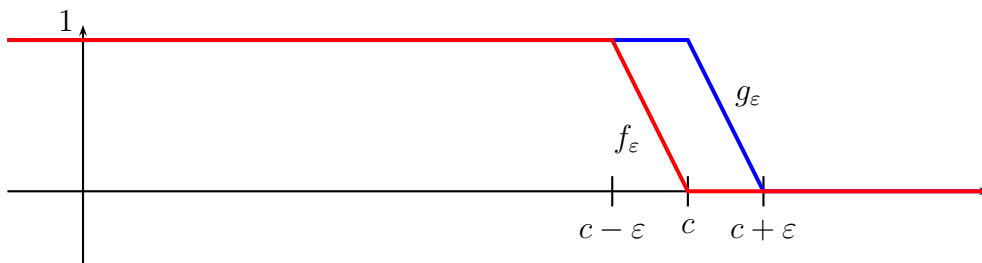


Abbildung 8.8: Stetige Approximationen von  $I_{(-\infty, c]}$ .

Es gilt

$$I_{(-\infty, c - \varepsilon]} \leq f_\varepsilon \leq I_{(-\infty, c]} \leq g_\varepsilon \leq I_{(-\infty, c + \varepsilon]}. \quad (8.3.2)$$

Konvergiert  $\mu_n$  schwach gegen  $\mu$ , dann folgt nach (8.3.1) und (8.3.2):

$$\begin{aligned} \liminf F_n(c) &\geq \liminf \int f_\varepsilon d\mu_n = \int f_\varepsilon d\mu \geq F(c - \varepsilon), & \text{und} \\ \limsup F_n(c) &\leq \limsup \int g_\varepsilon d\mu_n = \int g_\varepsilon d\mu \leq F(c + \varepsilon). \end{aligned}$$

Für  $\varepsilon \searrow 0$  erhalten wir

$$\limsup F_n(c) \leq F(c) = \lim_{\varepsilon \searrow 0} F(c + \varepsilon),$$

und

$$\liminf F_n(c) \geq F(c) = \lim_{\varepsilon \searrow 0} F(c - \varepsilon),$$

falls  $F$  bei  $c$  stetig ist.

„(2)  $\Rightarrow$  (3)“: Für  $u \in (0, 1)$  betrachten wir die minimalen und maximalen  $u$ -Quantile

$$\underline{G}(u) := \inf\{x \in \mathbb{R} \mid F(x) \geq u\}, \quad \text{und} \quad \overline{G}(u) := \inf\{x \in \mathbb{R} \mid F(x) > u\}$$

der Verteilung  $\mu$ , siehe Abschnitt 4.4. Entsprechend seien  $\underline{G}_n$  und  $\overline{G}_n$  die minimalen und maximalen  $u$ -Quantile der Verteilung  $\mu_n$ . Analog zum Beweis von Satz 4.20 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  $\mu$  bzw.  $\mu_n$  sind. Wir zeigen nun, dass aus (2) folgt:

*Behauptung:*  $\underline{G}_n \rightarrow \underline{G}$   $P$ -fast sicher und  $\overline{G}_n \rightarrow \overline{G}$   $P$ -fast sicher.

Damit ist dann 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)  $\underline{G} = \overline{G}$  und  $\underline{G}_n = \overline{G}_n$   $P$ -fast sicher, denn:

$$\begin{aligned} P[\underline{G} \neq \overline{G}] &= P[\underline{G} < \overline{G}] = P\left[\bigcup_{c \in \mathbb{Q}} \{\underline{G} \leq c < \overline{G}\}\right] \\ &\leq \sum_{c \in \mathbb{Q}} P[\{\underline{G} \leq c\} \setminus \{\overline{G} \leq c\}] = \sum_{c \in \mathbb{Q}} \underbrace{(P[\{\underline{G} \leq c\}] - P[\{\overline{G} \leq c\}])}_{=F(c)} = 0. \end{aligned}$$

(c) Wir zeigen nun:

$$\limsup \overline{G}_n(u) \leq \overline{G}(u), \quad \text{und} \quad \liminf \underline{G}_n(u) \geq \underline{G}(u). \quad (8.3.3)$$

Zum Beweis der ersten aussage genügt es zu zeigen, dass

$$\limsup \overline{G}_n(u) \leq c \quad \text{für alle } c > \overline{G}(u) \quad \text{mit } \mu[\{c\}] = 0 \quad (8.3.4)$$

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):

$$u < F(c) = \lim_{n \rightarrow \infty} F_n(c),$$

also existiert ein  $n_0 \in \mathbb{N}$  mit

$$F_n(c) > u \quad \text{für alle } n > n_0. \quad (8.3.5)$$

Aus (8.3.5) folgt

$$\overline{G}_n(u) \leq c \quad \text{für } n \geq n_0,$$

und somit

$$\limsup \overline{G}_n(u) \leq c.$$

Damit haben wir die erste Aussage in (8.3.3) bewiesen. Die zweite Aussage zeigt man auf ähnliche Weise.

(d) Aus (a)-(c) folgt  $P$ -fast sicher:

$$\limsup \underline{G}_n \stackrel{(a)}{\leq} \limsup \overline{G}_n \stackrel{(c)}{\leq} \overline{G} \stackrel{(b)}{=} \underline{G} \stackrel{(3)}{\leq} \liminf \overline{G}_n \stackrel{(a)}{\leq} \liminf \underline{G}_n,$$

also

$$\lim \underline{G}_n = \underline{G} \quad \text{und} \quad \lim \overline{G}_n = \overline{G}.$$

□

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:

### Existenz schwach konvergenter Teilfolgen

Eine Folge von Wahrscheinlichkeitsverteilungen auf einer *endlichen* Menge  $S = \{x_1, \dots, x_d\}$  können wir 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 beweisen nun ein Kriterium für die Existenz schwach konvergenter Teilfolgen für Folgen von Wahrscheinlichkeitsverteilungen auf  $\mathbb{R}^1$ . Dazu setzen wir voraus, dass die Masse nicht ins Unendliche abwandert:

**Definition.** Eine Folge  $\mu_n \in WV(\mathbb{R})$  heißt **straff** (engl. *tight*), falls zu jedem  $\varepsilon > 0$  ein  $c \in (0, \infty)$  existiert mit

$$\mu_n([-c, c]) \geq 1 - \varepsilon \quad \text{für alle } n \in \mathbb{N}.$$

Eine straffe Folge von Wahrscheinlichkeitsverteilungen ist also gleichmäßig auf Kompakta konzentriert. Die Masse kann daher für  $n \rightarrow \infty$  nicht ins Unendliche abwandern.

**Beispiel.** 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  $\sigma_n$  der Mittelwerte und Standardabweichungen beschränkt sind.

**Satz 8.6 (Helly-Bray).** *Jede straffe Folge  $\mu_n \in WV(\mathbb{R})$  hat eine schwach konvergente Teilfolge.*

**Bemerkung.** (1). Das Kriterium lässt sich deutlich verallgemeinern: Eine entsprechende Aussage gilt für Folgen von Wahrscheinlichkeitsverteilungen auf beliebigen vollständigen separablen metrischen Räumen (Satz von Prohorov, siehe z.B. *Billingsley: Convergence of probability measures*). Die endlichen Intervalle  $[-c, c]$  in der Definition von Straffheit ersetzt man in diesem Fall durch kompakte Mengen.

- (2). Der Raum  $WV(\overline{\mathbb{R}})$  aller Wahrscheinlichkeitsverteilungen auf  $[-\infty, \infty]$  ist sogar **kompakt** bezüglich der schwachen Topologie, d.h. **jede** Folge  $\mu_n \in WV(\overline{\mathbb{R}})$  hat eine schwach konvergente Teilfolge. Der Beweis verläuft analog zu dem von Satz 8.6. Es 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. Allgemeiner gilt: Ist  $S$  kompakt, dann ist  $WV(S)$  kompakt bzgl. der schwachen Konvergenz.

Wir beweisen nun den Satz von Helly-Bray:

*Beweis.* Sei  $\mu_n$  ( $n \in \mathbb{N}$ ) eine straffe Folge von Wahrscheinlichkeitsverteilungen 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  $x_1, x_2, \dots$  eine Abzählung von  $\mathbb{Q}$ . Wegen  $0 \leq F_n \leq 1$  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  $\overline{F}(x) := \lim_{k \rightarrow \infty} F_{n_k}(x)$ . Nach (1) existiert der Grenzwert, außerdem ist die Funktion  $\overline{F} : \mathbb{Q} \rightarrow [0, 1]$  Der Limes existiert nach 1. für  $x \in \mathbb{Q}$  und die Funktion  $\overline{F} : \mathbb{Q} \rightarrow [0, 1]$  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  $0 \leq F \leq 1$ .  
 (b)  $F_{n_k}(x) \rightarrow F(x)$  für alle  $x \in \mathbb{R}$ , an denen  $F$  stetig ist.



(3). Aus (a) folgt, dass durch

$$\mu[(a, b]] := F(b) - F(a), \quad -\infty < a \leq b < \infty,$$

ein positives Maß auf  $\mathbb{R}$  definiert wird mit

$$\mu[\mathbb{R}] = \lim_{c \rightarrow \infty} \mu[(-c, c]] \in [0, 1].$$

Wir zeigen nun, dass  $\mu$  eine *Wahrscheinlichkeitsverteilung* auf  $\mathbb{R}$  ist, falls die Folge  $(\mu_n)_{n \in \mathbb{N}}$  *straff* ist. Es gilt nämlich:

$$\mu[(-c, c]] = F(c) - F(-c) = \lim_{k \rightarrow \infty} (F_{n_k}(c) - F_{n_k}(-c)) = \lim_{k \rightarrow \infty} \mu_{n_k}[(-c, c]] \quad (8.3.6)$$

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

$$\mu_{n_k}[(-c, c]] \geq 1 - \varepsilon \quad \text{für alle } k.$$

Aus (8.3.6) folgt dann  $\mu[(-c, c]] \geq 1 - \varepsilon$ , falls  $c$  groß genug ist, und damit für  $\varepsilon \searrow 0$ :

$$\mu[\mathbb{R}] \geq 1, \quad \text{also } \mu(\mathbb{R}) = 1.$$

(4). Aus (b) folgt nun nach Satz 8.5, dass die Folge  $(\mu_{n_k})_{k \in \mathbb{N}}$  schwach gegen  $\mu$  konvergiert.

□

## 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:

**Satz 8.7 (Stetigkeitssatz, Konvergenzsatz von Lévy).** *Seien  $(\mu_n)_{n \in \mathbb{N}}$  Wahrscheinlichkeitsverteilungen auf  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  mit charakteristischen Funktionen*

$$\phi_n(t) = \int e^{itx} \mu_n(dx).$$

Dann gilt:

(1). *Konvergiert  $\mu_n$  schwach gegen eine Wahrscheinlichkeitsverteilung  $\mu$ , dann konvergieren auch die charakteristischen Funktionen:*

$$\phi_n(t) \rightarrow \phi(t) := \int e^{itx} \mu(dx) \quad \text{für alle } t \in \mathbb{R}.$$

- (2). Konvergiert umgekehrt  $\phi_n(t)$  für alle  $t \in \mathbb{R}$  gegen einen Limes  $\phi(t)$ , und ist  $\phi$  stetig bei  $t = 0$ , dann ist  $\phi$  die charakteristische Funktion einer Wahrscheinlichkeitsverteilung  $\mu$ , und  $\mu_n$  konvergiert schwach gegen  $\mu$ .

**Bemerkung.** (1). Die Stetigkeit von  $\phi$  bei 0 ist wesentlich. Zum Beispiel ist die Folge  $\mu_n = N(0, n)$  nicht schwach konvergent, aber die charakteristischen Funktionen konvergieren punktweise:

$$\phi_n(t) = e^{-\frac{t^2}{2n}} \xrightarrow{n \uparrow \infty} \begin{cases} 0 & \text{falls } t \neq 0 \\ 1 & \text{falls } t = 0 \end{cases}.$$

- (2). Eine Aussage wie im Satz gilt auch für Wahrscheinlichkeitsverteilungen auf  $\mathbb{R}^d$ . Hier definiert man die charakteristische Funktion  $\phi : \mathbb{R}^d \rightarrow \mathbb{C}$  durch

$$\phi(t) = \int_{\mathbb{R}^d} e^{it \cdot x} \mu(dx), \quad t \in \mathbb{R}^d.$$

*Beweis.* Der erste Teil der Aussage folgt unmittelbar aus  $e^{itx} = \cos(tx) + i \sin(tx)$ , 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  $\mu_n$  ( $n \in \mathbb{N}$ ) 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  $\lim_{|u| \rightarrow \infty} f(u) = 1$ , existiert eine Konstante  $a > 0$  mit  $f(u) \geq a$  für alle  $|u| \geq 1$ . Damit erhalten wir für  $\varepsilon > 0$ :

$$\begin{aligned} \mu_n \left[ |x| \geq \frac{1}{\varepsilon} \right] &= \mu_n [\{x \in \mathbb{R} \mid |\varepsilon x| \geq 1\}] \leq \frac{1}{a} \int \underbrace{\left( 1 - \frac{\sin \varepsilon x}{\varepsilon x} \right)}_{= \frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} 1 - \cos(xt) dt} \mu_n(dx) \\ & \stackrel{\text{Fubini}}{=} \frac{1}{2a\varepsilon} \int_{-\varepsilon}^{\varepsilon} (1 - \operatorname{Re}(\phi_n(t))) dt \stackrel{n \nearrow \infty}{\text{Lebesgue}} \frac{1}{2a\varepsilon} \cdot \int_{-\varepsilon}^{\varepsilon} (1 - \operatorname{Re}(\phi(t))) dt. \end{aligned} \tag{8.3.7}$$

Sei nun  $\delta > 0$  vorgegeben. Ist  $\varepsilon$  hinreichend klein, dann gilt wegen der vorausgesetzten Stetigkeit von  $\phi$  bei 0:

$$|1 - \operatorname{Re}(\phi(t))| = |\operatorname{Re}(\phi(0) - \phi(t))| \leq \frac{\delta a}{2} \quad \text{für alle } t \in [-\varepsilon, \varepsilon].$$

Also können wir die rechte Seite von (8.3.7) durch  $\delta/2$  abschätzen, und somit existiert ein  $n_0 \in \mathbb{N}$  mit

$$\mu_n \left[ |x| \geq \frac{1}{\varepsilon} \right] \leq \delta \quad \text{für alle } n \geq n_0. \quad (8.3.8)$$

Diese Aussage gilt natürlich auch, falls wir  $\varepsilon$  noch kleiner wählen. Zudem gilt (8.3.8) auch für alle  $n < n_0$ , falls  $\varepsilon$  klein genug ist. Also ist  $\mu_n$  ( $n \in \mathbb{N}$ ) straff.

(2). *Der Grenzwert jeder schwach konvergenten Teilfolge von  $(\mu_n)_{n \in \mathbb{N}}$  hat die charakteristische Funktion  $\phi$ .*

Zum beweis sei  $(\mu_{n_k})_{k \in \mathbb{N}}$  eine Teilfolge von  $(\mu_n)_{n \in \mathbb{N}}$  und  $\mu$  eine Wahrscheinlichkeitsverteilung mit  $\mu_{n_k} \xrightarrow{w} \mu$ . Dann gilt nach dem ersten Teil der Aussage des Satzes:

$$\phi_\mu(t) = \lim_{k \rightarrow \infty} \phi_{n_k}(t) = \phi(t) \quad \text{für alle } t \in \mathbb{R}.$$

(3). *Schwache Konvergenz von  $(\phi_n)_{n \in \mathbb{N}}$ .*

Nach dem Inversionsatz existiert höchstens eine Wahrscheinlichkeitsverteilung  $\mu$  mit charakteristischer Funktion  $\phi$ . Also konvergieren nach (2) alle schwach konvergenten Teilfolgen von  $(\mu_n)_{n \in \mathbb{N}}$  gegen denselben Limes  $\mu$ . Hieraus folgt aber, zusammen mit (1), dass  $(\mu_n)_{n \in \mathbb{N}}$  schwach gegen  $\mu$  konvergiert. Für  $f \in \mathcal{C}_b(S)$  hat nämlich jede Teilfolge von  $\int f d\mu_n$  eine gegen  $\int f d\mu$  konvergente Teilfolge, also gilt  $\int f d\mu_n \rightarrow \int f d\mu$ .

□

## 8.4 Der Zentrale Grenzwertsatz

Wir können nun den in Abschnitt 8.2 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.

## Zentraler Grenzwertsatz für Summen von i.i.d. Zufallsvariablen

**Satz 8.8 (Zentraler Grenzwertsatz – 1. Version).** Seien  $X_1, X_2, \dots \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$  unabhängige, identisch verteilte Zufallsvariablen mit Varianz  $\sigma^2$  und sei

$$S_n = X_1 + \dots + X_n.$$

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

$$\frac{S_n - E[S_n]}{\sigma \cdot \sqrt{n}} \xrightarrow{\mathcal{D}} Z,$$

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 8.12 unten.

(3). Im Fall  $\sigma^2 = 0$  gilt die Aussage auch. Hierbei interpretieren wir das Diracmaß  $\delta_m$  als degenerierte Normalverteilung  $N(m, 0)$ .

Wir beweisen nun den Zentralen Grenzwertsatz in der obenstehenden Form:

*Beweis.* O.B.d.A. sei  $E[X_i] = 0$ , ansonsten betrachten wir die zentrierten Zufallsvariablen  $\tilde{X}_i := X_i - E[X_i]$ . Nach dem Konvergenzsatz von Lévy genügt es zu zeigen, dass die charakteristischen Funktionen der standardisierten Summen  $\hat{S}_n$  punktweise gegen die charakteristische Funktion der Normalverteilung  $N(0, \sigma^2)$  konvergieren, d.h.

$$\phi_{\hat{S}_n}(t) \rightarrow \phi_{N(0, \sigma^2)}(t) = e^{-\frac{\sigma^2 t^2}{2}} \quad \forall t \in \mathbb{R}. \quad (8.4.1)$$

Da die Zufallsvariablen  $X_i$  unabhängig, identisch verteilt und zentriert sind, gilt für  $t \in \mathbb{R}$ :

$$\phi_{\hat{S}_n}(t) \stackrel{E[S_n]=0}{=} \phi_{S_n}\left(\frac{t}{\sqrt{n}}\right) \stackrel{X_i \text{ iid}}{=} \left(\phi_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

Aus  $X_1 \in \mathcal{L}^2$  folgt  $\phi_{X_1} \in C^2(\mathbb{R})$ , und

$$\phi_{X_1}(t) = E[e^{itX_1}] = 1 + itE[X_1] + \frac{(it)^2}{2}E[X_1^2] + o(t^2) = 1 - \frac{t^2\sigma^2}{2} + o(t^2),$$

wobei  $o$  für eine Funktion  $o: \mathbb{R}^+ \rightarrow \mathbb{C}$  mit  $\lim_{\varepsilon \downarrow 0} \frac{|o(\varepsilon)|}{\varepsilon} = 0$  steht. Damit erhalten wir:

$$\phi_{\hat{S}_n}(t) = \left(1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n.$$

Wir vermuten, dass dieser Ausdruck für  $n \rightarrow \infty$  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

$$\begin{aligned} \left| \prod_{i=1}^n z_i - \prod_{i=1}^n w_i \right| &= |(z_1 - w_1)z_2z_3 \cdots z_n + w_1(z_2 - w_2)z_3z_4 \cdots z_n + \dots + w_1 \cdots w_{n-1}(z_n - w_n)| \\ &\leq \sum_{i=1}^n |z_i - w_i|. \end{aligned}$$

Damit erhalten wir:

$$\begin{aligned} \left| \phi_{\hat{S}_n}(t) - \exp\left(-\frac{t^2\sigma^2}{2}\right) \right| &= \left| \left(1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n - \exp\left(-\frac{t^2\sigma^2}{2n}\right)^n \right| \\ &\leq n \cdot \left| 1 - \frac{t^2\sigma^2}{2n} + o\left(\frac{t^2}{n}\right) - \exp\left(-\frac{t^2\sigma^2}{2n}\right) \right|. \end{aligned}$$

Da die rechte Seite für  $n \rightarrow \infty$  gegen 0 konvergiert, folgt (8.4.1) und damit die Behauptung.  $\square$

**Beispiel.** (1). Sind  $X_1, X_2, \dots$  unabhängig mit  $P[X_i = 1] = p$  und  $P[X_i = 0] = 1 - p$ , dann ist  $S_n = \sum_i X_i$  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  $S_n = \sum_i X_i$  Poissonverteilt mit Parameter  $n\lambda$ . Der Zentrale Grenzwertsatz liefert in diesem Fall eine Normalapproximation für Poissonverteilungen mit großer Intensität (Übung).

(3). Sind  $X_1, X_2, \dots$  unabhängige,  $N(m, \sigma^2)$ -verteilte Zufallsvariablen, dann gilt

$$\hat{S}_n = \frac{X_1 + X_2 + \dots + X_n - nm}{\sqrt{n}} \sim N(0, \sigma^2)$$

für alle  $n \in \mathbb{N}$  (und nicht nur asymptotisch!).

