

# Stochastik für Informatiker – Vorlesungsskript

## Informationen zu dieser Version

Das vorliegende Skript zur Vorlesung „Stochastik für Informatik/Stochastik für Informatiker“ an der TU Berlin ist der Stand vom 10. Juli 2019. Im Vergleich zur Version vom 1. April wurden alle Kapitel überarbeitet. Es fehlen jetzt vor allem noch einige Lösungen sowie ein paar wenige Bilder. Das Skript wird in naher Zukunft als Lehrbuch überarbeitet werden (zu erscheinen voraussichtlich 2020 bei Springer Vieweg), die aktuelle Version enthält immer noch Fehler. Meldungen von Fehlern und weitere Hinweise nehme ich immer noch sehr gern entgegen.

Berlin, 10. Juli 2019, Noemi Kurt

## Einleitung

Das Wort *Stochastik* bezeichnet das mathematische Fachgebiet, welches sich mit den Gesetzmäßigkeiten des Zufalls beschäftigt. Die Stochastik unterteilt sich grob in die *Wahrscheinlichkeitstheorie* und die *Statistik*. Die Wahrscheinlichkeitstheorie bildet die theoretische Grundlage für die mathematische Formalisierung, Modellierung und Untersuchung von zufälligen Abläufen. Die Statistik hat die Analyse und Interpretation von Messwerten und Daten aus zufälligen Vorgängen zum Inhalt.

Der angesprochene *Zufall* kann dabei entweder ein “echter”, inhärenter Zufall oder ein lediglich angenommener Zufall sein. Letzterer kann zum Beispiel aus einem Mangel an Information herrühren, oder als Effekt vieler kleiner (einzeln nicht genau kontrollierbarer) Ereignisse auftreten. Für die mathematische Behandlung ist es dabei nicht relevant, woher die Zufälligkeit eines Prozesses genau stammt. Deshalb kann die philosophische Frage über die Existenz von echtem Zufall umgangen und gleichzeitig eine universelle Anwendbarkeit der Theorie erreicht werden.

Im ersten Teil dieses Buches werden wir uns mit der Wahrscheinlichkeitstheorie beschäftigen, da diese das Fundament bildet, auf dem alle Anwendungen aufbauen. Gleichzeitig liefert sie die Sprache, in die Probleme aus der Wirklichkeit übersetzt werden (müssen), um die Macht der Mathematik anwenden zu können. Dabei wird notgedrungen die Wirklichkeit idealisiert und formalisiert.

Für Anwendungen, welche insbesondere im zweiten und dritten Teil des Buches angesprochen werden, ist es deshalb wichtig, dass das mathematische Modell immer wieder mit der Wirklichkeit abgeglichen und gegebenenfalls angepasst wird. Dies geschieht in vielen Fällen mit Hilfe der Statistik. Teil 2 bietet deshalb einen Überblick zu den wichtigsten klassischen Grundproblemen der Statistik.

Wegen ihrer Universalität ist die Wahrscheinlichkeitstheorie selbst in Bereichen nützlich, in denen a priori gar kein Zufall herrscht. Jedoch kann eine künstliche Zufälligkeit z.B. bei Suchalgorithmen enorm hilfreich sein. Im dritten Teil des Buches werden wir einige Anwendungen kennenlernen, welche auf wahrscheinlichkeitstheoretischen Methoden basieren. Ein besonderer Fokus liegt dabei auf einem algorithmischen Ansatz.

Ein großer Anwendungsbereich der Stochastik und Statistik in unserer Zeit liegt im Bereich der künstlichen Intelligenz und von Big Data. Dieses Buch hat dabei nicht den Anspruch, in diese Gebiete einzuführen. Es soll jedoch die wahrscheinlichkeitstheoretischen und statistischen Grundlagen bereitstellen, die es Leser\*innen leicht machen sollte, Literatur oder Vorlesungen zu diesem Themenbereich zu verstehen und sich in diesen Bereichen zu vertiefen.

Zur Bearbeitung dieses Buches wird die Kenntnis von mathematischen Grundlagen vorausgesetzt, wie sie üblicherweise in den Grundvorlesungen der ersten beiden Semester vermittelt wird. Wichtig sind insbesondere die gute Beherrschung der Grundbegriffe und Notationen was Mengen und Funktionen angeht, sowie das souveräne Arbeiten mit Integralen, Reihen, linearen Gleichungssystemen und Matrizen.

Jedes Kapitel beginnt mit einem sehr knappen Ausblick auf seinen Inhalt, sowie einer Zusammenstellung der Lernziele des Kapitels und ggf. einem Hinweis auf besonders relevante Vorkenntnisse in diesem Kapitel. Die wichtigsten Konzepte werden an detailliert ausgeführten Rechenbeispielen illustriert. Am Ende jedes Kapitels findet sich eine Zusammenfassung der wichtigsten Ergebnisse, sowie einige passende Übungsaufgaben. Zu manchen (aber nicht allen) finden sich am Ende die Ergebnisse. Die Kapitel bauen insbesondere im ersten Teil aufeinander auf.