**Warum tritt die Normalverteilung im Limes auf?** Wie schon im letzten Beispiel bemerkt, gilt

$$X_i \sim N(0, \sigma^2) \text{ unabhängig} \Rightarrow \frac{X_1 + \dots + X_n}{\sqrt{n}} \sim N(0, \sigma^2).$$

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 8.9.** Sei  $\mu$  eine Wahrscheinlichkeitsverteilung auf  $\mathbb{R}$  mit  $\int x^2 \mu(dx) < \infty$ . Gilt

$$X, Y \sim \mu \text{ unabhängig} \Rightarrow \frac{X + Y}{\sqrt{2}} \sim \mu, \quad (8.4.2)$$

dann ist  $\mu$  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  $X_1, X_2, \dots$  unabhängige Zufallsvariablen mit Verteilung  $\mu$ . Aus der Voraussetzung (8.4.2) folgt  $E[X_i] = \int x \mu(dx) = 0$  für alle  $i \in \mathbb{N}$ , und durch Induktion:

$$\frac{(X_1 + \dots + X_n)}{\sqrt{n}} \sim \mu \quad \text{für } n = 2^k, k \in \mathbb{N}.$$

Wegen  $\int x^2 \mu(dx) < \infty$  sind die  $X_i$  quadratintegrierbar. Durch Anwenden des zentralen Grenzwertsatzes auf die standardisierten Summen folgt, dass  $\mu$  eine zentrierte Normalverteilung ist.  $\square$

## Normalapproximationen

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:

$$X_i \in \mathcal{L}^2 \text{ i.i.d. mit } E[X_i] = m, \text{ Var}[X_i] = \sigma^2.$$

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  $S_n = X_1 + \dots + X_n$ . 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 der 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  $n \rightarrow \infty$  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 untersuchen. 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  $\text{Var}[X_i] = 1$ , und sei  $a > 0$ . Mit einer ZGS-Approximation erhalten wir für große  $n$

$$\begin{aligned} P \left[ \frac{1}{n} \sum_{i=1}^n X_i \geq a \right] &= P \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \geq a\sqrt{n} \right] \\ &\approx \frac{1}{\sqrt{2\pi}} \int_{a\sqrt{n}}^{\infty} e^{-\frac{x^2}{2}} dx \\ &= e^{-\frac{na^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-a\sqrt{ny} - \frac{y^2}{2}} dy \quad (x = a\sqrt{n} + y) \\ &= e^{-\frac{na^2}{2}} \cdot \frac{1}{\sqrt{2\pi n}} \int_0^{\infty} e^{-az - \frac{z^2}{2n}} dz \quad (z = \sqrt{ny}) \\ &\sim \frac{1}{\sqrt{2\pi a^2 n}} \cdot \exp \left( -\frac{na^2}{2} \right) \end{aligned}$$

Dies ist aber **nicht** die korrekte Asymptotik für  $n \rightarrow \infty$ . Auf der exponentiellen Skala gilt nämlich

$$P \left[ \frac{1}{n} \sum_{i=1}^n X_i \geq a \right] \sim \exp(-nI(a)),$$

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 aus dem Beispiel oben trotzdem recht gut funktioniert, wenn die Zufallsvariablen  $X_i$  dritte Momente haben garantiert die folgende **Abschätzung der Konvergenzgeschwindigkeit im Zentralen Grenzwertsatz**:

**Satz 8.10 (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  $\text{Bin}(n, p)$  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  $p \rightarrow 0$  oder  $p \rightarrow 1$  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.



### Heavy Tails, Konvergenz gegen $\alpha$ -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  $X_1, X_2, \dots : \Omega \rightarrow \mathbb{R}$  unabhängige identisch verteilte absolutstetige Zufallsvariablen, deren Dichtefunktion

$$f_{X_i}(x) = |x|^{-\alpha-1} \quad \text{für alle } |x| \geq r$$

erfüllt. Da die Dichte für  $|x| \rightarrow \infty$  nur langsam abfällt, sind die Zufallsvariablen nicht quadratintegrierbar; sie sind aber integrierbar. Daher ergibt sich ein anderes asymptotisches Verhalten der charakteristischen Funktionen für  $t \rightarrow 0$ :

**Lemma 8.11.** Für  $t \rightarrow 0$  gilt

$$\phi_{X_i}(t) = 1 + imt - c|t|^\alpha + O(t^2)$$

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

$$\phi_{S_n}(t) = (1 - c|t|^\alpha + O(t^2))^n.$$

Um Konvergenz der charakteristischen Funktionen zu erhalten, müssen wir  $X_n$  nun mit  $n^{-1/\alpha}$  statt  $n^{-1/2}$  reskalieren:

$$\begin{aligned} \phi_{n^{-1/\alpha} S_n}(t) &= \phi_{S_n}(n^{-1/\alpha} t) = (1 - c|t|^\alpha n^{-1} + O(n^{-2/\alpha}))^n \\ &\rightarrow \exp(-c|t|^\alpha) \quad \text{für } n \rightarrow \infty. \end{aligned}$$

Nach dem Konvergenzsatz von Lévy folgt:

**Satz 8.12.** Für  $n \rightarrow \infty$  gilt

$$n^{-1/\alpha} S_n \xrightarrow{\mathcal{D}} \mu_{c,\alpha},$$

wobei  $\mu_{c,\alpha}$  die Wahrscheinlichkeitsverteilung mit charakteristischer Funktion

$$\phi_{c,\alpha}(t) = \exp(-c|t|^\alpha)$$

ist.

**Definition.** Seien  $\alpha \in (0, 2]$  und  $m \in \mathbb{R}$ . Die Wahrscheinlichkeitsverteilungen mit charakteristischer Funktion

$$\phi(t) = \exp(imt - c|t|^\alpha),$$

$c \in (0, \infty)$  heißen **symmetrische  $\alpha$ -stabile Verteilungen** mit Mittelwert  $m$ .

Die Dichten der  $\alpha$ -stabilen Verteilungen sind für  $\alpha \neq 1, 2$  nicht explizit berechnbar, fallen aber für  $|x| \rightarrow \infty$  wie  $|x|^{-\alpha-1}$  ab. Für  $\alpha = 1$  erhält man die Cauchyverteilungen, für  $\alpha = 2$  die Normalverteilungen. Satz 8.12 ist ein Spezialfall eines allgemeineren Grenzwertsatzes für Summen von zufallsvariablen mit polynomiellen Tails, siehe z.B. BREIMAN, THEOREM 9.34.

## Der Satz von Lindeberg-Feller

Wir wollen nun die Annahme fallenlassen, dass die Summanden  $X_i$  identisch verteilt sind, und zeigen, dass trotzdem ein zentraler Grenzwertsatz gilt. Sei

$$\widehat{S}_n = Y_{n,1} + Y_{n,2} + \dots + Y_{n,n} \quad \text{mit } Y_{n,i} \in \mathcal{L}^2(\Omega, \mathcal{A}, P).$$

Die Zufallsvariablen  $Y_{n,i}$  können etwa kleine Störungen oder Meßfehler beschreiben. Setzen wir

$$Y_{n,i} = \frac{X_i - E[X_i]}{\sqrt{n}} \quad \text{mit } X_i \in \mathcal{L}^2 \text{ unabhängig,} \quad (8.4.3)$$

so erhalten wir das Setup von oben.

**Satz 8.13 (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)  $\text{Var}[\widehat{S}_n] = \sum_{i=1}^n \text{Var}[Y_{n,i}] \xrightarrow{n \uparrow \infty} \sigma^2$ ,
- (iii)  $\gamma_{n,\varepsilon} := \sum_{i=1}^n E[Y_{n,i}^2; |Y_{n,i}| > \varepsilon] \xrightarrow{n \uparrow \infty} 0 \quad \forall \varepsilon > 0$ .

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 die 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: Sind  $X_i \in \mathcal{L}^2$  i.i.d. Zufallsvariablen mit  $E[X_i] = m$  und  $\text{Var}[X_i] = \sigma^2$ , und definieren wir  $Y_{n,i}$  wie in (8.4.3), dann gilt:

$$\text{Var}[\widehat{S}_n] = \frac{1}{n} \sum_{i=1}^n \text{Var}[X_i] = \text{Var}[X_1] = \sigma^2, \quad \text{für alle } n \in \mathbb{N},$$

und, für  $\varepsilon > 0$

$$\begin{aligned} \gamma_{n,\varepsilon} &= \sum_{i=1}^n E[Y_{n,i}^2; |Y_{n,i}| > \varepsilon] = \frac{1}{n} \sum_{i=1}^n E[|X_i - m|^2; |X_i - m| > \varepsilon\sqrt{n}] \\ &= E[|X_1 - m|^2; |X_1 - m| > \varepsilon\sqrt{n}] \rightarrow 0 \quad \text{für } n \rightarrow \infty, \end{aligned}$$

da  $X_1$  quadratintegrierbar ist.

(2). Die Bedingung (iii) ist insbesondere erfüllt, wenn

$$\sum_{i=1}^n E[|Y_{n,i}|^p] \xrightarrow{n \rightarrow \infty} 0 \quad \text{für ein } p > 2 \text{ gilt,}$$

denn für  $\varepsilon > 0$  ist  $E[Y_{n,i}^2; |Y_{n,i}| \geq \varepsilon] \leq \varepsilon^{2-p} E[|Y_{n,i}|^p]$ .

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

$$e^{ix} = 1 + ix - \frac{x^2}{2} + R(x) \quad \text{mit} \quad |R(x)| \leq \min\left(\frac{|x|^3}{6}, x^2\right). \quad (8.4.4)$$

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

$$|R_{n,i}| \leq E \left[ \min \left( \frac{|tY_{n,i}|^3}{6}, t^2 Y_{n,i}^2 \right) \right] \quad (8.4.5)$$

gilt.

(II) Wir zeigen  $\sum_{i=1}^n |R_{n,i}| \rightarrow 0$  für  $n \rightarrow \infty$ :

Für  $\varepsilon > 0$  gilt nach (8.4.5):

$$|R_{n,i}| \leq \frac{1}{6} \cdot E[|tY_{n,i}|^3; |Y_{n,i}| \leq \varepsilon] + E[|tY_{n,i}|^2; |Y_{n,i}| > \varepsilon].$$

Mit  $E[|tY_{n,i}|^3; |Y_{n,i}| \leq \varepsilon] \leq |t|^3 \varepsilon \cdot \sigma_{n,i}^2$  erhalten wir

$$\sum_{i=1}^n |R_{n,i}| \leq \frac{|t|^3 \varepsilon}{6} \sum_{i=1}^n \sigma_{n,i}^2 + t^2 \gamma_{n,\varepsilon},$$

und somit nach Voraussetzung (ii) und (iii)

$$\limsup_{n \rightarrow \infty} \sum_{i=1}^n |R_{n,i}| \leq \frac{\sigma^2 |t|^3}{6} \varepsilon.$$

Die Behauptung folgt für  $\varepsilon \rightarrow 0$ .

(III) Wir zeigen  $\sup_{1 \leq i \leq n} \sigma_{n,i}^2 \rightarrow 0$  für  $n \rightarrow \infty$ :

Für  $\varepsilon > 0$  und  $1 \leq i \leq n$  gilt

$$\sigma_{n,i}^2 = E[Y_{n,i}^2; |Y_{n,i}| \leq \varepsilon] + E[Y_{n,i}^2; |Y_{n,i}| > \varepsilon] \leq \varepsilon^2 + \gamma_{n,\varepsilon}$$

Wegen  $\gamma_{n,\varepsilon} \rightarrow 0$  für  $n \rightarrow \infty$  ergibt sich

$$\limsup_{n \rightarrow \infty} \sup_{1 \leq i \leq n} \sigma_{n,i}^2 \leq \varepsilon^2.$$

Die Behauptung folgt wieder für  $\varepsilon \rightarrow 0$ .

(b) **Hauptteil des Beweises:** Zu zeigen ist

$$\phi_{\hat{S}_n}(t) = \prod_{i=1}^n \phi_{n,i}(t) \xrightarrow{n \rightarrow \infty} \exp\left(-\frac{t^2 \sigma^2}{2}\right), \quad (8.4.6)$$

die Aussage folgt dann aus dem Konvergenzsatz von Lévy.

Wir zeigen:

$$\left| \prod_{i=1}^n \phi_{n,i}(t) - \prod_{i=1}^n \left( 1 - \frac{t^2 \sigma_{n,i}^2}{2} \right) \right| \xrightarrow{n \rightarrow \infty} 0, \quad \text{und} \quad (8.4.7)$$

$$\prod_{i=1}^n \left( 1 - \frac{t^2 \sigma_{n,i}^2}{2} \right) \xrightarrow{n \rightarrow \infty} e^{-\frac{t^2 \sigma^2}{2}}. \quad (8.4.8)$$

Daraus folgt (8.4.6), und damit die Behauptung.

*Beweis von (8.4.7):* 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

$$1 - \frac{t^2 \sigma_{n,i}^2}{2} \in (0, 1) \quad \text{für alle } n \geq n_0 \text{ und } 1 \leq i \leq n. \quad (8.4.9)$$

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 (8.4.8):* Wegen (8.4.9) erhalten wir

$$\begin{aligned} \log \left( \prod_{i=1}^n \left( 1 - \frac{t^2 \sigma_{n,i}^2}{2} \right) \right) &= \sum_{i=1}^n \log \left( 1 - \frac{t^2 \sigma_{n,i}^2}{2} \right) \\ &= - \sum_{i=1}^n \frac{t^2 \sigma_{n,i}^2}{2} + \sum_{i=1}^n \tilde{R}_{n,i}, \end{aligned}$$

wobei  $|\tilde{R}_{n,i}| \leq C \cdot (t^2 \sigma_{n,i}^2)^2$  mit  $\tilde{C} \in (0, \infty)$ . Die rechte Seite konvergiert nach Voraussetzung (ii) für  $n \rightarrow \infty$  gegen  $-\frac{t^2 \sigma^2}{2}$ , denn

$$\sum_{i=1}^n |\tilde{R}_{n,i}| \leq Ct^4 \cdot \sum_{i=1}^n \sigma_{n,i}^4 \leq Ct^4 \cdot \sum_{i=1}^n \sigma_{n,i}^2 \cdot \sup_{1 \leq i \leq n} \sigma_{n,i}^2 \rightarrow 0$$

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.

# Kapitel 9

## Multivariate Verteilungen und statistische Anwendungen

### 9.1 Mehrstufige Modelle

Seien  $(S_i, \mathcal{S}_i)$ ,  $1 \leq i \leq n$ , meßbare Räume. Wir wollen allgemeine Wahrscheinlichkeitsverteilungen auf dem Produktraum  $S_1 \times \dots \times S_n$  konstruieren und effektiv beschreiben. In Analogie zu diskreten, mehrstufigen Modellen versuchen wir diese in der Form

$$P(dx_1 \dots dx_n) = \mu(dx_1) p(x_1, dx_2) p((x_1, x_2), dx_3) \cdots p((x_1, \dots, x_{n-1}), dx_n)$$

darzustellen.

#### 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  $(T, \mathcal{T})$  messbare Räume und sei

$$\Omega := S \times T \quad \text{und} \quad \mathcal{A} := \mathcal{S} \otimes \mathcal{T} \quad \text{die produkt-}\sigma\text{-Algebra.}$$

Unser Ziel ist die Konstruktion einer Wahrscheinlichkeitsverteilung auf  $(\Omega, \mathcal{A})$  vom Typ

$$P(dxdy) = \mu(dx) p(x, dy).$$

**Definition.** *Eine Abbildung*

$$p : S \times \mathcal{T} \longrightarrow [0, 1], \quad (x, C) \mapsto p(x, C),$$

heißt **stochastischer Kern** (oder **Übergangswahrscheinlichkeit**), wenn gilt:

(i)  $p(x, \bullet)$  ist für jedes  $x \in S$  eine Wahrscheinlichkeitsverteilung auf  $(T, \mathcal{T})$ ,

(ii)  $p(\bullet, C)$  ist für jedes  $C \in \mathcal{T}$  eine meßbare Funktion auf  $(S, \mathcal{S})$ .

**Bemerkung (Diskreter Spezialfall).** Sind  $S$  und  $T$  abzählbar mit  $\mathcal{S} = \mathcal{P}(S)$ ,  $\mathcal{T} = \mathcal{P}(T)$ , dann ist  $p$  eindeutig festgelegt durch die Matrix mit Komponenten

$$p(x, y) := p(x, \{y\}) \quad (x \in S, y \in T).$$

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  $\mu$  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 9.1 (Fubini).** Sei  $\mu(dx)$  eine Wahrscheinlichkeitsverteilung auf  $(S, \mathcal{S})$  und  $p(x, dy)$  ein stochastischer Kern von  $(S, \mathcal{S})$  nach  $(T, \mathcal{T})$ . Dann existiert eine eindeutige Wahrscheinlichkeitsverteilung  $\mu \otimes p$  auf  $(\Omega, \mathcal{A})$  mit

$$(\mu \otimes p)[B \times C] = \int_B \mu(dx) p(x, C) \quad \text{für alle } B \in \mathcal{S}, C \in \mathcal{T}. \quad (9.1.1)$$

Für diese Wahrscheinlichkeitsverteilung gilt:

$$\int f d(\mu \otimes p) = \int \left( \int f(x, y) p(x, dy) \right) \mu(dx) \quad \text{für alle } \mathcal{A}\text{-messbaren } f : \Omega \rightarrow \mathbb{R}_+. \quad (9.1.2)$$

**Beweis.** (1). *Eindeutigkeit:* Das Mengensystem  $\{B \times C \mid B \in \mathcal{S}, C \in \mathcal{T}\}$  ist ein durchschnittsstabiler Erzeuger der Produkt- $\sigma$ -Algebra  $\mathcal{A}$ . Also ist die Wahrscheinlichkeitsverteilung  $\mu \otimes \nu$  durch (9.1.1) eindeutig festgelegt.

(2). *Existenz:* Wir wollen die Wahrscheinlichkeitsverteilung  $\mu \otimes \nu$  über (9.1.2) 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  $A = B \times C$  ( $B \in \mathcal{S}, C \in \mathcal{T}$ ) 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 (9.1.2) erfüllt. □

Als nächstes wollen wir die **Randverteilungen** des gerade konstruierten zweistufigen Modells berechnen. Sei also  $P := \mu \otimes p$ , und seien

$$\begin{aligned} X : S \times T &\rightarrow S & , & \quad Y : S \times T \rightarrow T \\ (x, y) &\mapsto x & \quad & \quad (x, y) \mapsto y' \end{aligned}$$

die Projektionen auf die 1. bzw. 2. Komponente. Wegen  $p(x, T) = 1$  gilt:

$$P[X \in B] = P[B \times T] = \int_B \mu(dx) p(x, T) = \mu[B] \quad \forall B \in \mathcal{S},$$

also ist die Verteilung  $P \circ X^{-1}$  der ersten Komponente gleich  $\mu$ . Für die Verteilung der zweiten Komponente erhalten wir

$$P[Y \in C] = P[S \times C] = \int_S \mu(dx) p(x, C) \quad \forall C \in \mathcal{T}.$$

**Definition.** Die durch

$$(\mu p)[C] := \int \mu(dx) p(x, C), \quad C \text{ in } \mathcal{T},$$

definierte Wahrscheinlichkeitsverteilung auf  $(T, \mathcal{T})$  heißt **Mischung** der Wahrscheinlichkeitsverteilungen  $p(x, \bullet)$  bezüglich  $\mu$ .

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  $\mu p$  die schon in Abschnitt 2.3 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  $\mu p$  sind also das Tensor- bzw. Matrixprodukt des Zeilenvektors  $\mu$  und der stochastischen Matrix  $p$ .

### Wichtige Spezialfälle

**Produktmaße:** Ist  $p(x, \bullet) \equiv \nu$  eine feste (von  $x$  unabhängige) Wahrscheinlichkeitsverteilung auf  $(T, \mathcal{T})$ , dann ist  $\mu \otimes p$  das Produkt  $\mu \otimes \nu$  der Wahrscheinlichkeitsverteilungen  $\mu$  und  $\nu$ . Der Satz von Fubini liefert also die Existenz des Produktmaßes, und die schon mehrfach verwendete Berechnungsformel

$$\int f d(\mu \otimes \nu) = \int_S \left( \int_T f(x, y) \nu(dy) \right) \mu(dx) \quad (9.1.3)$$

für die Integrale nicht-negativer oder integrierbarer messbarer Funktionen bzgl. des Produktmaßes. Die Integrationsreihenfolge kann man in diesem Fall vertauschen, denn wegen

$$(\mu \otimes \nu)[B \times C] = \mu[B] \nu[C] \quad \text{für alle } B \in \mathcal{S}, C \in \mathcal{T} \quad (9.1.4)$$

gilt  $(\nu \otimes \mu) \circ R^{-1} = \mu \otimes \nu$ , wobei  $R(x, y) = (y, x)$ , und damit nach dem Transformationssatz:

$$\begin{aligned}\int \left( \int f(x, y) \mu(dx) \right) \nu(dy) &\stackrel{\text{Fub.}}{=} \int f \circ R d(\nu \otimes \mu) \\ &= \int f d(\mu \otimes \nu) \\ &\stackrel{\text{Fub.}}{=} \int \left( \int f(x, y) \nu(dy) \right) \mu(dx).\end{aligned}$$

Durch wiederholte Anwendung dieses Arguments erhalten wir zudem:

**Korollar 9.2.** Seien  $(S_i, \mathcal{S}_i, \mu_i)$  Wahrscheinlichkeitsräume ( $1 \leq i \leq n$ ). Dann existiert eine eindeutige Wahrscheinlichkeitsverteilung  $\mu_1 \otimes \dots \otimes \mu_n$  auf  $(S_1 \times \dots \times S_n, \mathcal{S}_1 \otimes \dots \otimes \mathcal{S}_n)$  mit:

$$(\mu_1 \otimes \dots \otimes \mu_n)[B_1 \times \dots \times B_n] = \prod_{i=1}^n \mu_i[B_i] \quad \text{für alle } B_i \in \mathcal{S}_i \quad (1 \leq i \leq n).$$

Für alle produktmessbaren Funktionen  $f : S_1 \times \dots \times S_n \rightarrow [0, \infty)$  gilt:

$$\int f d(\mu_1 \otimes \dots \otimes \mu_n) = \int \dots \left( \int f(x_1, \dots, x_n) \mu_n(dx_n) \right) \dots \mu_1(dx_1),$$

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  $n = 2$ .  $\square$

**Deterministische Kopplung:** Gilt  $p(x, \bullet) = \delta_{f(x)}$  für eine messbare Funktion  $f : S \rightarrow T$ , 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  $\mu$  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  $\mu$  auf  $(S, \mathcal{S})$  heißt **Gleichgewicht (stationäre oder auch invariante Verteilung)** von  $p$ , falls  $\mu p = \mu$  gilt, d.h. falls

$$\int \mu(dx) p(x, B) = \mu[B] \quad \text{für alle } B \in \mathcal{S}.$$

**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. Für  $\alpha \geq 1$  existiert kein Gleichgewicht.

## Bedingte Dichten und Bayessche Formel

Wir betrachten nun Situationen mit nichttrivialer Abhängigkeit zwischen den Komponenten im kontinuierlichen Fall. Seien  $X : \Omega \rightarrow \mathbb{R}^n$  und  $Y : \Omega \rightarrow \mathbb{R}^m$  Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ , deren gemeinsame Verteilung absolutstetig ist mit Dichte  $f_{X,Y}$ , d.h.

$$P[x \in B, Y \in C] = \int_B \int_C f_{X,Y}(x, y) dy dx \quad \text{für alle } B \in \mathcal{B}(\mathbb{R}^n), C \in \mathcal{B}(\mathbb{R}^m).$$

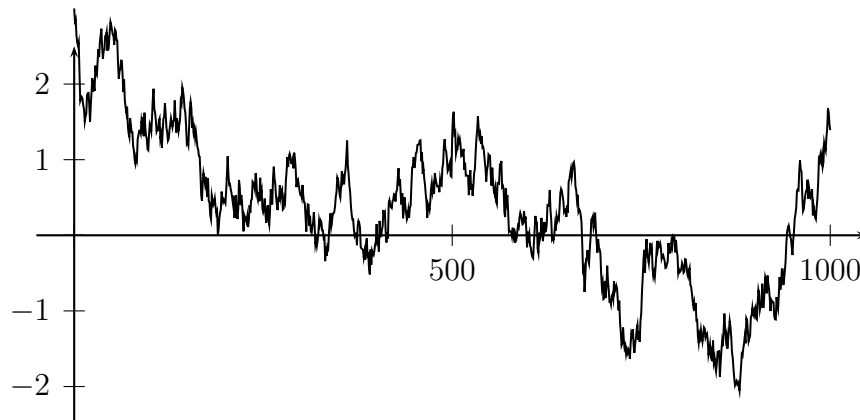


Abbildung 9.1: Simulation einer Trajektorie eines AR(1)-Prozesses mit Parametern  $\alpha = 0.8$  und  $\varepsilon^2 = 1.5$ .

Nach dem Satz von Fubini sind dann auch die Verteilungen von  $X$  und  $Y$  absolutstetig mit dichten

$$f_X(x) = \int_{\mathbb{R}^m} f_{X,Y}(x, y) dy$$

und

$$f_Y(y) = \int_{\mathbb{R}^n} f_{X,Y}(x, y) dx.$$

Obwohl bedingte Wahrscheinlichkeiten gegeben  $Y = y$  nicht im herkömmlichen Sinn definiert werden können, da das Ereignis  $\{Y = y\}$  eine Nullmenge ist, können wir die bedingte Dichte und die bedingte Verteilung von  $X$  gegeben  $Y$  in diesem Fall sinnvoll definieren. Anschaulich beträgt die Wahrscheinlichkeit, dass der Wert  $X$  in einem infinitesimal kleinen Volumenelement  $dx$  liegt, gegeben, dass der Wert von  $Y$  in einem entsprechenden infinitesimalen Volumenelement  $dy$  liegt:

$$\begin{aligned} P[X \in dx | Y \in dy] &= \frac{P[X \in dx, Y \in dy]}{P[Y \in dy]} = \frac{f_{X,Y}(x, y) dx dy}{f_Y(y) dy} \\ &= \frac{f_{X,Y}(x, y)}{f_Y(y)} dx \end{aligned}$$

Diese heuristische Überlegung motiviert die folgende Definition:

**Definition.** Die Funktion  $f_{X|Y} : \mathbb{R}^n \times \mathbb{R}^m \rightarrow [0, \infty]$  mit

$$f_{X|Y} = \begin{cases} \frac{f_{X,Y}(x, y)}{f_Y(y)} & \text{falls } f_Y(y) \neq 0 \\ f_X(x) & \text{falls } f_Y(y) = 0 \end{cases}$$

heißt **bedingte Dichte von  $X$  gegeben  $Y$** , und die von  $y$  abhängende Wahrscheinlichkeitsverteilung

$$\mu_{X|Y}(y, B) := \int_B f_{X|Y}(x, y) dx, \quad \text{für } B \in \mathcal{B}(\mathbb{R}^n),$$

heißt **bedingte Verteilung von  $X$  gegeben  $Y$** .

**Bemerkung.** (1). Für festes  $y$  ist die bedingte Dichte eine Wahrscheinlichkeitsdichte auf  $\mathbb{R}^n$ .

Da  $f_{X|Y}$  produktmessbar ist, ist die bedingte Verteilung  $\mu_{X|Y}$  nach dem Satz von Fubini ein *stochastischer Kern* von  $\mathbb{R}^m$  nach  $\mathbb{R}^n$ .

(2). Auf der Nullmenge  $\{y \in \mathbb{R}^m | f_Y(y) = 0\}$  sind  $f_{X|Y}(x|y)$  und  $\mu_{X|Y}(y, dx)$  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 9.3 (Bayessche Formel).** Für  $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$  mit  $f_X(x) > 0$  und  $f_Y(y) > 0$  gilt

$$f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{\int_{\mathbb{R}^n} f_X(x)f_{Y|X}(y|x) dx}.$$

*Beweis.* Aus der Definition folgt

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{f_{X,Y}(x, y)}{\int_{\mathbb{R}^n} f_{X,Y}(x, y) dx},$$

und damit die Behauptung. □

In Modellen der Bayesschen Statistik interpretiert man  $f_X(x)$  als Dichte der *a priori* angenommenen Verteilung eines unbekanntem 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

$$f_{X|Y}(x|y) = \text{const.}(y) \cdot f_X(x) \cdot f_{Y|X}(y|x)$$

$$\text{A posteriori Dichte} \propto \text{A priori Dichte} \times \text{Likelihood}$$

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

$$S \times T = \mathbb{R}^2 = \{(x, y) \mid x, y \in \mathbb{R}\}.$$

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 Bayes'schen 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 dann:

$$\begin{aligned} \mu(dx) &= f_X(x) \lambda(dx) \\ p(x, dy) &= f_{Y|X}(y|x) \lambda(dy) \end{aligned}$$

mit den Dichten

$$\begin{aligned} f_X(x) &:= \frac{1}{\sqrt{2\pi v}} e^{-\frac{x^2}{2v}} && \text{(Dichte der Verteilung der ersten Komponente } X), \\ f_{Y|X}(y|x) &:= \frac{1}{\sqrt{2\pi\varepsilon}} e^{-\frac{(y-x)^2}{2\varepsilon}} && \text{(bedingte Dichte der zweiten Komponente } Y \text{ gegeben } X = x). \end{aligned}$$

Die gemeinsame Verteilung von Signal und Beobachtungswert ist

$$\begin{aligned} (\mu \otimes p)(dxdy) &= \mu(dx) p(x, dy) \\ &= \frac{1}{2\pi\sqrt{v\varepsilon}} \exp\left(-\frac{(\varepsilon + v)x^2 - 2vxy + vy^2}{2v\varepsilon}\right) \lambda(dx)\lambda(dy) \\ &= \frac{1}{2\pi\sqrt{\det C}} \exp\left(-\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix} \cdot C^{-1} \begin{pmatrix} x \\ y \end{pmatrix}\right) \lambda^2(dx dy). \end{aligned}$$

D.h.  $\mu \otimes p$  ist eine zweidimensionale Normalverteilung mit Kovarianzmatrix

$$C = \begin{pmatrix} v & v \\ v & v + \varepsilon \end{pmatrix}.$$

Mit anderen Worten: Die gemeinsame Verteilung von  $X$  und  $Y$  ist absolutstetig bzgl. des zweidimensionalen Lebesguemaßes mit Dichte

$$f_{X,Y}(x, y) = f_X(x)f_{Y|X}(y|x) = \frac{1}{2\pi\sqrt{\det C}} \exp\left(-\frac{1}{2} \begin{pmatrix} x \\ y \end{pmatrix}^\top \cdot C^{-1} \begin{pmatrix} x \\ y \end{pmatrix}\right).$$

Als Dichte der Verteilung  $\mu_p$  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  $y$ :

$$\begin{aligned} f_{X|Y}(x|y) &:= \frac{f_{X,Y}(x, y)}{f_Y(y)} \\ &= \frac{f_X(x)f_{Y|X}(y|x)}{\int f_X(x)f_{Y|X}(y|x) \lambda(dx)} \\ &= \text{const}(y) \cdot \exp\left(-\frac{\varepsilon + v}{2v\varepsilon} \left(x - \frac{v}{v + \varepsilon}y\right)^2\right). \end{aligned} \quad (9.1.5)$$

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.

## 9.2 Summen unabhängiger Zufallsvariablen, Faltung

Seien  $X$  und  $Y$  unabhängige reellwertige Zufallsvariablen auf  $(\Omega, \mathcal{A}, P)$  mit Verteilungen  $\mu$  bzw.  $\nu$ . Wir wollen die Verteilung von  $X + Y$  bestimmen. Für diskrete Zufallsvariablen ergibt sich:

$$P[X + Y = z] = \sum_{x \in X(\Omega)} \underbrace{P[X = x, Y = z - x]}_{=P[X=x] \cdot P[Y=z-x]} = \sum_{x \in X(\Omega)} \mu(x)\nu(z - x) \quad (9.2.1)$$

Die Wahrscheinlichkeitsverteilung mit Massenfunktion

$$(\mu \star \nu)(z) = \sum_{x \in X(\Omega)} \mu(x) \nu(z - x)$$

heißt Faltung von  $\mu$  und  $\nu$ . Eine entsprechende Aussage erhält man auch im allgemeinen Fall:

### Verteilungen von Summen unabhängiger Zufallsvariablen

**Satz 9.4.** Seien  $X$  und  $Y$  unabhängige reellwertige Zufallsvariablen mit Verteilungen  $\mu$  bzw.  $\nu$ . Dann ist die Verteilung von  $X + Y$  die durch

$$(\mu \star \nu)[B] := \int \mu(dx) \nu[B - x], \quad B \in \mathcal{B}(\mathbb{R}),$$

definierte **Faltung** der Wahrscheinlichkeitsverteilungen  $\mu$  und  $\nu$ .

*Beweis.* Sei  $\tilde{B} := \{(x, y) \mid x + y \in B\}$ . Da  $X$  und  $Y$  unabhängig sind, erhalten wir mit dem Satz von Fubini

$$\begin{aligned} P[X + Y \in B] &= P[(X, Y) \in \tilde{B}] = (\mu \otimes \nu)[\tilde{B}] \\ &\stackrel{\text{Fubini}}{=} \int \mu(dx) \int \nu(dy) \underbrace{I_B(x + y)}_{=I_{B-x}(y)} = \int \mu(dx) \nu[B - x]. \end{aligned}$$

□

**Bemerkung.** Die Faltung  $\mu \star \nu$  zweier Wahrscheinlichkeitsverteilungen  $\mu$  und  $\nu$  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 9.4 dieselben Eigenschaften:

$$\mu \star \nu = \nu \star \mu \quad (\text{da } X + Y = Y + X) \quad (9.2.2)$$

$$(\mu \star \nu) \star \eta = \mu \star (\nu \star \eta) \quad (\text{da } (X + Y) + Z = X + (Y + Z)). \quad (9.2.3)$$

Im diskreten Fall ist  $\mu \star \nu$  nach (9.2.2) die Wahrscheinlichkeitsverteilung mit Gewichten

$$(\mu \star \nu)(z) = \sum_x \mu(x) \nu(z - x).$$

Eine entsprechende Berechnungsformel ergibt sich auch für absolutstetige Wahrscheinlichkeitsverteilungen:



**Lemma 9.5.** *Ist  $\nu$  absolutstetig mit Dichte  $g$ , dann ist auch  $\mu \star \nu$  absolutstetig mit Dichte*

$$\varrho(z) = \int \mu(dx) g(z-x).$$

*Ist zusätzlich auch  $\mu$  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  $\varrho$ . Die zweite Behauptung folgt unmittelbar.  $\square$

**Beispiel.** (1). Sind  $X$  und  $Y$  unabhängig, und  $\text{Bin}(n, p)$  bzw.  $\text{Bin}(m, p)$ -verteilt, dann ist  $X+Y$  eine  $\text{Bin}(n+m, p)$ -verteilte Zufallsvariable. Zum Beweis bemerkt man, daß die gemeinsame Verteilung von  $X$  und  $Y$  mit der gemeinsamen Verteilung von  $Z_1 + \dots + Z_n$  und  $Z_{n+1} + \dots + Z_{n+m}$  übereinstimmt, wobei die Zufallsvariablen  $Z_i$  ( $1 \leq i \leq n+m$ ) unabhängig und Bernoulli( $p$ )-verteilt sind. Also folgt:

$$\mu_{X+Y} = \mu_{Z_1+\dots+Z_n+Z_{n+1}+\dots+Z_{n+m}} = \text{Bin}(n+m, p) .$$

Als Konsequenz erhalten wir (ohne zu rechnen):

$$\text{Bin}(n, p) \star \text{Bin}(m, p) = \text{Bin}(n+m, p) ,$$

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.

$$\sum_{k=0}^l \binom{n}{k} \binom{m}{l-k} = \binom{n+m}{l} . \quad (9.2.4)$$

Die kombinatorische Formel (9.2.4) ist auch als *Vandermonde-Identität* bekannt.

- (2). Sind  $X$  und  $Y$  unabhängig und Poisson-verteilt mit Parametern  $\lambda$  bzw.  $\tilde{\lambda}$ , dann ist  $X + Y$  Poisson-verteilt mit Parameter  $\lambda + \tilde{\lambda}$ , denn nach der Binomischen Formel gilt für  $n \geq 0$ :

$$\begin{aligned} (\mu_X \star \mu_Y)(n) &= \sum_{k=0}^n \mu_X(k) \cdot \mu_Y(n-k) \\ &= \sum_{k=0}^n \frac{\lambda^k}{k!} e^{-\lambda} \cdot \frac{\tilde{\lambda}^{n-k}}{(n-k)!} e^{-\tilde{\lambda}} \\ &= e^{-\lambda+\tilde{\lambda}} \cdot \sum_{k=0}^n \frac{\lambda^k \tilde{\lambda}^{n-k}}{k! (n-k)!} \\ &= e^{-\lambda+\tilde{\lambda}} \cdot \frac{(\lambda + \tilde{\lambda})^n}{n!}. \end{aligned}$$

Also bilden auch die Poissonverteilungen eine Faltungshalbgruppe:

$$\text{Poisson}(\lambda) \star \text{Poisson}(\tilde{\lambda}) = \text{Poisson}(\lambda + \tilde{\lambda})$$

- (3). Sind  $X$  und  $Y$  unabhängig und normalverteilt mit Parametern  $(m, \sigma^2)$  bzw.  $(\tilde{m}, \tilde{\sigma}^2)$ , dann ist  $X + Y$  normalverteilt mit Parametern  $(m + \tilde{m}, \sigma^2 + \tilde{\sigma}^2)$ , siehe ???. Dies verifiziert man leicht mithilfe der charakteristischen Funktionen. Die Normalverteilungen bilden also eine zweiparametrische Faltungshalbgruppe.

## Wartezeiten, Gamma-Verteilung

] Seien  $T_1, T_2, \dots$  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

$$S_n = T_1 + \dots + T_n$$

bis zum  $n$ -ten Ereignis ist dann

$$\mu_{S_n} = \mu_{T_1} \star \mu_{T_2} \star \dots \star \mu_{T_n}.$$

Insbesondere ist die Verteilung von  $S_2$  absolutstetig mit Dichte

$$(f \star f)(s) = \int_{\mathbb{R}} \underbrace{f(x)}_{=0 \text{ für } x < 0} \underbrace{f(s-x)}_{=0 \text{ für } x > s} = \int_0^s \lambda^2 e^{-\lambda x} e^{-\lambda(s-x)} dx = \lambda^2 e^{-\lambda s} \int_0^s dx = \lambda^2 s e^{-\lambda s}$$

für  $s \geq 0$ , bzw.  $(f \star f)(s) = 0$  für  $s < 0$ . Durch Induktion ergibt sich allgemein:

**Lemma 9.6.** 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

$$\Gamma(n) := \int_0^\infty t^{n-1} e^{-t} dx \stackrel{n \in \mathbb{N}}{=} (n-1)! .$$

**Definition.** Die Wahrscheinlichkeitsverteilung auf  $\mathbb{R}_+$  mit Dichte  $f_{\lambda,n}$  heißt **Gammaverteilung** mit Parametern  $\lambda, n \in (0, \infty)$ .

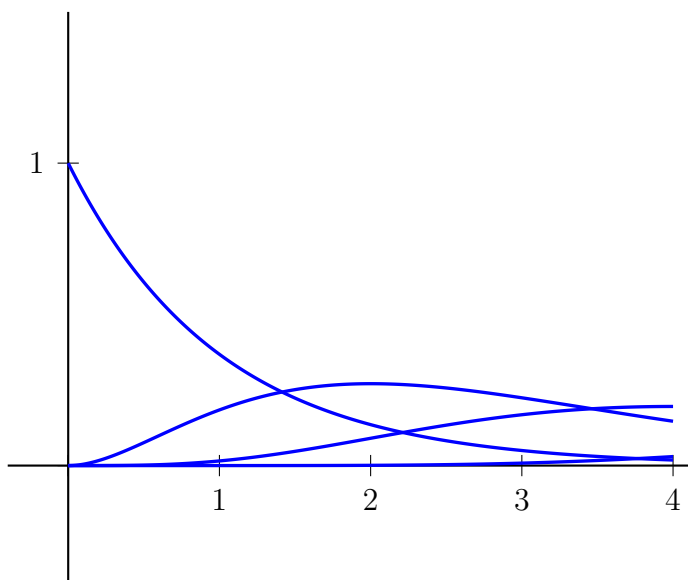


Abbildung 9.2: Dichtefunktionen der Gammaverteilung  $\Gamma_{1,n}$  für verschiedene  $n$ .

Die Gammaverteilung ist auch für nicht-ganzzahlige  $n$  definiert,  $\Gamma$  ist dann die Eulersche Gammafunktion. Für  $n = 1$  ergibt sich die Exponentialverteilung als Spezialfall der Gammaverteilung. Allgemein gilt:

$$\Gamma(\lambda, r) \star \Gamma(\lambda, s) = \Gamma(\lambda, r + s),$$

d.h. die Gammaverteilungen mit festem Parameter  $\lambda$  bilden eine Faltungshalbgruppe.

Durch Anwenden des zentralen Grenzwertsatzes auf die Zufallsvariable  $S_n$  erhalten wir:

**Korollar 9.7 (Normalapproximation der Gammaverteilungen).** Sei  $\lambda > 0$ . Dann gilt für  $\Gamma(\lambda, n)$  verteilte Zufallsvariablen  $S_n$ :

$$n^{-1/2} \cdot (S_n - n\lambda^{-1}) \xrightarrow{\mathcal{D}} N(0, \lambda^{-2}) \quad \text{für } n \rightarrow \infty.$$

**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**  $\lambda$ . 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.

## 9.3 Transformationen, Gaußmodelle und Parameterschätzung

### Der Dichtetransformationssatz

Allgemein gibt es zwei ganz verschiedene Arten, eine Wahrscheinlichkeitsverteilung  $\mu(dx)$  zu transformieren:

(1). Koordinatentransformation:  $y = \phi(x), \quad \mu(dx) \rightarrow \mu \circ \phi^{-1}(dy)$

(2). Maßwechsel durch Dichte:  $\mu(dx) \rightarrow \varrho(x)\mu(dx).$

In bestimmten regulären Fällen lassen sich beide Transformationen in Beziehung setzen: Ein Koordinatenwechsel hat denselben Effekt wie eine absolutstetige Maßtransformation mit einer geeigneten Dichte  $\varrho$ . Wir demonstrieren dies hier im Fall absolutstetiger Verteilungen im  $\mathbb{R}^d$ . Die entsprechende Koordinatentransformationsformel verwenden wir dann, um multivariate Normalverteilungen, und verschiedene für die Statistik zentrale Verteilungen zu untersuchen.

Seien  $S, T \subseteq \mathbb{R}^n$  offen, und sei  $X : \Omega \rightarrow S$  eine Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  mit absolutstetiger Verteilung  $\mu_X$  mit Dichte  $f_X$ .

**Satz 9.8 (Mehrdimensionaler Dichtetransformationssatz).** *Ist  $\phi : S \rightarrow T$  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  $\det D\phi^{-1}(y) = \det\left(\frac{\partial x_i}{\partial y_j}\right)$  die Jacobideterminante der Koordinatentransformation ist.

*Beweis.* Die Behauptung folgt aus dem Transformationssatz der multivariaten Analysis:

$$\begin{aligned} P[\phi(X) \in B] &= P[X \in \phi^{-1}(B)] \\ &= \int_{\phi^{-1}(B)} f_X(x) dx \stackrel{\text{Subst.}}{=} \int_B f_X(\phi^{-1}(y)) \cdot |\det D\phi^{-1}(y)| dy. \end{aligned}$$

□

**Beispiel (Sukzessive Wartezeiten).** Seien  $T$  und  $\tilde{T}$  unabhängige, zum Parameter  $\lambda > 0$  exponentialverteilte Zufallsvariablen (z.B. sukzessive Wartezeiten), und sei  $S = T + \tilde{T}$ . Nach dem Dichtetransformationssatz gilt dann

$$\begin{aligned} f_{T,S}(t, s) &= f_{T,\tilde{T}}(t, s-t) \cdot \left| \det \frac{\partial(t, s-t)}{\partial(t, s)} \right| \\ &\propto e^{-\lambda t} \cdot I_{(0,\infty)}(t) \cdot e^{-\lambda(s-t)} \cdot I_{(0,\infty)}(s-t) \\ &= e^{-\lambda s} \cdot I_{(0,s)}(t). \end{aligned}$$

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  $[0, S]!$

Wir betrachten nun verschiedene weiterreichende Anwendungen des Dichtetransformationssatzes.