Noemi Kurt

# Inhaltsverzeichnis

<b>I</b>	<b>Grundlagen der Wahrscheinlichkeitstheorie</b>	<b>1</b>
<b>1</b>	<b>Endliche Wahrscheinlichkeitsräume</b>	<b>3</b>
1.1	Wahrscheinlichkeiten und Ereignisse	4
1.2	Häufigkeiten und Wahrscheinlichkeiten	13
	Aufgaben	14
<b>2</b>	<b>Bedingte Wahrscheinlichkeiten und Unabhängigkeit</b>	<b>17</b>
2.1	Bedingte Wahrscheinlichkeiten	18
2.2	Mehrstufige Experimente und Formel von der Gesamtwahrscheinlichkeit	20
2.3	Unabhängigkeit von Ereignissen	23
2.4	Bayes'sche Umkehrformel	26
	Aufgaben	29
<b>3</b>	<b>Diskrete Zufallsvariablen und Verteilungen</b>	<b>31</b>
3.1	Zufallsvariablen, Verteilung, Verteilungsfunktion	32
3.2	Bernoulli-Verteilung, Binomialverteilung, geometrische Verteilung	39
3.3	Poisson-Verteilung und Approximation	45
3.4	Zipf-Verteilung und skalenfreie Netzwerke	47
	Aufgaben	49
<b>4</b>	<b>Gemeinsame Verteilung, Unabhängigkeit von Zufallsvariablen</b>	<b>51</b>
4.1	Gemeinsame Verteilung, bedingte Verteilung	52
4.2	Unabhängigkeit von Zufallsvariablen	55
	Aufgaben	59
<b>5</b>	<b>Kenngrößen für Zufallsvariablen</b>	<b>61</b>
5.1	Erwartungswert	62
5.2	Varianz	67
5.3	Markov- und Chebyshev-Ungleichung	70
5.4	Kovarianz und Korrelation	71
	Aufgaben	76
<b>6</b>	<b>Zufallsvariablen mit Dichte</b>	<b>79</b>
6.1	Zufallsvariablen, Verteilungsfunktion, Dichten	80
6.2	Exponentialverteilung	85
6.3	Normalverteilung	87
6.4	Mehrdimensionale Verteilungen	90
	Aufgaben	94
<b>7</b>	<b>Grenzwertsätze</b>	<b>97</b>
7.1	Gesetz der großen Zahlen	97
7.2	Zentraler Grenzwertsatz und Anwendungen	99
	Aufgaben	103
<b>II</b>	<b>Einführung in die Statistik</b>	<b>105</b>

<b>8</b>	<b>Parameterschätzung</b>	107
8.1	Einige Grundbegriffe	108
8.2	Beispiele und Eigenschaften von Schätzern	111
8.3	Maximum Likelihood-Schätzung	114
8.4	Empirische Korrelation und Regressionsgerade	118
	Aufgaben	122
<b>9</b>	<b>Konfidenzbereiche</b>	123
9.1	Grundprinzipien und Beispiele	124
9.2	Einschub: $\chi^2$ -Verteilung und $t$ -Verteilung	128
9.3	Konfidenzintervalle für Erwartungswert und Varianz	130
	Aufgaben	131
<b>10</b>	<b>Hypothesentests</b>	133
10.1	Grundprinzipien, $p$ -Wert, Fehler erster und zweiter Art	134
10.2	$t$ -Test	138
10.3	$\chi^2$ -Test	143
	Aufgaben	148
<b>III</b>	<b>Markov-Ketten und stochastische Algorithmen</b>	151
<b>11</b>	<b>Markov-Ketten</b>	153
11.1	Markov-Ketten und stochastische Matrizen	154
11.2	Invariante Verteilungen	160
11.3	Langzeitverhalten von Markov-Ketten	164
11.4	Anwendungen	166
	Aufgaben	169
<b>12</b>	<b>Randomisierte Algorithmen: Beispiele und Anwendungen</b>	171
12.1	Erwartete Laufzeit von Quicksort	172
12.2	Erfolgswahrscheinlichkeit des MinCut	174
12.3	Ein probabilistischer Existenzbeweis: Graphenfärbung	177
	Aufgaben	179
<b>13</b>	<b>Verzweigungsprozesse und erzeugende Funktionen</b>	181
13.1	Galton-Watson Verzweigungsprozesse	182
13.2	Wahrscheinlichkeitserzeugende Funktionen	184
13.3	Aussterbewahrscheinlichkeit im Galton-Watson Prozess	185
13.4	Weitere Anwendungen der wahrscheinlichkeitserzeugenden Funktion	188
	Aufgaben	189