### Multivariate Normalverteilungen und multivariater ZGS

Sei  $Z = (Z_1, Z_2, \dots, Z_d)$  mit unabhängigen,  $N(0, 1)$ -verteilten Zufallsvariablen  $Z_i$ . Die Verteilung des Zufallsvektors  $Z$  ist dann absolutstetig bzgl. des Lebesguemaßes im  $\mathbb{R}^d$  mit Dichte

$$f_Z(x) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} = (2\pi)^{-\frac{d}{2}} e^{-\frac{|x|^2}{2}} \quad (d\text{-dimensionale Standardnormalverteilung}).$$

Sei nun  $m \in \mathbb{R}^d$  und  $\sigma \in \mathbb{R}^{d \times d}$  eine  $d \times d$ -Matrix. Wir betrachten den Zufallsvektor

$$Y = \sigma Z + m.$$

Wir zeigen zunächst, dass  $Y$  Erwartungswert  $m$  und Kovarianzmatrix  $C = \sigma\sigma^T$  hat, und berechnen die charakteristische Funktion: **Erwartungswert:**  $E[Y_i] = \sum_{k=1}^d \sigma_{ik} E[Z_k] + m_i = m_i$

**Kovarianz:** 
$$\begin{aligned} \text{Cov}(Y_i, Y_j) &= \text{Cov}(\sum_k \sigma_{ik} Z_k + m_i, \sum_l \sigma_{jl} Z_l + m_j) \\ &= \sum_{k,l} \sigma_{ik} \sigma_{jl} \cdot \text{Cov}(Z_k, Z_l) = \sum_k \sigma_{ik} \sigma_{jk} = C_{ij}. \end{aligned}$$

**Charakteristische Funktion:** Für einen Vektor  $p \in \mathbb{R}^d$  gilt

$$\begin{aligned} \varphi_Y(p) &:= E[e^{ip \cdot Y}] = E\left[e^{i(\sigma^T p) \cdot Z}\right] e^{ip \cdot m} = e^{-\frac{1}{2}|\sigma^T p|^2 + ip \cdot m} \\ &= e^{-\frac{1}{2}p \cdot C p + ip \cdot m}. \end{aligned} \tag{9.3.1}$$

Ist  $\sigma$  regulär, dann können wir die Dichte der Verteilung von  $Y$  sofort mithilfe des Transformationssatzes explizit berechnen:

$$\begin{aligned} f_Y(y) &= f_X(\sigma^{-1}(y - m)) \cdot |\det \sigma^{-1}| \\ &= \frac{1}{\sqrt{(2\pi)^d |\det C|}} \exp\left(-\frac{1}{2}(y - m)C^{-1}(y - m)\right). \end{aligned}$$

Auch im  $\mathbb{R}^d$  ist eine Wahrscheinlichkeitsverteilung durch ihre charakteristische Funktion eindeutig festgelegt, s. z.B. Bauer: Wahrscheinlichkeitstheorie. Allgemein (also auch für nicht reguläre  $\sigma$ ) können wir die Verteilung von  $Y$  auch über die Fourierinversionsformel berechnen.

**Definition.** Sei  $m \in \mathbb{R}^d$  und  $C \in \mathbb{R}^{d \times d}$  eine symmetrische, nicht-negativ definite Matrix. Die Verteilung  $N(m, C)$  im  $\mathbb{R}^d$  mit charakteristischer Funktion  $\phi_Y = \exp(-\frac{1}{2}p \dot{C} p + ipm)$  heißt *d-dimensionale Normalverteilung mit Mittel  $m$  und Kovarianzmatrix  $C$* .

**Bemerkung/Übung.** Mithilfe von charakteristischen Funktionen beweist man die folgenden Transformationsformeln und Charakterisierungen für multivariate Normalverteilungen:

(1). Für  $a \in \mathbb{R}^k$  und  $A \in \mathbb{R}^{k \times d}$  gilt

$$X \sim N(m, C) \Rightarrow AX + a \sim N(Am + a, ACA^T).$$

(2). Folgende Aussagen sind äquivalent:

- $X \sim N(0, C)$  ist multivariat normalverteilt mit Kovarianzmatrix  $C$ .
- $p \cdot X \sim N(0, p \cdot C p) \quad \forall p \in \mathbb{R}^d$ .

Auch im  $\mathbb{R}^d$  gilt ein zentraler Grenzwertsatz :

**Satz 9.9 (Multivariater zentraler Grenzwertsatz).** Seien  $X_1, X_2, \dots : \Omega \rightarrow \mathbb{R}^d$  unabhängige, identisch verteilte, quadratintegrierbare Zufallsvektoren auf  $(\Omega, \mathcal{A}, P)$ , und sei  $S_n = X_1 + \dots + X_n$ . Dann gilt

$$\frac{S_n - E[S_n]}{\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim N(0, C),$$

wobei  $C_{jk} = \text{Cov}(X_{1,j}, X_{1,k})$  die Kovarianzmatrix der Zufallsvektoren  $X_i$  ist.

Der Beweis basiert auf folgender Charakterisierung der schwachen Konvergenz von Zufallsvektoren:

**Lemma 9.10 (Cramér-Wold Device).** Für Zufallsvariablen  $Y, Y_1, Y_2, \dots : \Omega \rightarrow \mathbb{R}^d$  gilt:

$$Y_n \xrightarrow{\mathcal{D}} Y \iff p \cdot Y_n \xrightarrow{\mathcal{D}} p \cdot Y \quad \forall p \in \mathbb{R}^d.$$

*Beweisskizze.* Die Richtung „ $\Rightarrow$ “ ist klar, da  $Y \mapsto p \cdot Y$  stetig ist. Umgekehrt gilt:

$$p \cdot Y_n \xrightarrow{\mathcal{D}} p \cdot Y \implies E[\exp(ip \cdot Y_n)] \rightarrow E[\exp(ip \cdot Y)] \quad \forall p \in \mathbb{R}^d.$$

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

$$F_n(x_1, \dots, x_d) := P[Y_{n,1} \leq x_1, \dots, Y_{n,d} \leq x_d], \quad (x_1, \dots, x_d) \in \mathbb{R}^d.$$

Wir beweisen nun den zentralen Grenzwertsatz:

*Beweis.* Für  $p \in \mathbb{R}^d$  gilt nach dem eindimensionalen zentralen Grenzwertsatz:

$$\begin{aligned} p \cdot \left( \frac{S_n - E[S_n]}{\sqrt{n}} \right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (p \cdot X_i - E[p \cdot X_i]) \\ &\xrightarrow{\mathcal{D}} N(0, \text{Var}[p \cdot X_1]) = N(0, p \cdot Cp), \end{aligned}$$

da

$$\text{Var}[p \cdot X_1] = \text{Cov} \left[ \sum_k p_k X_{1,k}, \sum_l p_l X_{1,l} \right] = \sum_{k,l} p_k p_l C_{kl} = p \cdot Cp.$$

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

$$(S_n - E[S_n]) / \sqrt{n} \xrightarrow{\mathcal{D}} Y.$$

□

**Beispiel (Vom Random walk zur Brownschen Bewegung).** Sei  $S_n = X_1 + \dots + X_n$ , wobei die  $X_i$  unabhängige Zufallsvariablen mit

$$E[X_i] = 0 \quad \text{und} \quad \text{Var}[X_i] = 1$$

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

$$\tilde{S}_t^{(n)} := \frac{1}{\sqrt{n}} S_{nt}, \quad t \in \mathbb{R}_+.$$

#### GRAPHIK SKALIRTER RANDOM WALK

Aus dem Zentralen Grenzwertsatz folgt:

$$\tilde{S}_t^{(n)} = \sqrt{t} \frac{1}{\sqrt{nt}} S_{nt} \xrightarrow{\mathcal{D}} \sim N(0, t) \quad \text{für jedes feste } t \in \mathbb{R}_+,$$

d.h. die eindimensionalen Randverteilungen der Prozesse  $\tilde{S}^{(n)} = (\tilde{S}_t^{(n)})_{t \geq 0}$  konvergieren. Allgemeiner zeigt man mithilfe des multivariaten zentralen Grenzwertsatzes, dass auch endlich dimensionale Randverteilungen schwach konvergieren:

$$\left( \tilde{S}_{t_1}^{(n)}, \tilde{S}_{t_2}^{(n)}, \dots, \tilde{S}_{t_k}^{(n)} \right) \xrightarrow{\mathcal{D}} (B_{t_1}, \dots, B_{t_k}), \quad \text{für alle } 0 \leq t_1 < t_2 < \dots < t_k, k \in \mathbb{N},$$

wobei  $(B_{t_1}, \dots, B_{t_k})$  multivariat normalverteilt ist mit

$$E[B_{t_j}] = 0 \quad \text{und} \quad \text{Cov}[B_{t_j}, B_{t_k}] = \min(t_j, t_k).$$

Eine noch allgemeinere Aussage erhält man mithilfe eines **funktionalen zentralen Grenzwertsatzes** (Invarianzprinzip von Donsker, ZGS auf dem Banachraum  $C([0, 1], \mathbb{R})$ ): Der gesamte stochastische Prozess  $(\tilde{S}_t^{(n)})_{0 \leq t \leq 1}$  konvergiert in Verteilung gegen eine **Brownsche Bewegung**  $(B_t)_{0 \leq t \leq 1}$ . Mehr dazu in den weiterführenden Vorlesungen »Stochastische Prozesse« und »Grundzüge der stochastischen Analysis«.

Wir betrachten noch eine weitere Anwendung des Dichtetransformationssatzes auf Normalverteilungen.

**Beispiel ( $\chi^2$ -Verteilungen).** Wir berechnen nun die Verteilung vom Quadrat des Abstandes vom Ursprung eines standardnormalverteilten Zufallsvektors im  $\mathbb{R}^d$ :

$$Z = (Z_1, \dots, Z_d) \sim N(0, I_d), \quad \|Z\|^2 = \sum_{i=1}^d Z_i^2.$$



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 mit  $Y = \phi(x) := x^2$ :

$$f_{Z_i^2}(y) = \sqrt{\frac{2}{\pi}} e^{-\frac{y}{2}} \cdot I_{(0,\infty)}(y) \cdot \frac{1}{2\sqrt{y}},$$

d.h.  $Z_i^2$  ist  $\Gamma(\frac{1}{2}, \frac{1}{2})$ -verteilt. Da die Zufallsvariablen  $Z_i^2$ ,  $1 \leq i \leq d$ , unabhängig sind, folgt:

$$\|Z\|^2 = \sum_{i=1}^d Z_i^2 \sim \Gamma\left(\frac{1}{2}, \frac{d}{2}\right).$$

**Definition.** Die Gamma-Verteilung mit Parametern  $\frac{1}{2}$  und  $\frac{d}{2}$  heißt auch **Chiquadrat-Verteilung**  $\chi^2(d)$  mit  $d$  **Freiheitsgraden**.

## Parameterschätzung im Gaußmodell

Angenommen, wir beobachten reellwertige Meßwerte (Stichproben, Daten), die von einer unbekannt Wahrscheinlichkeitsverteilung  $\mu$  auf  $\mathbb{R}$  stammen. Ziel der Statistik ist es, Rückschlüsse auf die zugrundeliegende Verteilung aus den Daten zu erhalten. Im einfachsten Modell (Gaußmodell) 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  $X_1(\omega) = x_1, \dots, X_n(\omega) = x_n$  zu schätzen, und zu quantifizieren.

### Problemstellung: Schätzung des Erwartungswerts

- Schätze  $m$  auf der Basis von  $n$  unabhängigen Stichproben  $X_1(\omega), \dots, X_n(\omega)$  von  $\mu$ .
- Herleitung von Konfidenzintervallen.

Im mathematischen Modell interpretieren wir die Beobachtungswerte als Realisierungen von unabhängigen Zufallsvariablen  $X_1, \dots, X_n$ . Da wir die tatsächliche Verteilung nicht kennen, untersuchen wir alle in Betracht gezogenen Verteilungen simultan:

$$X_1, \dots, X_n \sim N(m, v) \quad \text{unabhängig unter } P_{m,v}. \quad (9.3.2)$$

Ein naheliegender Schätzer für  $m$  ist der *empirische Mittelwert*

$$\bar{X}_n(\omega) := \frac{X_1(\omega) + \dots + X_n(\omega)}{n}.$$

Wir haben oben bereits gezeigt, dass dieser Schätzer *erwartungstreu (unbiased)* und *konsistent* ist, d.h. für alle  $m, v$  gilt:

$$E_{m,v}[\bar{X}_n] = m$$

und

$$\bar{X}_n \rightarrow m \quad P_{m,v}\text{-stochastisch für } n \rightarrow \infty.$$

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  $\bar{X}_n$ :

$$\begin{aligned} X_i \sim N(m, v) \text{ unabh.} &\Rightarrow X_1 + \dots + X_n \sim N(nm, nv) \\ &\Rightarrow \bar{X}_n \sim N\left(m, \frac{v}{n}\right) \\ &\Rightarrow \frac{\bar{X}_n - m}{\sqrt{v/n}} \sim N(0, 1) \end{aligned}$$

Bezeichnet  $\Phi$  die Verteilungsfunktion der Standardnormalverteilung, dann erhalten wir

$$P_{m,v} \left[ |\bar{X}_n - m| < q \sqrt{\frac{v}{n}} \right] = N(0, 1)(-q, q) = 2 \left( \Phi(q) - \frac{1}{2} \right) \quad \text{für alle } m \in \mathbb{R}.$$

**Satz 9.11.** *Im Gaußmodell (9.3.2) mit bekannter Varianz  $v$  ist das zufällige Intervall*

$$\left( \bar{X}_n - \Phi^{-1}(\alpha) \sqrt{\frac{v}{n}}, \bar{X}_n + \Phi^{-1}(\alpha) \sqrt{\frac{v}{n}} \right)$$

ein  $(2\alpha - 1) \cdot 100\%$  **Konfidenzintervall** für  $m$ , d.h.

$$P_{m,v}[m \in \text{Intervall}] \geq 2\alpha - 1 \quad \text{für alle } m \in \mathbb{R}.$$

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 unbekanntem 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

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{und} \quad V_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Um ein Konfidenzintervall für  $m$  zu erhalten, bestimmen wir mithilfe des Transformationssatzes die gemeinsame Verteilung von  $\bar{X}_n$  und  $V_n$ :

**Lemma 9.12.**  $\bar{X}_n$  und  $V_n$  sind unabhängig unter  $P_{m,v}$  mit Verteilung

$$\bar{X}_n \sim N\left(m, \frac{v}{n}\right) \quad , \quad \frac{n-1}{v} V_n \sim \chi^2(n-1).$$

*Beweis.* Wir führen eine lineare Koordinatentransformation  $Y = OX$  durch, wobei  $O$  eine orthogonale  $n \times n$ -Matrix vom Typ

$$O = \begin{pmatrix} \frac{1}{\sqrt{n}} & \dots & \frac{1}{\sqrt{n}} \\ \text{beliebig} \end{pmatrix}$$

ist. Eine solche Matrix erhalten wir durch Ergänzen des normierten Vektors  $(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$  zu einer Orthonormalbasis des  $\mathbb{R}^n$ . In den neuen Koordinaten gilt:

$$\begin{aligned} \bar{X}_n &= \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{\sqrt{n}} Y_1, \quad \text{und} \\ (n-1)V_n &= \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n X_i^2 - n\bar{X}_n^2 = \|X\|_{\mathbb{R}^n}^2 - n\bar{X}_n^2 \\ &\stackrel{O \text{ orthogonal}}{=} \|Y\|_{\mathbb{R}^n}^2 - Y_1^2 = \sum_{i=2}^n Y_i^2. \end{aligned}$$

Da die Zufallsvariablen  $X_i$  ( $1 \leq i \leq n$ ) unabhängig und  $N(m, v)$ -verteilt sind, ist der Zufallsvektor  $X = (X_1, \dots, X_n)$  multivariat normalverteilt mit Mittel  $(m, \dots, m)$  und Kovarianzmatrix  $v \cdot I_n$ . Nach dem Transformationssatz folgt

$$Y \sim N \left( O \begin{pmatrix} m \\ \vdots \\ m \end{pmatrix}, v \cdot O I_n O^T \right) = N \left( \begin{pmatrix} m\sqrt{n} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, v \cdot I_n \right).$$

Also sind  $Y_1, \dots, Y_n$  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

$$\bar{X}_n = \frac{Y_1}{\sqrt{n}} \quad \text{und} \quad \frac{n-1}{v} V_n = \sum_{i=2}^n \left( \frac{Y_i}{\sqrt{v}} \right)^2$$

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  $\bar{X}_n \pm q \cdot \sqrt{\frac{v}{n}}$  erhalten, wobei  $q$  ein geeignetes Quantil der Standardnormalverteilung ist. Daher liegt es nahe, zu versuchen, bei unbekannter Varianz Konfidenzintervalle vom Typ  $\bar{X}_n \pm q \cdot \sqrt{\frac{V_n}{n}}$  herzuleiten. Es gilt:

$$P_{m,v} \left[ |\bar{X}_n - m| \geq q \sqrt{\frac{V_n}{n}} \right] = P_{m,v} [|T_{n-1}| \geq q] \quad \text{mit}$$

$$T_{n-1} := \frac{\sqrt{n} \cdot (\bar{X}_n - m)}{\sqrt{V_n}}.$$

Die Zufallsvariable  $T_{n-1}$  heißt **Studentsche  $t$ -Statistik mit  $n - 1$  Freiheitsgraden**.<sup>1</sup> 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 9.13** (Student<sup>2</sup>). *Die Verteilung von  $T_n$  ist absolutstetig mit Dichte*

$$f_{T_n}(t) = B\left(\frac{1}{2}, \frac{n}{2}\right)^{-1} \cdot n^{-1/2} \cdot \left(1 + \frac{t^2}{2}\right)^{-n/2} \quad (t \in \mathbb{R}).$$

»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

$$\bar{X}_n \pm q \cdot \sqrt{\frac{V_n}{n}}$$

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 Transformationsatzes 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  $f_{T_{n-1}}$ .

<sup>1</sup>In der Statistik bezeichnet man eine meßbare Funktion der Beobachtungsdaten als Statistik - ein (Punkt-) Schätzer ist eine Statistik, die zum Schätzen eines unbekanntem Parameters verwendet wird, ein Konfidenzintervall nennt man auch Intervallschätzer.

<sup>2</sup>Synonym von W. S. Gosset, der als Angestellter der Guinness-Brauerei nicht publizieren durfte.

Der Satz folgt dann nach Lemma 9.12 mit

$$Z := \frac{\bar{X}_n - m}{\sqrt{v/n}} \quad \text{und} \quad Y := \frac{n-1}{v} V_n.$$

□

**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 5.1 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)

### Exkurs zu 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  $X_1, \dots, X_n$  von einer unbekanntem Verteilung vorliegen und wir gehen davon aus, daß die zugrundeliegende Verteilung aus einer Familie  $\mu_\theta$  ( $\theta \in \Theta$ ) 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  $X_1, \dots, X_n$  ist dann das Produktmaß  $\mu_\theta^n = \bigotimes_{i=1}^n \mu_\theta$ . Sei nun  $\Theta_0$  eine Teilmenge des Parameterbereichs. Wir wollen entscheiden zwischen der

$$\text{Nullhypothese } H_0: \quad \gg \theta \in \Theta_0 \ll$$

und der

$$\text{Alternative } H_1: \quad \gg \theta \notin \Theta_0 \ll$$

Ein **Hypothesentest** für ein solches Problem ist bestimmt durch eine meßbare Teilmenge  $C \subseteq \mathbb{R}^n$  (den **Verwerfungsbereich**) mit zugehöriger Entscheidungsregel:

$$\text{Akzeptiere } H_0 \iff (X_1, \dots, X_n) \notin C.$$

**Beispiel (t-Test).** Seien  $X_1, X_2, \dots, X_n$  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:

$$\text{Nullhypothese } H_0: \quad \gg m = m_0 \ll, \quad \Theta_0 = \{m_0\} \times \mathbb{R}^+.$$

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 *Studentischen t-Statistik* unterhalb einer angemessen zu wählenden Konstanten  $c$  liegt, bzw. verworfen, falls

$$|T_{n-1}| = \left| \frac{\sqrt{n} \cdot (\bar{X}_n - m_0)}{\sqrt{V_n}} \right| > c$$

gilt.

Seien nun allgemein  $X_1, X_2, \dots$  unter  $P_\theta$  unabhängige Zufallsvariablen mit Verteilung  $\mu_\theta$ . Bei einem Hypothesentest können zwei Arten von Fehlern auftreten:

**Fehler 1. Art:**  $H_0$  wird verworfen, obwohl wahr. Die Wahrscheinlichkeit dafür beträgt:

$$P_\theta[(X_1, \dots, X_n) \in C] = \mu_\theta^n[C], \quad \theta \in \Theta_0.$$

**Fehler 2. Art:**  $H_0$  wird akzeptiert, obwohl falsch. Die Wahrscheinlichkeit beträgt:

$$P_\theta[(X_1, \dots, X_n) \notin C] = \mu_\theta^n[C^C], \quad \theta \in \Theta \setminus \Theta_0.$$

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  $\alpha$  liegen. Die Wahrscheinlichkeit

$$\mu_\theta^n[C], \quad \theta \in \Theta \setminus \Theta_0,$$

daß kein Fehler 2. Art auftritt, sollte unter dieser Voraussetzung möglichst groß sein.

**Definition.** Die Funktion

$$G(\theta) = P_\theta[(X_1, \dots, X_n) \in C] = \mu_\theta^n[C]$$

heißt **Gütefunktion** des Tests. Der Test hat **Niveau**  $\alpha$ , falls

$$G(\theta) \leq \alpha \quad \text{für alle } \theta \in \Theta_0$$

gilt. Die Funktion  $G(\theta)$  mit  $\theta \in \Theta_1$  heißt **Macht** des Tests.

Aus Satz 9.13 und der Symmetrie der Studentschen  $t$ -Verteilung folgt unmittelbar:

**Korollar 9.14.** Der Studentsche  $t$ -Test hat Niveau  $\alpha$  falls  $c$  ein  $(1 - \frac{\alpha}{2})$ -Quantil der Studentschen  $t$ -Verteilung mit  $n - 1$  Freiheitsgraden ist.

Allgemeiner gilt:

**Satz 9.15 (Korrespondenz Konfidenzintervalle  $\leftrightarrow$  Hypothesentests).** Für einen reellwertigen Parameter  $\gamma = c(\theta)$ , ein Irrtumsniveau  $\alpha \in (0, 1)$ , und messbare Abbildungen (Statistiken)  $\hat{\gamma}, \varepsilon : \mathbb{R}^n \rightarrow \mathbb{R}$  sind äquivalent:

(i) Das Intervall

$$[\hat{\gamma}(X_1, \dots, X_n) - \varepsilon(X_1, \dots, X_n), \hat{\gamma}(X_1, \dots, X_n) + \varepsilon(X_1, \dots, X_n)]$$

ist ein  $(1 - \alpha) \cdot 100\%$  Konfidenzintervall für  $\gamma$ .

(ii) Für jedes  $\gamma_0 \in \mathbb{R}$  ist

$$C = \{(x_1, \dots, x_n) : |\hat{\gamma}(x_1, \dots, x_n) - \gamma_0| > \varepsilon(x_1, \dots, x_n)\}$$

der Verwerfungsbereich eines Test der Nullhypothese  $\gamma = \gamma_0$  zum Niveau  $\alpha$ .

**Beweis.** Das Intervall ist genau dann ein Konfidenzintervall für  $\gamma$  zum Irrtumsniveau  $\alpha$ , wenn

$$P_\theta [|\hat{\gamma}(X_1, \dots, X_n) - c(\theta)| > \varepsilon(X_1, \dots, X_n)] \leq \alpha \quad \forall \theta \in \Theta$$

gilt, also wenn der entsprechende Test der Nullhypothesen  $c(\theta) = \gamma_0$  für jedes  $\gamma_0$  Niveau  $\alpha$  hat. □

# Kapitel 10

## Bedingte Erwartungen

### 10.1 Bedingen auf diskrete Zufallsvariablen

Zur Analyse von stochastischen Modellen mit Abhängigkeiten verwendet man in der Regel bedingte Wahrscheinlichkeiten und Erwartungswerte gegeben die Werte von Zufallsvariablen. Beispielsweise beschreibt man einen stochastischen Prozess  $X_n, n \in \mathbb{N}$ , durch die bedingten Verteilungen des nächsten Zustands  $X_{n+1}$  gegeben den Verlauf  $X_{0:n} = (X_0, X_1, \dots, X_n)$  bis zur Zeit  $n$ . Wir betrachten zunächst das Bedingen auf den Ausgang einer diskreten Zufallsvariable  $Y : \Omega \rightarrow S, S$  abzählbar. In diesem Fall können wir die *bedingte Wahrscheinlichkeitsverteilung*

$$P[A | Y = z] = \frac{P[A \cap \{Y = z\}]}{P[Y = z]}, \quad A \in \mathcal{A},$$

und die *bedingten Erwartungswerte*

$$E[X | Y = z] = \frac{E[X; Y = z]}{P[Y = z]}, \quad X \in \mathcal{L}^1(\Omega, \mathcal{A}, P),$$

für alle  $z \in S$  mit  $P[Y = z] > 0$  auf elementare Weise wie in Abschnitt 2.1 definieren. Für  $z \in S$  mit  $P[Y = z] = 0$  sind die bedingten Wahrscheinlichkeiten nicht definiert.

### Bedingte Erwartungen als Zufallsvariablen

Es wird sich als praktisch erweisen, die bedingten Wahrscheinlichkeiten und Erwartungswerte nicht als Funktion des Ausgangs  $z$ , sondern als Funktion der Zufallsvariable  $Y$  zu interpretieren. Die bedingten Wahrscheinlichkeiten und Erwartungswerte sind dann selbst Zufallsvariablen:



**Definition.** Sei  $X : \Omega \rightarrow \mathbb{R}$  eine Zufallsvariable mit  $E[X^-] < \infty$ , und  $Y : \Omega \rightarrow S$  eine diskrete Zufallsvariable. Die durch

$$E[X | Y] := g(Y) = \sum_{z \in S} g(z) \cdot I_{\{Y=z\}}$$

mit

$$g(z) := \begin{cases} E[X | Y = z] & \text{falls } P[Y = z] > 0 \\ \text{beliebig} & \text{falls } P[Y = z] = 0 \end{cases}$$

$P$ -fast sicher eindeutig definierte Zufallsvariable  $E[X | Y]$  heißt (**Version der**) **bedingte(n) Erwartung von  $X$  gegeben  $Y$** . Für ein Ereignis  $A \in \mathcal{A}$  heißt die Zufallsvariable

$$P[A | Y] := E[I_A | Y]$$

(**Version der**) **bedingte(n) Wahrscheinlichkeit von  $A$  gegeben  $Y$** .

Die bedingte Erwartung  $E[X | Y]$  und die bedingte Wahrscheinlichkeit  $P[A | Y]$  sind also Zufallsvariablen mit den Werten  $E[X | Y = z]$  bzw.  $P[A | Y = z]$  auf den Mengen  $\{Y = z\}$ ,  $z \in S$  mit  $P[Y = z] > 0$ . Auf jeder der Nullmengen  $\{Y = z\}$ ,  $z \in S$  mit  $P[Y = z] = 0$ , wird der bedingten Erwartung ein willkürlicher konstanter Wert zugewiesen, d.h. die Definition ist nur  $P$ -fast überall eindeutig. Wir fassen zunächst einige elementare Eigenschaften der so definierten bedingten Erwartung zusammen:

**Lemma 10.1 (Eigenschaften der bedingten Erwartung).**

- (1). Die Abbildung  $X \mapsto E[X | Y]$  ist  $P$ -fast sicher linear und monoton.
- (2). Sind  $X$  und  $Y$  unabhängig, dann gilt  $E[X | Y] = E[X]$   $P$ -fast sicher;
- (3). Herausziehen, was bekannt ist:

Für alle  $f : S \rightarrow \mathbb{R}$  mit  $f(Y) \cdot X \geq 0$  bzw.  $f(Y) \cdot X \in \mathcal{L}^1$  gilt

$$E[f(Y) \cdot X | Y] = f(Y) \cdot E[X | Y] \quad P\text{-fast sicher.}$$

Insbesondere gilt

$$E[f(Y) | Y] = f(Y) \quad P\text{-fast sicher.}$$

*Beweis.* (2). Sind  $X$  und  $Y$  unabhängig, dann gilt

$$E[X | Y = z] = \frac{E[X \cdot I_{\{Y=z\}}]}{P[Y = z]} = E[X]$$

für alle  $z \in S$  mit  $P[Y = z] > 0$ , also  $E[X | Y] = E[X]$   $P$ -fast sicher. Die ebenso elementaren Beweise von (1) und (3) werden dem Leser als Übung überlassen.  $\square$

Anschaulich können wir die zweite Aussage folgendermaßen interpretieren: Sind  $X$  und  $Y$  unabhängig, dann liefert die Kenntnis des Wertes  $Y(\omega)$  keine zusätzlichen Informationen über  $X(\omega)$ . Daher ist die beste  $L^2$ -Prognose für  $X(\omega)$  wie im unbedingten Fall durch den Erwartungswert  $E[X]$  gegeben.

### Formel von der totalen Wahrscheinlichkeit

Die aus Satz 2.1 bekannte Formel von der totalen Wahrscheinlichkeit können wir mithilfe der obigen Definition in kompakter Weise schreiben.

**Satz 10.2 (Formel von der totalen Wahrscheinlichkeit).** Sei  $Y : \Omega \rightarrow S$  eine diskrete Zufallsvariable mit Verteilung  $\mu(z) = P[Y = z]$ . Für alle messbaren  $X : \Omega \rightarrow \mathbb{R}_+$  gilt:

$$E[X] = \sum_{z: \mu(z) \neq 0} E[X | Y = z] \mu(z) = E[E[X | Y]]$$

Insbesondere gilt

$$P[A] = E[P[A | Y]] \quad \text{für alle } A \in \mathcal{A}.$$

*Beweis.* Wegen  $\Omega = \dot{\bigcup}_{z \in S} \{Y = z\}$  gilt nach dem Transformationssatz

$$\begin{aligned} E[X] &= \sum_{z \in S} E[X; Y = z] = \sum_{z: \mu(z) \neq 0} E[X; Y = z] \\ &= \sum_{z: \mu(z) \neq 0} E[X | Y = z] \cdot \mu(z) = \sum_{z: \mu(z) \neq 0} g(z) \cdot \mu(z) \\ &= E[g(Y)], \end{aligned}$$

wobei  $g : S \rightarrow \mathbb{R}$  eine beliebige Funktion mit  $g(z) = E[X | Y = z]$  für alle  $z \in S$  mit  $\mu(z) \neq 0$  ist. Die Aussage folgt wegen  $g(Y) = E[X | Y]$   $P$ -fast sicher.  $\square$

**Bemerkung.** Für  $X \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$  folgt aus der Monotonie der bedingten Erwartung

$$|E[X | Y]| \leq E[|X| | Y]$$

und damit die Ungleichung

$$E[|E[X | Y]|] \leq E[E[|X| | Y]] = E[|X|].$$

Die Abbildung  $X \mapsto E[X | Y]$  ist also eine Kontraktion auf  $L^1(\Omega, \mathcal{A}, P)$ . Die Aussage von Satz 10.2 gilt entsprechend auch für  $X \in \mathcal{L}^1$ .

## Bedingte Varianz

Sei nun  $X : \Omega \rightarrow \overline{\mathbb{R}}$  eine bzgl.  $P$  integrierbare Zufallsvariable

**Definition.**

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

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

Ist  $X$  quadratintegrierbar, dann gelten die folgenden Aussagen:

**Lemma 10.3.** Für  $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$  gilt:

(1).  **$L^2$ -Kontraktivität:**  $E[|E[X | Y]|^2] \leq E[X^2]$ .

(2).  $\text{Var}[X | Y] = E[X^2 | Y] - E[X | Y]^2$   *$P$ -fast sicher.*  
 Insbesondere folgt für  $z \in S$  mit  $\mu(z) \neq 0$ :

$$\text{Var}[X | Y] = \text{Var}[X | Y = z] \quad \text{auf } \{Y = z\}. \quad (10.1.1)$$

*Beweis.* (1). folgt aus Satz 10.2, da für alle  $z \in S$  mit  $P[Y = z] \neq 0$  auf  $\{Y = z\}$  gilt:

$$|E[X | Y]|^2 = |E[X | Y = z]|^2 \leq E[X^2 | Y = z] = E[X^2 | Y].$$

(2). Nach Lemma 10.1, (1) und (3), ergibt sich dann ähnlich wie für die unbedingte Varianz:

$$\begin{aligned} \text{Var}(X | Y) &= E[X^2 | Y] - 2 \cdot E[X \cdot E[X | Y] | Y] + E[E[X | Y]^2 | Y] \\ &= E[X^2 | Y] - E[X | Y]^2 \quad P\text{-fast sicher} \end{aligned}$$

□

Die folgende Zerlegungsformel kann häufig verwendet werden, um Varianzen zu berechnen oder abzuschätzen:

**Satz 10.4 (Formel von der bedingten Varianz).** Für eine Zufallsvariable  $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$  gilt:

$$\begin{aligned} \text{Var}[X] &= E[\text{Var}[X | Y]] + \text{Var}[E[X | Y]] \\ &= \sum_{z:\mu(z) \neq 0} \text{Var}[X | Y = z] \cdot \mu(z) + \sum_{z:\mu(z) \neq 0} (E[X | Y = z] - E[X])^2 \cdot \mu(z). \end{aligned}$$

*Beweis.* Es gilt

$$\begin{aligned}\text{Var}[X] &= E[X^2] - E[X]^2 = E[E[X^2 | Y]] - E[E[X | Y]]^2 \\ &= E[E[X^2 | Y]] - E[E[X | Y]^2] + E[E[X | Y]^2] - E[E[X | Y]]^2 \\ &= E[\text{Var}[X | Y]] + \text{Var}[E[X | Y]].\end{aligned}$$

Der zweite Teil der Behauptung folgt nun aus (10.1.1) und der entsprechenden Eigenschaft für die bedingte Erwartung.  $\square$

### Anwendung auf zufällige Summen

Als erste Anwendung betrachten wir eine Summe

$$S_N(\omega) := \sum_{i=1}^{N(\omega)} X_i(\omega)$$

von unabhängigen, identisch verteilten Zufallsvariablen  $X_i \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$  mit zufälliger Anzahl  $N$  von Summanden. Hierbei sei  $N : \Omega \rightarrow \{0, 1, 2, \dots\}$  eine von den  $X_i$  unabhängige Zufallsvariable. Seien  $m = E[X_i]$  und  $\sigma^2 = \text{Var}[X_1]$ . Wir berechnen nun die verschiedenen Kenngrößen der Verteilung von  $S_N$ .

*Berechnung des Erwartungswertes:* Da  $S_k$  und  $N$  unabhängig sind, erhalten wir

$$E[S_N | N = k] = E[S_k | N = k] = E[S_k] = k \cdot m \quad \text{für alle } k \in \mathbb{N},$$

also  $E[S_N | N] = N \cdot m$  und damit nach Satz 10.2:

$$E[S_N] = E[E[S_N | N]] = E[N] \cdot m.$$

*Berechnung der Varianz:* Erneut folgt wegen der Unabhängigkeit von  $S_k$  und  $N$ :

$$\text{Var}[S_N | N = k] = \text{Var}[S_k | N = k] = \text{Var}[S_k] = k \cdot \sigma^2,$$

also  $\text{Var}[S_N | N] = N \cdot \sigma^2$ , und damit nach Satz 10.4:

$$\text{Var}[S_N] = E[\text{Var}[S_N | N]] + \text{Var}[E[S_N | N]] = E[N] \cdot \sigma^2 + \text{Var}[N] \cdot m^2.$$

*Berechnung der momentenerzeugenden Funktion:* Für  $t \in \mathbb{R}$  gilt

$$\begin{aligned}M_{S_N}(t) &= E[e^{tS_N}] = E[E[e^{tS_N} | N]] = E\left[\prod_{i=1}^N E[e^{tX_i}]\right] \\ &= E[E[e^{tX_1}]^N] = E[M_{X_1}(t)^N] = M_N(\log M_{X_1}(t)).\end{aligned}$$

Mithilfe von  $M_{S_N}$  kann man die Momente der zufälligen Summe  $S_N$  berechnen:

$$E[S_N^m] = M_{S_N}^{(m)}(0) \quad \text{für alle } m \in \mathbb{N}.$$

Im Prinzip erhält man die Verteilung von  $S_N$  durch Laplaceinversion, was aber nicht immer praktikabel ist. Nehmen die Zufallsvariablen  $X_i$  nur nichtnegative ganzzahlige Werte an, kann man statt der momentenerzeugenden Funktion die erzeugende Funktion verwenden, und daraus die Verteilung berechnen. Wir gehen darauf im folgenden Abschnitt ein.

### Charakterisierende Eigenschaften der bedingten Erwartung

Zum Abschluss dieses Abschnitts beweisen wir eine alternative Charakterisierung der bedingten Erwartung gegeben eine diskrete Zufallsvariable  $Y : \Omega \rightarrow S$ ,  $S$  abzählbar. Diese Charakterisierung werden wir in Abschnitt ?? verwenden, um bedingte Erwartungen für allgemeine Bedingungen zu definieren. Sei  $X : \Omega \rightarrow \mathbb{R}_+$  eine nicht negative (bzw. integrierbare) Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ .

**Satz 10.5.** *Eine reellwertige Zufallsvariable  $\bar{X} \geq 0$  (bzw.  $\bar{X} \in \mathcal{L}^1$ ) auf  $(\Omega, \mathcal{A}, P)$  ist genau dann eine Version der bedingten Erwartung  $E[X | Y]$ , wenn gilt:*

(I)  $\bar{X} = g(Y)$  für eine Funktion  $g : S \rightarrow \mathbb{R}$

(II)  $E[\bar{X} \cdot f(Y)] = E[X \cdot f(Y)]$  für alle nichtnegativen bzw. beschränkten Funktionen  $f : S \rightarrow \mathbb{R}$ .

*Beweis.* Ist  $\bar{X}$  eine Version von  $E[X | Y]$ , dann gilt (I). Außerdem folgt nach Lemma 10.1 (3) und der Formel von der totalen Wahrscheinlichkeit:

$$E[\bar{X} \cdot f(Y)] = E[E[X | Y] \cdot f(Y)] = E[E[X \cdot f(Y) | Y]] = E[X \cdot f(Y)]$$

für jede nichtnegative, bzw. beschränkte, Funktion  $f : S \rightarrow \mathbb{R}$ .

Umgekehrt folgt aus (I), dass  $\bar{X} = g(z)$  auf  $\{Y = z\}$  gilt. Ist außerdem (II) erfüllt, dann folgt weiter

$$\begin{aligned} g(z) &= E[\bar{X} | Y = z] = \frac{E[\bar{X} \cdot I_{\{z\}}(Y)]}{P[Y = z]} \\ &= \frac{E[X \cdot I_{\{z\}}(Y)]}{P[Y = z]} = E[X | Y = z] \end{aligned}$$

für alle  $z \in S$  mit  $P[Y = z] > 0$ , d.h.  $\bar{X} = g(Y)$  ist eine Version der bedingten Erwartung  $E[X | Y]$ .  $\square$

Eine Charakterisierung der bedingten Erwartung wie in Satz 10.5 werden wir in Abschnitt ?? verwenden, um bedingte Erwartungen für allgemeine Bedingungen zu definieren. In einigen Fällen können die charakterisierenden Eigenschaften auch direkt überprüft werden, um bedingte Erwartungen zu identifizieren:

**Beispiel (Summen austauschbarer Zufallsvariablen).** Seien  $X_1, X_2, \dots, X_n \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$  integrierbare Zufallsvariablen, deren gemeinsame Verteilung invariant unter Koordinatenpermutationen ist, d.h.  $(X_{\pi(1)}, X_{\pi(2)}, \dots, X_{\pi(n)}) \sim (X_1, X_2, \dots, X_n)$  für alle  $\pi \in \mathcal{S}_n$ . Zufallsvariablen mit dieser Eigenschaft heißen **austauschbar** – beispielsweise sind unabhängige identisch verteilte Zufallsvariablen austauschbar. Wir zeigen:

$$E[X_i | S_n] = \frac{1}{n} S_n \quad P\text{-fast sicher für alle } i = 1, \dots, n,$$

wobei  $S_n = X_1 + \dots + X_n$ . Zum Beweis überprüfen wir, dass  $\bar{X}_i := \frac{1}{n} S_n$  die Bedingungen (I) und (II) aus Satz 10.5 für  $Y = S_n$  erfüllt. (I) ist offensichtlich. Zudem gilt wegen der Austauschbarkeit für jede beschränkte messbare Funktion  $f: \mathbb{R} \rightarrow \mathbb{R}$ :

$$E[X_i \cdot f(S_n)] = E[X_j \cdot f(S_n)] \quad \text{für alle } i, j = 1, \dots, n,$$

also

$$E\left[\frac{1}{n} S_n \cdot f(S_n)\right] = \frac{1}{n} \sum_{j=1}^n E[X_j \cdot f(S_n)] = E[X_i \cdot f(S_n)]$$

für alle  $i = 1, \dots, n$ , d.h. (II) ist auch erfüllt.

## 10.2 Erzeugende Funktionen, Verzweigungsprozesse und Erneuerungen

Wir wollen die Methoden aus dem letzten Abschnitt nun verwenden, um Verzweigungs- und Erneuerungsprozesse zu untersuchen. Ein wichtiges Hilfsmittel sind in beiden Fällen erzeugende Funktionen:

### Erzeugende Funktionen von ganzzahligen Zufallsvariablen

Sei  $X: \Omega \rightarrow \{0, 1, 2, \dots\}$  eine auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  definierte Zufallsvariable mit nichtnegativen *ganzzahligen* Werten.

**Definition.** Die durch

$$G(s) = E[s^X] = \sum_{k=0}^{\infty} P[X = k] s^k, \quad s \in [-1, 1],$$

definierte Funktion heißt **erzeugende Funktion** der Zufallsvariable  $X$  bzw. der Folge  $\mu(k) = P[X = k]$  der Gewichte von  $X$ .

Durch Vergleich mit der geometrischen Reihe sieht man, dass der Konvergenzradius der Potenzreihe stets größer oder gleich 1 ist. Also ist die erzeugende Funktion analytisch auf  $(-1, 1)$ , und es gilt

$$P[X = k] = \frac{G^{(k)}(0)}{k!} \quad \text{für alle } k = 0, 1, 2, \dots$$

Kennen wir also die erzeugende Funktion explizit, dann können wir die Gewichte der Verteilung berechnen.

Durch zweimaliges Ableiten zeigt man zudem, dass  $G$  monoton und konvex auf  $[0, 1]$  ist. Für  $s \in (0, 1]$  gilt nach Definition  $G(s) = M(\log s)$ . Daher lassen sich aus der erzeugenden Funktion die Momente von  $X$  berechnen – beispielsweise gilt  $E[X] = G'(1-)$  (linksseitige Ableitung von  $G(s)$  bei  $s = 1$ ), falls der Erwartungswert endlich ist.

Für die erzeugende Funktion einer Summe  $X + Y$  von unabhängigen, nichtnegativen, ganzzahligen Zufallsvariablen  $X$  und  $Y$  gilt offensichtlich

$$G_{X+Y}(s) = G_X(s) \cdot G_Y(s) \quad \text{für alle } s \in [-1, 1].$$

Somit ist die erzeugende Funktion der Faltung

$$(\mu * \nu)(k) = \sum_{i=0}^k \mu(i) \nu(k-i) \quad (k = 0, 1, 2, \dots)$$

zweier Wahrscheinlichkeitsverteilungen  $\mu$  und  $\nu$  auf  $\mathbb{N} \cup \{0\}$  das Produkt der einzelnen erzeugenden Funktionen.

Erzeugende Funktionen können in verschiedenen Situationen für explizite Berechnungen verwendet werden. Wir demonstrieren die hier in einigen grundlegenden Beispielen. viele weitere entsprechende Anwendungen finden sich in den Wahrscheinlichkeitstheorie-Lehrbüchern von Feller und Grimmett/Stirzacker.

## Erzeugende Funktionen zufälliger Summen

Sind  $N, X_1, X_2, \dots : \Omega \rightarrow \{0, 1, 2, \dots\}$  unabhängige Zufallsvariablen, dann erhalten wir für die Summe  $S_N = \sum_{i=1}^N X_i$ :

$$G_{S_N}(s) = E[s^{S_N}] = E[E[s^{S_N} | N]] = E[G(s)^N] = G_N(G(s)), \quad (10.2.1)$$

wobei  $G$  die erzeugende Funktion der Summanden  $X_i$  ist. Für die Verteilung von  $S_N$  ergibt sich

$$P[S_N = k] = \frac{1}{k!} (G_N \circ G)^{(k)}(0) \quad \text{für alle } k \geq 0.$$

**Beispiel (Ausdünnungseigenschaft von Poissonverteilungen).** Ein Huhn lege eine mit Parameter  $\lambda > 0$  Poissonverteilte Anzahl  $N$  von Eiern, von denen aus jedem unabhängig voneinander und von  $N$  mit Wahrscheinlichkeit  $p$  ein Küken schlüpfe. Die erzeugende Funktion der Poissonverteilung ist

$$G_N(s) = E[s^N] = \sum_{k=0}^{\infty} s^k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = e^{\lambda(s-1)}.$$

Die Anzahl der geschlüpften Küken ist  $S_N = \sum_{i=1}^N X_i$ , wobei die  $X_i$  untereinander und von  $N$  unabhängige, Bernoulli( $p$ )-verteilte Zufallsvariablen sind. Wir erhalten also

$$G_{S_N}(s) = G_N(G_{X_1}(s)) = G_N(1 - p + p \cdot s) = e^{p\lambda \cdot (s-1)},$$

d.h. die Zahl der geschlüpften Küken ist wieder Poissonverteilt mit Parameter  $p \cdot \lambda$ . eine ausgedünnte Poissonverteilung ist also wieder eine Poissonverteilung!

## Galton-Watson-Verzweigungsprozesse

Wir betrachten das folgende Modell für ein zufälliges Populationswachstum: Alle Individuen der Population erzeugen unabhängig voneinander eine zufällige Anzahl von Nachkommen in der nächsten Generation mit Verteilung  $\nu$ . Hierbei sei  $\nu$  eine feste Wahrscheinlichkeitsverteilung auf  $\{0, 1, 2, \dots\}$  mit  $\nu[\{2, 3, \dots\}] \neq 0$ . Dieses Modell wurde 1889 von Galton und Watson eingeführt, um die Aussterbewahrscheinlichkeit englischer Adelstitel zu untersuchen. Ähnlich wie beim Random Walk handelt es sich um ein fundamentales stochastisches Modell mit unzähligen Erweiterungen und Anwendungen, z.B. auf das Wachstum von Zellpopulationen, die Ausbreitung von Epidemien, die Zunahme der Neutronenzahl in einem Reaktor, oder auch die näherungsweise Berechnung von genetischen Abständen oder der Anzahl von Zuständen in einem großen zufälligen Graphen (z.B. dem Internet), die man in einer bestimmten Anzahl von Schritten erreichen



kann. Die Nachkommensstruktur eines einzelnen Individuums bestimmt einen zufälligen verwurzelten Baum, s. Grafik 10.2. Dementsprechend spielen Verzweigungsprozesse auch eine zentrale Rolle bei der Analyse diverser stochastischer Modelle auf Bäumen, s. z.B. [Peres: Probability on trees].

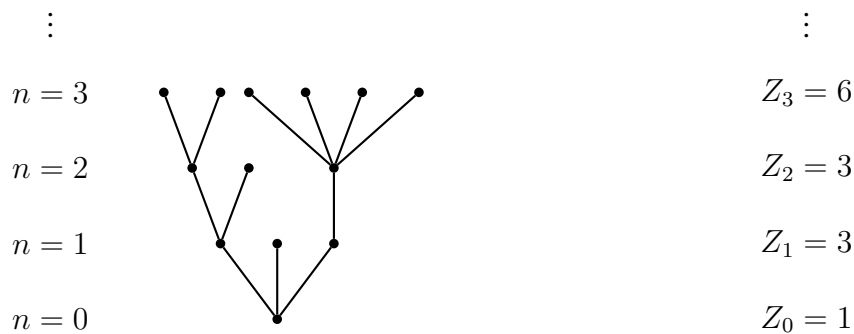


Abbildung 10.1: Beispiel für einen Galton-Watson-Prozess.

Wir beschreiben die Nachkommenszahlen der einzelnen Individuen in der  $(n - 1)$ -ten Generation eines Verzweigungsprozesses durch unabhängige Zufallsvariablen

$$N_i^n : \Omega \rightarrow \{0, 1, 2, \dots\}, \quad i = 1, 2, \dots,$$

mit Verteilung  $\nu$ . Für die Gesamtzahl der Individuen in der  $n$ -ten Generation erhalten wir die folgende rekursive Darstellung:

$$Z_n = \sum_{i=1}^{Z_{n-1}} N_i^n \quad \text{für alle } n \geq 1.$$

Ohne wesentliche Einschränkungen nehmen wir  $Z_0 = 1$  an – enthält die Anfangspopulation stattdessen mehrere Individuen, dann erzeugen diese voneinander unabhängige, identisch verteilte Unterpopulationen. Da  $Z_{n-1}$  nur von den Zufallsvariablen  $N_i^k$  für  $k \leq n - 1$  abhängt, sind  $Z_{n-1}$  und  $N_i^n$  ( $i \in \mathbb{N}$ ) unabhängige Zufallsvariablen. durch Bedingen auf  $Z_{n-1}$  erhalten wir für die mittleren Populationsgrößen die Rekursion

$$E[Z_n] = E[Z_{n-1}] \cdot m,$$

wobei  $m := \sum_{i=1}^{\infty} i \cdot \nu(i)$  die mittlere Nachkommenszahl eines Individuums ist. Wir unterscheiden die folgenden Fälle:

- $m > 1$  : Exponentielles Wachstum der Erwartungswerte (superkritisch)
- $m = 1$  : Erwartungswerte konstant (kritisch)
- $m < 1$  : Exponentieller Abfall der Erwartungswerte (subkritisch)

Wir wollen nun untersuchen, mit welcher Wahrscheinlichkeit die Population in den einzelnen Fällen ausstirbt. Nach (10.2.1) gilt für die erzeugenden Funktionen der Populationsgrößen die Rekursionsformel

$$G_{Z_n}(s) = E \left[ s^{\sum_{i=1}^{Z_{n-1}} N_i^n} \right] = G_{Z_{n-1}}(G(s)),$$

wobei  $G$  die erzeugende Funktion der Verteilung  $\nu$  der Anzahl  $N_i^n$  der Kinder eines Individuums ist. Per Induktion folgt wegen  $G_{Z_1}(s) = G(s)$ :

$$G_{Z_n}(s) = \underbrace{G(G(\dots G(s)\dots))}_{n\text{-mal}} = G^n(s) \quad \text{für alle } n \in \mathbb{N} \text{ und } s \in [0, 1].$$

Für die Wahrscheinlichkeiten  $\pi_n$ , dass der Prozess zur Zeit  $n$  ausgestorben ist, erhalten wir die Rekursionsformel

$$\pi_n = P[Z_n = 0] = G_{Z_n}(0) = G^n(0) = G(\pi_{n-1}) \quad (10.2.2)$$

Sei nun  $\pi$  die Wahrscheinlichkeit, dass die Population in endlicher Zeit ausstirbt. Da die Ereignisse  $\{Z_n = 0\}$  monoton wachsend sind, gilt

$$\pi = P \left[ \bigcup_n \{Z_n = 0\} \right] = \lim_{n \rightarrow \infty} \pi_n.$$

Da  $G$  auf  $[0, 1]$  stetig ist, folgt aus (10.2.2)

$$\pi = G(\pi),$$

d.h. die Aussterbewahrscheinlichkeit  $\pi$  ist ein Fixpunkt der erzeugenden Funktion. Wie oben bemerkt, ist die erzeugende Funktion  $G : [0, 1] \rightarrow [0, 1]$  strikt konvex mit  $G(1) = 1$  und  $G'(1-) = E[N_i^n] = m$ . Hieraus folgt, dass 1 im Fall  $m \leq 1$  der einzige Fixpunkt von  $G$  in  $[0, 1]$  ist, während im superkritischen Fall  $m > 1$  ein weiterer Fixpunkt  $\pi^* \in [0, 1)$  existiert, siehe auch Grafik ???. Aus den Skizzen erkennt man zudem, dass die Aussterbewahrscheinlichkeit  $\pi = \lim \pi_n$  der kleinste Fixpunkt von  $G$  auf  $[0, 1]$  ist. Also stirbt der Prozess im subkritischen bzw. kritischen Fall mit Wahrscheinlichkeit 1 aus, während er im superkritischen Fall mit einer strikt positiven Wahrscheinlichkeit überlebt.

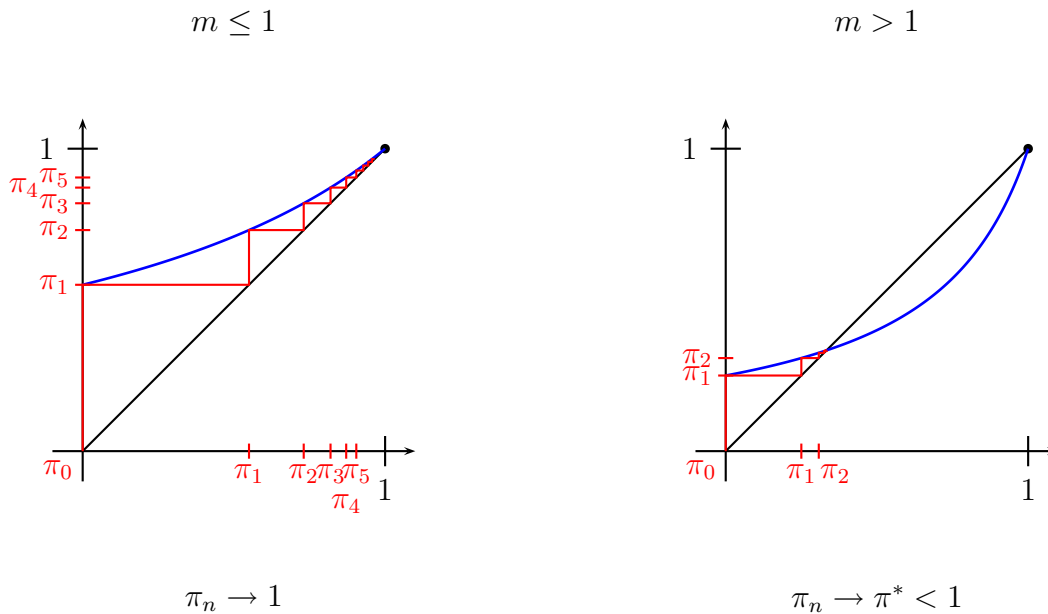


Abbildung 10.2: Erzeugendenfunktionen von Galton-Watson-Prozessen mit unterschiedlichen Verteilungen für die Anzahl der Nachkommen. In Rot: Fixpunktiteration

**Beispiel (Geometrische nachkommensverteilung).** Ist die Verteilung

$$\nu(k) = p^k(1 - p) \quad (k = 0, 1, 2 \dots)$$

der Anzahl der Nachkommen eine geometrische Verteilung mit Parameter  $p \in (0, 1)$ , dann ergibt sich

$$G(s) = \sum_{k=0}^{\infty} s^k p^k (1 - p) = \frac{1 - p}{1 - ps} \quad \text{für alle } s \in [0, 1].$$

Fixpunkte dieser Funktion sind 1 und  $\frac{1-p}{p}$ . Für  $1 - p \geq p$  (subkritischer Fall) ist 1 der einzige Fixpunkt in  $[0, 1]$ , also stirbt die Population  $P$ -fast sicher aus. Im superkritischen Fall  $1 - p < p$  beträgt die Aussterbewahrscheinlichkeit  $\pi$  dagegen nur  $\frac{1-p}{p}$ .

### Rekurrente Ereignisse und Erneuerungsgleichung

Als weitere Anwendung von erzeugenden Funktionen betrachten wir eine Folge von unvorhersehbaren Ereignissen, die zu diskreten Zeitpunkten  $n \in \mathbb{N}$  eintreten. Die Ereignisse bezeichnen wir auch als „Erneuerungen“ (engl. renewals), und denken dabei z.B. an den wiederholten Ausfall und Austausch eines Verschleißteils in einer Maschine, oder das wiederholte Abarbeiten einer

Warteschlange. Wir beschreiben den Zeitpunkt, an dem die  $k$ -te Erneuerung stattfindet, durch eine Zufallsvariable

$$S_k = T_1 + T_2 + \dots + T_k.$$

$T_1$  ist also der Zeitpunkt der ersten Erneuerung, und für  $k \geq 2$  ist  $T_k$  der zeitliche Abstand der  $(k-1)$ -ten unter der  $k$ -ten Erneuerung. In einem einfachen Modell nehmen wir an, dass  $T_1, T_2, \dots : \Omega \rightarrow \mathbb{N}$  unabhängige Zufallsvariablen sind, und, dass  $T_2, T_3, \dots$  identisch verteilt sind (aber nicht  $T_1$ !). Wir wollen nun die Wahrscheinlichkeiten  $p_n$  der Ereignisse

$$A_n = \{\exists k \in \mathbb{N} : S_k = n\} \quad \text{„Erneuerung zur Zeit } n\text{“}$$

aus den Verteilungen der Wartezeiten berechnen. Bedingen auf den Wert von  $T_1$  liefert für  $n \geq m$ :

$$\begin{aligned} P[A_n | T_1 = m] &= P[\exists k \in \mathbb{N} : T_1 + \dots + T_k = n | T_1 = m] \\ &= P[\exists k \in \mathbb{N} : T_2 + \dots + T_k = n - m | T_1 = m] \\ &= P[\exists k \in \mathbb{N} : T_2 + \dots + T_k = n - m], \end{aligned}$$

und damit

$$P[A_n | T_1 = m] = P[A_{n-m+1} | T_1 = 1] = P[A_{n-m+1} | A_1].$$

Nach der Formel von der totalen Wahrscheinlichkeit erhalten wir für  $n \in \mathbb{N}$ :

$$p_n = \sum_{m=1}^n q_{n-m} \cdot P[T_1 = m] \quad (10.2.3)$$

mit  $q_n := P[A_{n+1} | A_1]$ . Um die bedingten Wahrscheinlichkeiten  $q_n$  zu berechnen, bedingen wir zusätzlich auf  $T_2$ . Da  $T_2, T_3, \dots$  unabhängig und identisch verteilt sind, gilt für  $n \geq m$ :

$$\begin{aligned} P[A_{n+1} | A_1 \cap \{T_2 = m\}] &= P[\exists k \in \mathbb{N} : T_1 + \dots + T_k = n + 1 | T_1 = 1, T_2 = m] \\ &= P[\exists k \geq 2 : T_3 + \dots + T_k = n - m | T_1 = 1, T_2 = m] \\ &= P[\exists k \geq 2 : T_3 + \dots + T_k = n - m] \\ &= P[\exists k \geq 2 : T_2 + \dots + T_{k-1} = n - m] \\ &= P[A_{n-m+1} | A_1] = q_{n-m}. \end{aligned}$$

Wegen

$$q_n = P[A_{n+1} | A_1] = \sum_{m=1}^n P[A_{n+1} | A_1 \cap \{T_2 = m\}] \cdot P[T_2 = m]$$

erhalten wir

$$q_n = \sum_{m=1}^n q_{n-m} \cdot P[T_2 = m] \quad \text{für alle } n \geq 1. \quad (10.2.4)$$

Die Gleichungen (10.2.3) und (10.2.4) heißen *Erneuerungsgleichungen*. Auf den rechten Seiten dieser Gleichungen stehen (wegen  $T_1, T_2 \geq 1$ ) die Faltungen der Folge  $q_n, n \in \mathbb{N}$  mit der Folge der Gewichte der Wartezeiten  $T_1$  bzw.  $T_2$ . Daher ist es zweckmäßig, zu den erzeugenden Funktionen

$$G_p(s) = \sum_{n=1}^{\infty} p_n s^n$$

und

$$G_q(s) = \sum_{n=0}^{\infty} q_n s^n$$

überzugehen. Für  $|s| < 1$  erhalten wir aus (10.2.3)

$$G_p(s) = G_q(s) \cdot G_{T_1}(s).$$

Aus (10.2.4) ergibt sich, da die rechte Seite für  $n = 0$  verschwindet:

$$G_q(s) - 1 = \sum_{n=1}^{\infty} q_n s^n = G_q(s) \cdot G_{T_2}(s).$$

Es folgt  $G_q(s) = (1 - G_{T_2}(s))^{-1}$ , und damit

$$G_p(s) = \frac{G_{T_1}(s)}{1 - G_{T_2}(s)}. \quad (10.2.5)$$

(10.2.5) liefert den gesuchten Zusammenhang zwischen der Verteilung der Wartezeiten, und den Wahrscheinlichkeiten  $p_n$ , dass zur Zeit  $n$  eine Erneuerung stattfindet.

Sei nun die Verteilung der Lebensdauern  $T_2, T_3, \dots$  vorgegeben. Dann können wir untersuchen, welche Verteilung die Anfangswartezeiten  $T_1$  haben muss, damit die Wahrscheinlichkeiten  $p_n$  nicht von  $n$  abhängen (*Stationarität*). Für  $\alpha \in [0, 1]$  gilt  $p_n = \alpha$  für alle  $n \in \mathbb{N}$  genau dann, wenn

$$G_p(s) = \sum_{n=1}^{\infty} p_n s^n = \frac{\alpha}{1+s} \quad \text{für alle } s \in (-1, 1),$$

d.h. wenn

$$G_{T_1}(s) = \alpha \cdot \frac{1 - G_{T_2}(s)}{1-s} \quad \text{für alle } s \in (-1, 1). \quad (10.2.6)$$

Da  $G_{T_1}$  und  $G_{T_2}$  erzeugende Funktionen von Wahrscheinlichkeitsverteilungen sind, muss dann gelten:

$$\begin{aligned} 1 &= G_{T_1}(1) = \lim_{s \uparrow 1} G_{T_1}(s) \\ &= \alpha \lim_{s \uparrow 1} \frac{G_{T_2}(s) - 1}{s - 1} = \alpha G'_{T_2}(1-) \\ &= \alpha \cdot E[T_2]. \end{aligned}$$

Also muss  $T_2$  endlichen Erwartungswert haben, und

$$\alpha = 1/E[T_2] \quad (10.2.7)$$

gelten. Dies ist auch anschaulich plausibel: Im stationären Fall ist die Erneuerungswahrscheinlichkeit zu einem festen Zeitpunkt der Kehrwert des mittleren zeitlichen Abstandes zwischen zwei Erneuerungen. Gilt (10.2.7), dann ergibt sich aus (10.2.6) durch Koeffizientenvergleich:

$$P[T_1 = n] = \alpha \cdot \left( 1 - \sum_{k=1}^n P[T_2 = k] \right) = \frac{P[T_2 > n]}{E[T_2]}. \quad (10.2.8)$$

Die Folge  $p_n$  der Erneuerungswahrscheinlichkeiten ist also genau dann konstant, wenn die Verteilung von  $T_1$  durch (10.2.8) gegeben ist („stationärer Erneuerungsprozess“). weiter kann man ausgehend von (10.2.6) zeigen, dass für *beliebige* Verteilungen der ersten Erneuerungszeit die Wahrscheinlichkeiten  $p_n$  für  $n \rightarrow \infty$  gegen  $1/E[T_2]$  konvergieren („asymptotische Stationarität“), falls der Erwartungswert endlich ist und keine *Periodizität* auftritt, d.h.

$$\lim_{n \rightarrow \infty} P[T_2 = n] = 1.$$

Den Beweis dieses **Erneuerungssatzes** über erzeugende Funktionen findet man im Klassiker von W.Feller (An Introduction to Probability Theory and its Applications, Vol. 1).

### 10.3 Bedingen auf allgemeine Zufallsvariablen

Ist  $Y$  eine reellwertige Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  mit stetiger Verteilungsfunktion, dann gilt  $P[Y = z] = 0$  für alle  $z \in \mathbb{R}$ . Bedingte Wahrscheinlichkeiten gegeben  $Y = z$  können daher nicht wie für diskrete Zufallsvariablen definiert werden. Alternativ könnte man versuchen, bedingte Wahrscheinlichkeiten gegeben  $Y$  als Grenzwert zu definieren:

$$P[A | Y = z] = \lim_{h \rightarrow 0} P[A | z - h \leq Y \leq z + h]. \quad (10.3.1)$$

Dies ist in bestimmten Fällen möglich, aber im allgemeinen ist die Existenz des Grenzwertes nicht gewährleistet.

Stattdessen definiert man bedingte Erwartungen gegeben allgemeine Zufallsvariablen  $Y$  mithilfe der Charakterisierung aus Satz 10.5. Bedingte Wahrscheinlichkeiten gegeben  $Y$  erhält man als Spezialfall bedingter Erwartungen:

$$P[A | Y] := E[I_A | Y]. \quad (10.3.2)$$

Bedingte Wahrscheinlichkeiten wie in (10.3.1) sind im Allgemeinen nicht im herkömmlichen Sinn definiert. Es ist allerdings ausgehend von (10eq10) allgemein möglich für ein festes Ereignis  $A$  die Abbildung  $z \mapsto P[A | Y = z]$  bis auf Modifikation auf Nullmengen bzgl. der Verteilung von  $Y$  zu definieren.

### Das Faktorisierungslemma

Wir beweisen zunächst eine wichtige maßtheoretische Aussage. Diese wird es uns u.A. ermöglichen, die charakterisierenden Eigenschaften bedingter Erwartungen aus Satz 10.5 noch etwas eleganter zu formulieren:

**Satz 10.6 (Faktorisierungslemma).** Sei  $(S, \mathcal{S})$  ein messbarer Raum und  $Y : \Omega \rightarrow S$  eine Abbildung. Eine Abbildung  $X : \Omega \rightarrow \mathbb{R}$  ist genau dann  $\sigma(Y)$ -messbar, wenn

$$X = f(Y) = f \circ Y$$

für eine  $\mathcal{S}$ -messbare Funktion  $f : S \rightarrow \mathbb{R}$  gilt.

$$\begin{array}{ccc} & X & \\ & \curvearrowright & \\ (\Omega, \sigma(Y)) & \xrightarrow{Y} & (S, \mathcal{S}) \longrightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R})) \end{array}$$

*Beweis.* (1). Sei  $X = f \circ Y$  für eine messbare Funktion  $f$ , dann gilt

$$X^{-1}(B) = Y^{-1}(f^{-1}(B)) \in \sigma(Y) \quad \text{für alle } B \in \mathcal{B}(\mathbb{R}),$$

da  $f^{-1}(B) \in \mathcal{S}$ . Daher ist  $X$   $\sigma(Y)$ -messbar.

(2). Für die umgekehrte Richtung müssen wir zeigen, dass aus der  $\sigma(Y)$ -Messbarkeit von  $X$  folgt, dass  $X$  eine messbare Funktion von  $Y$  ist. Dazu gehen wir schrittweise vor („maßtheoretische Induktion“):

(a) Ist  $X = I_A$  eine Indikatorfunktion mit  $A \in \sigma(Y)$ , dann gilt  $A = Y^{-1}(B)$  mit  $B \in \mathcal{S}$ , und damit

$$X(\omega) = I_{Y^{-1}(B)}(\omega) = I_B(Y(\omega)) \quad \text{für alle } \omega \in \Omega.$$

(b) Für  $X = \sum_{i=1}^n c_i I_{A_i}$  mit  $A_i \in \sigma(Y)$  und  $c_i \in \mathbb{R}$  gilt entsprechend

$$X = \sum_{i=1}^n c_i I_{B_i}(Y),$$

wobei  $B_i$  Mengen aus  $S$  mit  $A_i = Y^{-1}(B_i)$  sind.

- (c) Für eine beliebige nichtnegative,  $\sigma(Y)$ -messbare Abbildung  $X : \Omega \rightarrow \mathbb{R}$  existiert eine Folge  $X_n$  von  $\sigma(Y)$ -messbaren Elementarfunktionen mit  $X_n \nearrow X$ . nach (b) gilt  $X_n = f_n(Y)$  mit  $\mathcal{S}$ -messbaren Funktionen  $f_n$ . Damit folgt:

$$X = \sup X_n = \sup f_n(Y) = f(Y),$$

wobei  $f = \sup f_n$  wieder  $\mathcal{S}$  messbar ist.

- (d) Für eine allgemeine  $\sigma(Y)$ -messbare Abbildung  $X : \Omega \rightarrow \mathbb{R}$  sind sowohl  $X^+$  als auch  $X^-$  messbare Funktionen von  $Y$ , also auch  $X$  selbst.

□

Mithilfe des Faktorisierungslemmas können wir die *charakterisierenden Eigenschaften* (I) und (II) bedingter ERwartungen gegeben eine diskrete Zufallsvariable  $Y$  aus Satz 10.5 wie folgt umformulieren:

$\bar{X}$  ist genau dann eine Version von  $E[X | Y]$ , wenn gilt:

(i)  $\bar{X}$  ist  $\sigma(Y)$ -messbar,

(ii)  $E[\bar{X}; A] = E[X; A]$  für alle  $A \in \sigma(Y)$ .

Die Äquivalenz von (I) und (i) folgt aus dem Faktorisierungslemma, und die Äquivalenz von (II) und (ii) ergibt sich durch maßtheoretische Induktion, denn (ii) besagt gerade, dass

$$E[\bar{X} \cdot I_B(Y)] = E[X \cdot I_B(Y)] \quad \text{für alle } B \in \mathcal{S} \text{ gilt.}$$

### Definition allgemeiner bedingter Erwartung

Eine bemerkenswerte Konsequenz der Charakterisierung bedingter Erwartungen durch die Bedingungen (i) und (ii) ist, dass die *bedingte ERwartung*  $E[X | Y]$  von der Zufallsvariablen  $Y$  nur über die von  $Y$  erzeugte  $\sigma$ -Algebra  $\sigma(Y)$  abhängt! Sind zwei Zufallsvariablen  $Y$  und  $Z$  Funktionen voneinander, dann ist  $\sigma(Y) = \sigma(Z)$ , und damit stimmen auch die bedingten Erwartungen  $E[X | Y]$  und  $E[X | Z]$  überein (mit Wahrscheinlichkeit 1). Daher liegt es nahe, gleich von der bedingten Erwartung gegeben eine  $\sigma$ -Algebra zu sprechen. Die  $\sigma$ -Algebra (z.B.  $\sigma(Y)$  oder  $\sigma(Y_1, \dots, Y_n)$ ) beschreibt dann die zur Verfügung stehende „Information“, auf die bedingt wird.

Die Charakterisierung bedingter Erwartung durch (i) und (ii) können wir sofort auf den Fall allgemeiner bedingter Erwartungen gegeben eine  $\sigma$ -Algebra oder gegeben beliebige Zufallsvariablen übertragen. Sei dazu  $X : \Omega \rightarrow \overline{\mathbb{R}}$  eine nichtnegative (oder integrierbare) Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ .



**Definition (Bedingte Erwartung, allgemein).** (1). Sei  $\mathcal{F} \subseteq \mathcal{A}$  eine  $\sigma$ -Algebra. Eine nicht-negative (bzw. integrierbare) Zufallsvariable  $\bar{X} : \Omega \rightarrow \bar{\mathbb{R}}$  heißt **Version der bedingten Erwartung**  $E[X | \mathcal{F}]$ , falls gilt:

- (a)  $\bar{X}$  ist  $\mathcal{F}$ -messbar, und  
 (b)  $E[\bar{X}; A] = E[X; A]$  für alle  $A \in \mathcal{F}$ .

(2). Für beliebige Zufallsvariablen  $Y, Y_1, Y_2, \dots, Y_n$  auf  $(\Omega, \mathcal{A}, P)$  definieren wir

$$\begin{aligned} E[X | Y] &:= E[X | \sigma(Y)], \\ E[X | Y_1, \dots, Y_n] &:= E[X | (Y_1, \dots, Y_n)] = E[X | \sigma(Y_1, \dots, Y_n)]. \end{aligned}$$

(3). Für ein Ereignis  $A \in \mathcal{A}$  definieren wir

$$P[A | \mathcal{F}] := E[I_A | \mathcal{F}], \quad \text{und entsprechend} \quad P[A | Y] = E[A | Y].$$

**Bemerkung.** Durch maßtheoretische Induktion zeigt man, dass Bedingung (b) äquivalent ist zu:

- (b')  $E[\bar{X} \cdot Z] = E[X \cdot Z]$  für alle nichtnegativen (bzw. beschränkten)  $\mathcal{F}$ -messbaren  $Z : \Omega \rightarrow \mathbb{R}$ .

**Satz 10.7 (Existenz und Eindeutigkeit der bedingten Erwartung).** Sei  $X \geq 0$  oder  $X \in \mathcal{L}^1$ , und  $\mathcal{F} \subseteq \mathcal{A}$  eine  $\sigma$ -Algebra.

- (1). Es existiert eine Version der bedingten Erwartung  $E[X | \mathcal{F}]$ .  
 (2). Zwei Versionen stimmen  $P$ -fast sicher überein.

*Beweis.* Die Existenz kann man unmittelbar aus dem Satz von Radon-Nikodym folgern, s. z.B. [A.Klenke, Wahrscheinlichkeitstheorie]. Wir geben stattdessen in Abschnitt ?? einen Existenzbeweis, der mit elementaren Methoden auskommt, siehe ??.

Zum Beweis der Eindeutigkeit seien  $\bar{X}$  und  $\tilde{X}$  zwei Versionen der bedingten Erwartung  $E[X | \mathcal{F}]$ . Dann sind  $\bar{X}$  und  $\tilde{X}$  beide  $\mathcal{F}$ -messbar, und

$$E[\bar{X}; A] = E[\tilde{X}; A] \quad \text{für alle } A \in \mathcal{F}.$$

Hieraus folgt  $\bar{X} = \tilde{X}$   $P$ -fast sicher. □

**Bemerkung (Probleme mit Ausnahmemengen).** Man beachte, dass die bedingte Erwartung  $E[X | \mathcal{F}]$  und damit auch die bedingte Wahrscheinlichkeit  $P[A | \mathcal{F}]$  nur für jede feste Zufallsvariable  $X$  bzw. jedes feste Ereignis  $A$  bis auf Modifikation auf Nullmengen eindeutig definiert

sind. Ein weiteres Problem ist, dass wir allgemein zwar bedingte Erwartungen gegeben eine Zufallsvariable  $Y$  definieren können, aber nicht solche gegeben das Ereignis  $Y = z$  für festes  $z$ . In vielen Fällen kann man die beschriebenen Probleme durch Auswahl einer „regulären Version der bedingten Verteilung gegeben  $Y$ “ umgehen. Wir kommen darauf in Korollar 10.3.6 zurück.

Obwohl  $E[X|Y = z]$  für festes  $z$  i.A. nicht definiert ist, kann man die Funktion  $z \mapsto E[X|Y = z]$  bis auf Modifikation auf Nullmengen bzgl. der Verteilung von  $Y$  sinnvoll definieren:

**$E[X | Y = z]$ , Definition und Warnung:**

Ist  $Y : \Omega \rightarrow S$  eine Zufallsvariable mit Werten in einem messbaren Raum  $(S; \mathcal{S})$ , dann ist jede Version der bedingten Erwartung  $E[X | Y]$  nach Definition  $\sigma(Y)$ -messbar. Also gilt nach dem Faktorisierungslemma:

$$E[X | Y] = g(Y) \quad \text{für eine messbare Funktion } g : S \rightarrow \mathbb{R}. \quad (10.3.3)$$

Da die Versionen der bedingten Erwartung bis auf Modifikation auf  $P$ -Nullmengen eindeutig festgelegt sind, ist die Funktion  $g$  bis auf Modifikation auf  $\mu_Y$ -Nullmengen eindeutig festgelegt. In Anlehnung an den diskreten Fall setzt man manchmal:

$$E[X | Y = z] := g(z). \quad (10.3.4)$$

Genauer definieren wir für eine nichtnegative Zufallsvariable  $X$ :

**Definition.** Eine messbare Funktion  $g : S \rightarrow \mathbb{R}^+$  heißt **Version der bedingten Erwartung**  $z \mapsto E[X | Y = z]$  von  $X$  gegeben  $Y = z$ , wenn gilt:

$$E[X ; Y \in B] = \int_B \delta(z) \mu_Y(dz) \quad \text{für alle } B \in \mathcal{S}. \quad (10.3.5)$$

Die charakterisierende Bedingung (10.3.5) ist nichts anderes als eine allgemeine Variante der **Formel von der totalen Wahrscheinlichkeit**. Mithilfe des Transformationssatzes sieht man, dass  $g$  genau dann (10.2.3) erfüllt, wenn  $g(Y)$  eine Versio von  $E[X | Y]$  ist.

**Warnung:** Bei der Definition ist zu beachten, dass  $E[X|Y = z]$  für ein festes  $z$  im Allgemeinen nicht definiert ist, sondern nur die Funktion  $z \mapsto E[X | Y = z]$  modulo Modifikation auf  $\mu_Y$ -Nullmengen! Das formale Rechnen mit bedingten Erwartungen wie in (10.3.4) ist daher eine häufige Fehlerquelle.

Trotz dieser Gefahren ist die Notation  $E[X | Y = z]$  oft nützlich, um Argumentationen transparenter zu machen, oder um anschauliche Überlegungen in mathematische Formeln zu übersetzen. Wir werden sie daher auch hier gelegentlich verwenden.

## Diskreter und absolutstetiger Fall

In einigen Fällen kann man die Definition direkt anwenden, um bedingte Erwartungswerte zu berechnen. Wir betrachten zunächst noch einmal den Spezialfall eine *diskreten Bedingung*:

Gilt  $\mathcal{F} = \sigma(\{H_i | i \in \mathbb{N}\})$  für eine disjunkte Zerlegung  $\Omega = \bigcup_{i \in \mathbb{N}} H_i$  in abzählbar viele messbare Teilmengen („Hypothesen“)  $H_i \in \mathcal{A}$ , dann sind  $\mathcal{F}$ -messbare Zufallsvariablen konstant auf jeder der Mengen  $H_i$ . aus der Definition der bedingten ERwartung folgt dann

$$E[X | \mathcal{F}] = E[X | H_i] \quad \text{auf } H_i$$

für alle  $i \in \mathbb{N}$  mit  $P[H_i] > 0$ .

**Beispiel (Unbedingte Erwartungen).** Die bedingte Erwartung einer Zufallsvariable  $X$  gegeben die triviale  $\sigma$ -Algebra  $\{\emptyset, \Omega\}$  ist der Erwartungswert von  $X$ .

**Beispiel (Bedingen auf eine Partition).** Ist  $P = \mathcal{U}_{[0,1]}$  die Gleichverteilung auf  $[0, 1)$ , und  $\mathcal{F} = \sigma(\{[t_{i-1}, t_i] | i = 1, \dots, n\})$  die von einer Partition  $0 = t_0 < t_1 < t_2 < \dots < t_n = 1$  erzeugte  $\sigma$ -Algebra, dann ist die bedingte Erwartung  $E[g | \mathcal{F}]$  einer integrierbaren Funktion  $g : [0, 1) \rightarrow \mathbb{R}$  durch

$$E[g | \mathcal{F}] = \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} g(u) du \quad \text{auf } [t_{i-1}, t_i)$$

definierte Funktion.

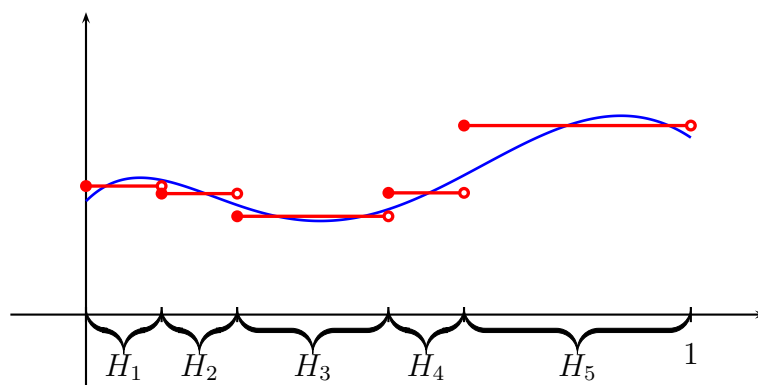


Abbildung 10.3: Die Dichte von  $X(\omega)$  ist hier blau dargestellt und  $E[X | \mathcal{F}]$  in rot.

Ist die gemeinsame Verteilung aller relevanten Zufallsvariablen absolutstetig, dann kann man bedingte Erwartungen mithilfe von bedingten Dichten berechnen:

**Satz 10.8 (Berechnung bedingter Erwartungen im absolutstetigen Fall).** Seien  $X : \Omega \rightarrow \mathbb{R}^n$  und  $Y : \Omega \rightarrow \mathbb{R}^m$  Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ , deren gemeinsame Verteilung  $\mu_{X,Y}$  absolutstetig ist, und sei  $h : \mathbb{R}^n \times \mathbb{R}^m \rightarrow [0, \infty]$  messbar. Dann ist

$$E[h(X, Y) | Y](\omega) = \int_{\mathbb{R}^n} h(x, Y(\omega)) f_{X|Y}(x|Y(\omega)) dx \quad (10.3.6)$$

eine Version der bedingten Erwartung von  $h(X, Y)$  gegeben  $Y$ .

*Beweis.* Nach dem Satz von Fubini ist die rechte Seite von (10.3.6) eine messbare Funktion von  $Y(\omega)$ , und es gilt

$$\begin{aligned} E[g(Y) \cdot \int h(x, Y) f_{X|Y}(x|Y) dx] &= \int \int g(y) h(x, y) f_{X|Y}(x|y) f_Y(y) dx dy \\ &= E[g(Y) h(X, Y)] \end{aligned}$$

für jede messbare Funktion  $g : \mathbb{R}^m \rightarrow [0, \infty]$ . □

Mit der Notation aus (10.3.4) lautet die Aussage des Satzes:

$$E[h(X, Y) | Y = z] = \int_{\mathbb{R}^n} h(x, z) f_{X|Y}(x|z) dx \quad \text{für } \mu_Y\text{-fast alle } z \in S.$$

Um die bedingte Erwartung zu berechnen, müssen wir also den uns bekannten Wert von  $Y$  einsetzen, und die Funktion bzgl. der bedingten Dichte  $f_{X|Y}$  nach  $x$  integrieren.

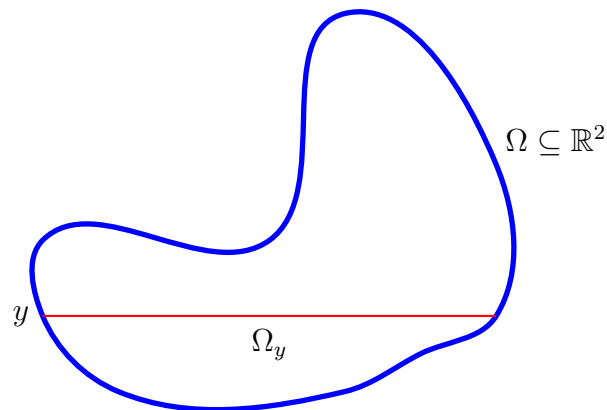
**Beispiel (Bedingen auf eine Koordinate).** Ist  $P = \mathcal{U}_\Omega$  die Gleichverteilung auf einer beschränkten, messbaren Menge  $\Omega \subseteq \mathbb{R}^2$ , und ist

$$Y : \Omega \rightarrow \mathbb{R}, \quad Y(x, y) = y,$$

die Projektion auf die zweite Komponente, dann gilt

$$E[h|Y](x, y) = \frac{1}{\lambda(\Omega_y)} \int_{S-y} h(x, y) dx \quad P\text{-fast sicher} \quad (10.3.7)$$

für jede integrierbare Funktion  $h : \Omega \rightarrow \mathbb{R}$ . Hierbei ist  $\Omega_y = \{x \in \mathbb{R} | (x, y) \in \Omega\}$  der  $y$ -Schnitt von  $\Omega$ . Bedingen auf  $Y$  entspricht hier also dem normierten „Herausintegrieren“ der komplementären Koordinate  $x$ .

Abbildung 10.4: In Rot: Der  $y$ -Schnitt der Menge  $\Omega$ .

### Reguläre bedingte Verteilungen

Beim Bedingen auf diskrete Zufallsvariablen konnten wir bedingte Wahrscheinlichkeitsverteilungen auf elementare Weise definieren. Für allgemeine Zufallsvariablen sind die bedingten Wahrscheinlichkeiten

$$P[X \in B | Y] = E[I_B | Y]$$

für jede feste messbare Menge  $B$  nur bis auf Modifikation auf  $P$ -Nullmengen eindeutig definiert. Dies ist ein Nachteil, da die Ausnahmemenge von  $B$  abhängen kann, und im Allgemeinen überabzählbar viele messbare Mengen existieren. Die bedingte Verteilung von  $X$  gegeben  $Y$  ist daher zunächst nicht definiert. Im absolutstetigen Fall können wir das Problem umgehen, indem wir die über die bedingte Dichte gegebene Version

$$\mu_{X|Y}(y, dx) := f_{X|Y}(x|y)dx$$

der bedingten Verteilung verwenden. Aus Satz 10.8 folgt unmittelbar, dass wir bedingte Wahrscheinlichkeiten gegeben  $Y$  aus  $\mu_{X|Y}$  berechnen können:

**Korollar 10.9.** *Ist die gemeinsame Verteilung der Zufallsvariablen  $X : \Omega \rightarrow \mathbb{R}^n$  und  $Y : \Omega \rightarrow \mathbb{R}^m$  absolutstetig, dann ist  $\mu_{X|Y}$  eine **reguläre Version der bedingten Verteilung von  $X$  gegeben  $Y$** , d.h.*

(1).  $\mu_{X|Y}$  ist ein stochastischer Kern von  $\mathbb{R}^m$  nach  $\mathbb{R}^n$ .

(2). Für jedes  $B \in \mathcal{B}(\mathbb{R}^n)$  ist

$$P[X \in B | Y] = \mu_{X|Y}(Y, B)$$

eine Version der bedingten Wahrscheinlichkeit von  $\{X \in B\}$  gegeben  $Y$ .

**Bemerkung (Existenz von regulären Versionen bedingter Verteilungen).** Die Existenz von regulären Versionen von bedingten Verteilungen gegeben eine Zufallsvariable  $Y$  kann man allgemein beweisen, wenn  $Y$  Werte in einem vollständigen, separablen, metrischen Raum (kurz: polnischen Raum) annimmt, siehe z.B. [Breiman, Ch. 4.3.]. Eine explizite Berechnung über bedingte Dichten ist natürlich im Allgemeinen nicht möglich.

Wenn wir uns auf eine bestimmte reguläre Version  $\mu_{X|Y}$  festlegen, dann können wir (10.3.2) als Definition der bedingten Wahrscheinlichkeiten  $P[X \in B|Y = z]$  für *alle*  $z \in S$  verwenden. Die Festlegung auf eine bestimmte reguläre Version der bedingten Verteilung ist im Allgemeinen willkürlich. Manchmal gibt es aber eine kanonische Version, die sich auszeichnet. Dies ist zum Beispiel der Fall, wenn die Dichte der gemeinsamen Verteilung von  $X$  und  $Y$  eine stetige Version hat.

**Beispiel (Bivariate Normalverteilung).** Ist  $(X, Y)$  bivariat normalverteilt mit Mittel  $(0, 0)$  und Kovarianzmatrix  $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ ,  $\rho \in (-1, 1)$ , dann gilt

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right).$$

Für ein festes  $x \in \mathbb{R}$  folgt

$$f_{Y|X}(y|x) \propto f_{X,Y}(x, y) \propto \exp\left(-\frac{(y - \rho x)^2}{2(1-\rho^2)}\right)$$

als Funktion von  $y$ . Also ist

$$\mu_{Y|X}(x, \bullet) = N(\rho x, 1 - \rho^2)$$

eine kanonische reguläre Version der bedingten Verteilung von  $Y$  gegeben  $X$ .

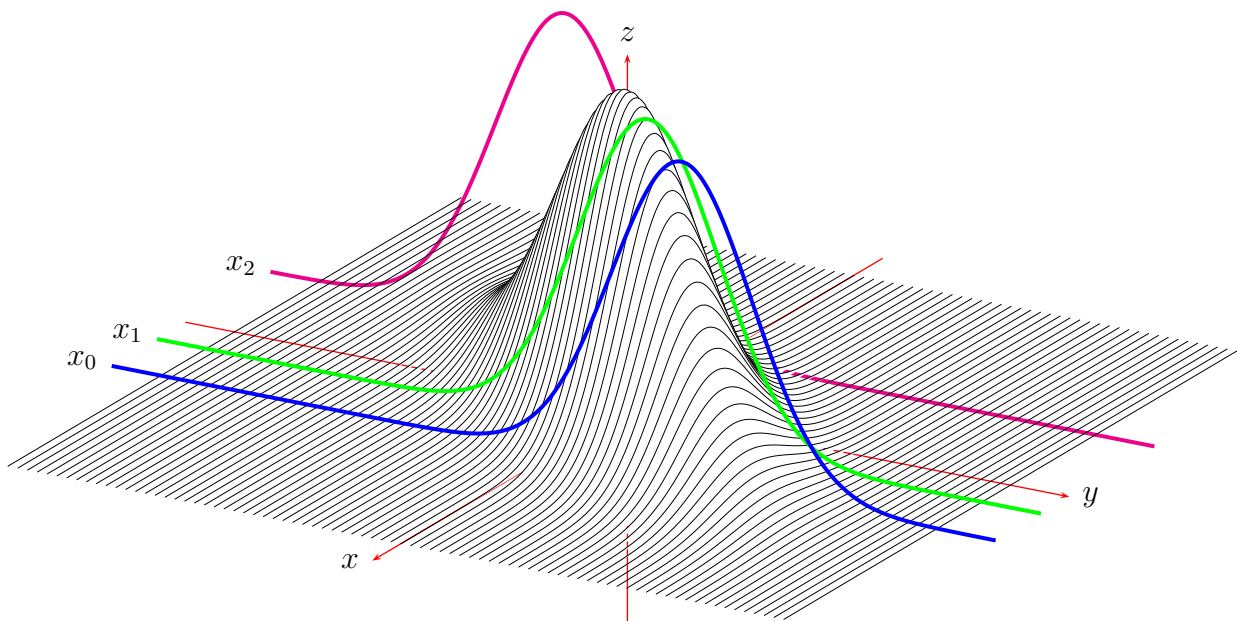


Abbildung 10.5: Die Dichte  $f_{X,Y}(x, y)$  und in Blau, Grün und Magenta  $f_{Y|X}(y|x_i)$  für ein  $i \in \{0, 1, 2\}$ . Man beachte, dass  $f_{Y|X}(y|x_i) \propto f_{X,Y}(x_i, y)$ .

**Beispiel (Grenzen naiven Bedingens).** Sei  $(X, Y)$  gleichverteilt auf dem Viertelkreis

$$S = \{(x, y) \in \mathbb{R}^2 \mid x > 0, y > 0, x^2 + y^2 < 1\}.$$

Wir versuchen auf zwei Arten eine „bedingte Verteilung von  $X$  gegeben  $X = Y$ “ zu berechnen. Dazubetrachten wir die Zufallsvariablen  $V = Y - X$  und  $W = Y/X$ . Wegen  $f_{X,Y} \propto I_S$  erhalten wir mithilfe des Dichtetransformationssatzes für fast jedes  $v$ :

$$\begin{aligned} f_{X|V}(x|v) &\propto f_{X,V}(x, v) = f_{X,Y}(x, v+x) \cdot \left| \det \frac{\partial(x, v+x)}{\partial(x, v)} \right| \\ &\propto I_S(x, v+x), \end{aligned}$$

wobei „ $\propto$ “ für „proportional als Funktion von  $x$ “ steht. Wählen wir die normierte rechte Seite als kanonische Version der bedingten Dichte, so ergibt sich

$$f_{X|V}(x|0) \propto I_S(x, x) = I_{(0,1/\sqrt{2})}(x).$$

Gegeben  $Y - X = 0$  ist  $X$  also gleichverteilt auf  $(0, 1/\sqrt{2})$ .

Andererseits erhalten wir für fast jedes  $w$ :

$$\begin{aligned} f_{X|W}(x|w) &\propto f_{X,W}(x, w) = f_{X,Y}(x, wx) \cdot \left| \det \frac{\partial(x, wx)}{\partial(x, w)} \right| \\ &\propto I_S(x, wx) \cdot x. \end{aligned}$$

Wählen wir wieder die rechte Seite als kanonische Version, so ergibt sich

$$f_{X|W}(x|1) \propto I_S(x, x)\dot{x} = x \cdot I_{(0,1/\sqrt{2})}(x).$$

Die bedingte Verteilung von  $X$  gegeben  $Y/X = 1$  unterscheidet sich also von der bedingten Verteilung von  $X$  gegeben  $Y - X = 0$ . Bedingte Wahrscheinlichkeiten gegeben  $X = Y$  sind daher nicht wohldefiniert!

Eine anschauliche Erklärung für das Phänomen ist, dass wir in den beiden Fällen oben auf unterschiedliche infinitesimale Umgebungen der Diagonale  $\{(x, y) \in S | x = y\}$  bedingen, wie die folgende Grafik veranschaulicht:

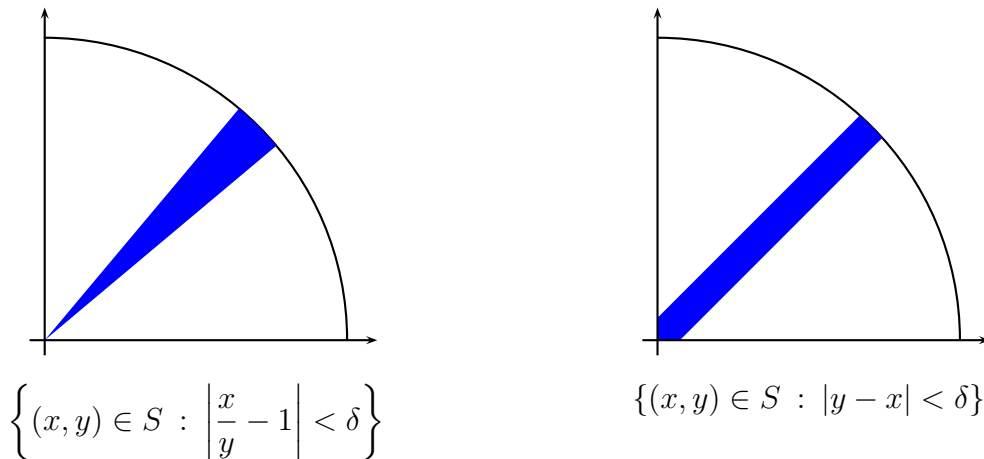


Abbildung 10.6: Zwei verschiedene Arten die Diagonale zu approximieren



# Stichwortverzeichnis

- 0-1 Gesetz von Kolmogorov, 181
- 0-1-Experimente
  - abhängige, 40
  - unabhängige, 40, 50
- $\sigma$ -Additivität, 12
- $\sigma$ -Algebra, 11
- a posteriori degree of belief, 46
- a priori degree of belief, 46
- abhängige 0-1-Experimente, 40
- absolutstetig, 202
- Acceptance-Rejection-Verfahren, 73
- Additivität, endliche, 12
- Akzeptanzwahrscheinlichkeit, 72
- Akzeptanzzeit, 73
- Algebra, 114
- arithmetisches Mittel, 193
- asymptotisch
  - e Zufallsvariable, 184
- asymptotische Äquivalenz von Folgen, 67
- Atome, 128
- Bayessche Regel, 46
- Bayessche Statistik, 46
- Bedingte Erwartung, 329
  - Definition
    - Diskrete -, 313
- bedingte Erwartung, 43
- bedingte Verteilung, 43
- bedingte Wahrscheinlichkeit, 43
- Benfordsches Gesetz, 20
- Bernoulli-Verteilung, 40
  - n-dimensionale, 50
- Bernstein-Ungleichung, 58
- Bias, 207
- Bildmaß, 121
- Binomialverteilung, 25
  - Poissonapproximation, 26
  - Varianz, 80
- Brown'sche Bewegung, 304
- Brownsche Bewegung, 112, 304
- Cauchy-Schwarz-Ungleichung in  $\mathcal{L}^2$ , 77
- Čebyšev-Ungleichung, 82
- Charakteristische Funktion
  - Ableitungen der -, 252
  - Lévys Inversionsformel, 253
- charakteristische Funktion, 249
- Cramér-Wold Device, 303
- degree of belief
  - a posteriori, 46
  - a priori, 46
- Detailed Balance-Bedingung, 89
- Dichte
  - bedingte -, 293
  - Wahrscheinlichkeits-, 128, 198
- diskrete Zufallsvariable, 22

- gemeinsame Verteilung, 63
- Unabhängigkeit, 63
- diskretes Modell, 11
  - mehrstufiges, 47
- durchschnittsstabil, 114, 115
- Dynkinsystem, 117
  - das von  $\mathcal{J}$  erzeugte -, 117
- Ehrenfest-Modell, 53, 90
- Einschluss-/Ausschlussprinzip, 14
- Elementarereignis, 8
- empirische Mittel, 204
- empirische Varianz, 204
- empirische Verteilung, 18, 236
- empirische Verteilungsfunktion, 236
- empirisches Mittel, 234
- Entropie, 241
- Ereignis, 8
  - Verteilungen für unabhängige Ereignisse, 57
  - asymptotisches -, 180
  - Elementar-, 8
  - Ereignisse und ihre Wahrscheinlichkeit, 10
  - Indikatorfunktion, 36
  - Unabhängigkeit, 55
- Erfolgswahrscheinlichkeit, 24
- Ergodensatz, 100
- Erneuerungsgleichung, 325
- Erneuerungsprozess
  - stationärer -, 326
- Erwartung, bedingte, 43
- Erwartungswert, 36
  - elementarer ZVn, 187
  - der Poissonverteilung, 37
  - Linearität, 39
  - Monotonie, 39
- Erzeugende Funktion, 319
- Euler'sche Beta-Funktion, 308
- Faltung von  $W'$ -Verteilungen, 296
- Faltungshalbgruppe, 297
- Fehler
  - 1. und 2. Art, 310
- Fluss in Markovketten, 89
- Fouriertransformation, 250
- gemeinsame Verteilung, 63, 166
- geometrische Verteilung, 57
- Gesetz der großen Zahlen, 58
  - für Markov-Ketten, 100
  - schwaches, 82
  - starkes, 82
- Gesetz großer Zahlen
  - für Bernoulli-Experimente, 106
  - Starkes -
    - ohne Integrierbarkeit, 234
    - Kolmogorovs -, 230
- gewichtetes Mittel, 38
- Gewichtung der möglichen Fälle, 14
- Gibbs-Sampler, 93
- Gleichgewichte von Markov-Ketten, 88
- Gleichgewichtsverteilung, 89
  - Konvergenz, 97
- Gleichverteilung, 17
  - reellwertiger Zufallsvariablen, 71
  - Simulation, 28
- Häufigkeitsverteilung der Anfangsziffern von Zahlen, 21
- Histogramm, 238
- hypergeometrische Verteilung, 28, 49
- Hypothese

- Alternativ-, 309  
 Null-, 309  
 Hypothesen, 44  
 Hypothesentest, 310  
  
 Importance Sampling, 85  
 Indikatorfunktion, 120  
 Indikatorfunktion einer Ereignisses, 36  
 Inverse  
     linksstetige verallgemeinerte -, 136  
 irreduzible stochastische Matrix, 98  
  
 kanonisches Modell, 179  
 Kern, stochastischer, 51  
 Konfidenzintervall, 160, 306, 311  
 Konfidenzniveau, 160  
 Kongruenzgenerator, linearer, 29  
 konsistente Schätzfolge, 83  
 Konvergenz  
     - in Verteilung, 262  
     fast sicher -, 217  
     schnelle stochastische -, 219  
     schwache -, 262  
     stochastische -, 217  
 Konvergenz ins Gleichgewicht, 96, 97  
 Konvergenz, stochastische, 82  
 Konvergenzsatz für endliche Markov-Ketten, 100  
 Korrelationskoeffizient, 78  
 Korrelationskoeffizienten, 208  
 Kovarianz, 78, 208  
 Kumulantenerzeugende Funktion, 258  
 kumulative Verteilungsfunktion, 71  
  
 $\mathcal{L}^2$ -Raum von diskreten Zufallsvariablen, 77  
 $\mathcal{L}^2$ -Skalarprodukt, 77  
 Lévy's Inversionsformel, 253  
  
 Laplace-Modell, 17  
 Laplacetransformation, 250  
 Legendre-Fenchel-Transformation, 258  
 Lemma  
     - von Borel-Cantelli  
         1. Teil, 104  
         2. Teil, 105  
     - von Fatou, 193  
 likelihood, 46  
 linearer Kongruenzgenerator, 29  
  
 Münzwurf, 9  
     abhängige Münzwürfe, 52  
     endlich viele faire Münzwürfe, 17  
     Markov-Kette, 90  
     zwei faire Münzwürfe, 56  
 Markov-Kette, 51  
     bei einem Münzwurf, 90  
     Bewegungsgesetz, 51  
     Fluss, 89  
     Gesetz der großen Zahlen, 100  
     Gleichgewicht, 88  
     Konstruktion mit vorgegebenen Gleichgewichtsverteilungen, 92  
     Konvergenzsatz für endliche Markov-Ketten, 100  
     Metropolis-Kette, 93  
     Monte Carlo-Verfahren, 100  
     Simulation mit vorgegebenem Gleichgewicht, 95  
     Stationarität, 89  
     zeitlich homogene, 88  
 Massenfunktion, 14, 122  
     einer diskreten Zufallsvariable, 22  
     eines mehrstufigen diskreten Modells, 47

- Matrix
- stochastische / Übergangs-, 88
  - irreduzible stochastische, 98
  - stochastische, 51
  - Stochastische -, 288
- Median, 136
- mehrstufiges diskretes Modell, 47
- Markov-Kette, **siehe** Markov-Kette
  - Produktmodell, 50
  - Wahrscheinlichkeitsverteilung, 47
- Menge aller möglichen Fälle, 8
- messbar
- e Abbildung, 119
- messbarer Raum, 114
- Messraum, 114
- Metropolis-Algorithmus, 95
- Metropolis-Kette, 93
- Konvergenz, 100
- Minorisierungsbedingung, 97
- Mischung, 289
- Mittel
- arithmetisches, 38
  - gewichtetes, 38
- Modell
- Bayes'sches -, 294
  - Ehrenfest-, 262
- Moment
- p-te -, 198
- Momentenerzeugende Funktion
- logarithmische -, 258
  - Reihenentwicklung der -, 252
- momentenerzeugende Funktionen, 249
- Monte Carlo-Schätzer, 75, 83
- Approximationsfehler, 75
  - eines mehrdimensionalen Integrals, 84
  - erwartungstreuer, 75
  - für Wahrscheinlichkeiten, 84
  - mittlere quadratische Fehler, 75
- Monte Carlo-Verfahren, 75
- für Markov-Ketten, 100
- Monte-Carlo
- Approximation, 267
- Multinomialkoeffizient, 240
- Nullmenge, 101
- Ordnungsstatistik, 160, 172
- $P$ -fast sicher, 101
- Paradoxon
- Sankt-Petersburg-, 38
  - Simpson-, 45
- Periode eines Zustands, 98
- Periodizität, 326
- Perkolation, 183
- Permutationen
- zufällige, **siehe** Zufallspermutationen
- Poissonapproximation der Binomialverteilung, 26
- Poissonverteilung, 27
- Erwartungswert, 37
- Produkt
- von Wahrscheinlichkeitsverteilungen, 178
- Produkt von Wahrscheinlichkeitsverteilungen, 50
- Produktmaß
- endliches -, 162
- Produktmodell, 50
- Prozess
- Autoregressiver -, 291
  - autoregressiver -, 214

- Ornstein-Uhlenbeck-, 291
- Poisson-, 300
- Pseudo-Zufallszahlengenerator, 28
- QQ-Plot, 237
- Quantil, 135
- Stichproben-, 136
- Quantil-Quantil-Plot, 237
- Rückkehrzeit, 67
- Random Walk, 67, 228, 304
- auf den ganzen Zahlen, 64
- auf einem Gitter, 52
- auf Graphen, 91
- Bewegungsverlauf, 67
- Rekurrenz, 182
- Rekurrenz von -s, 181
- symmetrischer, 67
- Trefferzeit, 67
- unbeschränkte Oszillation von -s, 182
- Verteilung der Positionen zur Zeit  $n$ , 66
- zyklischer, 90
- Randverteilung, 162
- reellwertige Zufallsvariable, 71
- gleichverteilt, 71
- Unabhängigkeit, 71
- Reflektionsprinzip, 68
- Relative Kompaktheit, 274
- renormierte Stichprobenvarianz, 235
- Rucksackproblem, 94
- Sankt-Petersburg-Paradoxon, 38
- Satz
- vom iterierten Logarithmus, 229
- von Berry-Esséen, 280
- von Bochner, 253
- von Chernoff, 257
- von Fubini, 288
- von Helly-Bray, 272
- von Lebesgue, 194
- von Prohorov, 272
- von de Moivre/Laplace, 144
- von der majorisierten Konvergenz, 194
- 0-1 - von Kolmogorov, 181
- Eindeutigkeits-, 115
- Formel von der totalen Wahrscheinlichkeit, 44
- Fortsetzungs- von Carathéodory, 115
- Konvergenz- von Lévy, 273
- Lévys Inversionsformel, 253
- Lemma von Fatou, 193
- Quellenkodierungs- von Shannon, 247
- Skorokhod - Darstellung, 268
- Stetigkeits-, 273
- Transformations-, 195
- Eindimensionaler Dichte-, 133
- Mehrdimensionaler Dichte-, 300
- Zentraler Grenzwert-
- $\mathcal{L}^2$ -Version, 276
- von Lindeberg-Feller, 282
- Multivariater -, 303
- Schätzer, 160, 305
- erwartungstreuer -, 306
- konsistenter -, 306
- Schätzfolge
- konsistente, 83
- Schwaches Gesetz der großen Zahlen, 82
- Selbstbefruchtung von Pflanzen, 52
- Shift-Register-Generatoren, 34
- $\sigma$

- Additivität, 103
- Stetigkeit, 103
- Subadditivität, 104
- $\sigma$ -Additivität von Wahrscheinlichkeitsverteilungen, 12
- $\sigma$ -Algebra
  - asymptotische -, 180
  - Borel'sche -, 113
  - die von  $\mathcal{I}$  erzeugte -, 113
  - Produkt-, 114
- $\sigma$ -endlich, 202
- Signalverarbeitung, 294
- Simpson-Paradoxon, 45
- Simulated Annealing, 95
  - Algorithmus, 96
- Simulation
  - exponentialverteilter ZVn, 124
- Simulation einer diskreten Verteilung
  - direkt, 72
- Simulation einer Markov-Kette mit vorgegebenem Gleichgewicht, 95
- Simulation von Gleichverteilungen, 28
- Simulationsverfahren, 71
  - Acceptance-Rejection-Verfahren, 72
  - direktes Verfahren, 71
- Standardabweichung, 76
- starkes Gesetz der großen Zahlen, 82
- Stationarität von Markov-Ketten, 89
- Statistik, 160
- Stichprobe
  - nquantil, 136
  - empirische Verteilung der -, 136
- Stirlingsche Formel, 66, 143
- stochastische Konvergenz, 82
- stochastische Matrix, 51, 88
  - irreduzibel, 98
- Stochastischer Kern, 287
- stochastischer Kern, 51
- symmetrischer Random Walk, 67
- Tail
  - event, 180
  - field, 180
- Test
  - Gütefunktion eines -s, 311
  - Hypothesen-, 311
  - Macht eines -s, 311
  - Niveau eines -s, 311
  - $t$ -, 310
- Transformationssatz, 37
- Trefferzeit, 67
  - Verteilung, 68
- Übergangsmatrix, 88
- unabhängige 0-1-Experimente, 40, 50
- Unabhängige Zufallsvariablen, 63
- Unabhängigkeit, 43
  - von Mengensystemen, 152
  - von Zufallsvariablen, 155
- Ereignis
  - Verteilung, 57
- reellwertiger Zufallsvariablen, 71
- von Ereignissen, 55
- Unabhängigkeit von diskreten Zufallsvariablen, 63, 64
- Unabhängigkeit von Ereignissen, 25, 56
- Ungleichung
  - Čebyšev-, 221
  - Cauchy-Schwarz-, 206, 209
  - Čebyšev-, 82, 220

- Exponentielle Čebyšev-Markov-, 221  
 Jensen'sche -, 222  
 Markov-, 220  
 Unkorreliertheit, 79  
 Vandermonde-Identität, 297  
 Varianz, 76  
   Definition, 203  
   der Binomialverteilung, 80  
   Reduktion durch Importance Sampling, 85  
   Stichproben-, 307  
   von Summen, 80  
 Variationsdistanz von Wahrscheinlichkeitsverteilungen, 96  
 Verteilung  
    $\alpha$ -stabile -, 282  
   - einer Zufallsvariablen, 121  
   -sfunktion, 122  
   bedingte, 43  
   bedingte -, 293  
   Beta-, 173  
   Cauchy-, 135  
   direkte Simulation einer diskreten Verteilung, 72  
   empirische -, 236, 267  
   Exponential-, 123, 129, 205  
   für unabhängige Ereignisse, 57  
   Gamma-, 299  
   Gleich-, 129  
   invariante -, 291  
   Multinomial-, 240  
   Normal-, 130  
   Rand-, 162  
   Standardnormal-  
     mehrdimensionale -, 165  
   stationäre -, 291  
   stetige -, 128  
   Students- $t$ -, 308  
   Uniforme -, 129  
    $\chi^2$ -, 305  
 Verteilungsfunktion, kumulative, 71  
 Verwerfungsbereich, 310  
 Würfelwurf, 23  
 Wahrscheinlichkeit, 8  
   Akzeptanz-, 72  
   bedingte, 43  
   Erfolgs-, 24  
 Wahrscheinlichkeits  
   -maß  
     Faltung von -en, 296  
     straffe Folge von -en, 271  
 Wahrscheinlichkeitsraum, 12  
 Wahrscheinlichkeitsverteilung, 12, 14, 121  
   einer diskreten Zufallsvariable, 22  
   der Anfangsziffern von Zahlen, 21  
   der Trefferzeiten, 68  
   des Maximums, 70  
   diskrete, 14  
   eines mehrstufigen diskreten Modells, 47  
   endliche Additivität, 12  
   gemeinsame, 63  
   geometrische, 57  
   Gleichverteilung / Laplace-Modell, 17  
   Produkt, 50  
   Variationsdistanz, 96  
 Warteschlange, 26  
 Wartezeit, 298  
 wesentlich, 245

Ziehen mit Zurücklegen, **siehe** Binomialverteilung

Ziehen ohne Zurücklegen, **siehe** hypergeometrische Verteilung

Zufallspermutationen, 35

Zufallsvariable, 9, 22, 119

- asymptotische -, 184
- austauschbare -n, 318
- diskrete, 22
- Elementare -n, 186
- reellwertige, 37, 71
- Standardabweichung, 76
- unabhängige, 63
- Varianz, 76

Zufallsvorgang, 8

- diskreter, 10

Zufallszahlen aus  $[0,1)$ , 35

Zufallszahlengenerator, 28, 71

- Kombinationen, 35

zyklischer Random Walk, 90

Zylindermenge, 114