<b>14 Warteschlangenmodelle und Markov-Ketten in stetiger Zeit</b> .....	191
14.1 Warteschlangenmodelle in stetiger Zeit .....	192
14.2 Eigenschaften der Exponentialverteilung .....	194
14.3 Markov-Ketten in stetiger Zeit .....	197
Aufgaben .....	200
<b>15 Simulation von Zufallsvariablen, Monte-Carlo Methode</b> ..	203
15.1 Simulation von Zufallszahlen, Pseudozufallszahlen .....	203
15.2 Monte-Carlo Simulation .....	206
Aufgaben .....	210
<b>16 Markov-Ketten Monte Carlo und Konvergenzgeschwindigkeit</b> .....	211
16.1 Grundideen .....	212
16.2 Metropolis-Algorithmus .....	212
16.3 Konvergenzgeschwindigkeit und Mischzeiten .....	215
Aufgaben .....	217
<b>A Tabellen</b> .....	219
<b>B Lösungen der Aufgaben</b> .....	223
<b>Literaturverzeichnis</b> .....	229



# Teil I

## Grundlagen der Wahrscheinlichkeitstheorie





# 1 Endliche Wahrscheinlichkeitsräume

---

## Übersicht

1.1	Wahrscheinlichkeiten und Ereignisse .....	4
1.2	Häufigkeiten und Wahrscheinlichkeiten .....	13

---

In diesem Kapitel werden Wahrscheinlichkeitsräume eingeführt. Dies ist der mathematische Rahmen, in dem üblicherweise wahrscheinlichkeitstheoretische Fragestellungen formuliert werden. Es werden Grundbegriffe und Notationen eingeführt, um die Wirklichkeit in einer Art und Weise zu formalisieren und zu modellieren, welche die präzise mathematische Behandlung möglich macht.

### Lernziele

- Die Grundbegriffe und Standardnotation für Wahrscheinlichkeiten und Ereignisse kennen
- Geeignete Wahrscheinlichkeitsräume für einfache Zufallsexperimente angeben können
- Wahrscheinlichkeiten für Ereignisse in endlichen Wahrscheinlichkeitsräumen berechnen können.

### Vorkenntnisse

Es werden Grundbegriffe und Notation von Mengen und Funktionen benötigt, wie sie beispielsweise in einführenden Vorlesungen der Analysis behandelt werden.

## 1.1 Wahrscheinlichkeiten und Ereignisse

Die Wahrscheinlichkeitstheorie beschäftigt sich mit den mathematischen Gesetzmäßigkeiten zufälliger Abläufe, sogenannter Zufallsexperimente.

### Definition 1.1 (Zufallsexperiment)

Ein **Zufallsexperiment** ist ein reproduzierbarer Vorgang mit zufälligem oder (a priori) unbestimmtem Ausgang. ♦

Mit dieser Definition, die zu einem gewissen Grad an alltägliche Vorstellungen appelliert, schließt man insbesondere Vorgänge mit ein, die vielleicht nicht tatsächlich zufällig ablaufen, aber für praktische Zwecke nicht exakt genug beschrieben werden können. Wirft man beispielsweise einen Würfel, so könnte man das Ergebnis theoretisch berechnen, wenn man die Ausgangslage des Würfels sowie sämtliche wirkenden Kräfte genau kennen würde. Da dies nicht möglich ist, bezeichnen wir das Ergebnis eines Würfelvorgangs als zufällig. Einige Beispiele für Zufallsexperimente:

- Werfen einer Münze
- Lottoziehung
- Auswahl einer Stichprobe von  $k$  Bauteilen für einen Funktionstest, aus einer Gesamtzahl von  $n$  gleichartigen Teilen aus einer Produktion
- Auswahl von Personen für eine repräsentative Umfrage
- Zufällige Auswahl eines Ortes und eines Zeitpunkts zur Messung der aktuellen Luftschadstoffbelastung

Auf den ersten Blick mögen Experimente wie Münzwurf uninteressant wirken, jedoch kann man kompliziertere zufällige Vorgänge oft auf solche einfachen Teilabläufe reduzieren, für die man dann interessante Größen einfach berechnen kann. Münzwürfe dienen z.B. zur Modellierung von zufälligen Entscheidungen. Abfolgen von zufälligen Entscheidungen können zu geeigneten komplexen Algorithmen, z.B. zufälligen Sortieralgorithmen, „zusammengebaut“ werden.

### Definition 1.2 (Ergebnis, Ereignis)

Ein möglicher Ausgang eines Zufallsexperiments ist ein **Ergebnis**. Die Menge aller möglichen Ergebnisse eines bestimmten Zufallsexperiment wird üblicherweise mit dem griechischen Buchstaben  $\Omega$  bezeichnet. Für ein einzelnes Ergebnis schreiben wir entsprechend  $\omega \in \Omega$ . Ein **Ereignis** ist eine Teilmenge von  $\Omega$ , in Mengenschreibweise  $A \subseteq \Omega$ . ♦

Ein Ereignis besteht also aus einem oder mehreren Ergebnissen. Die leere Menge  $\emptyset$  wird ebenfalls als Ereignis aufgefasst, obwohl sie keine Elemente enthält. Formal gesehen ist  $\emptyset$  eine Teilmenge jeder Menge. Will man mathematisch ganz sauber arbeiten, darf man eigentlich nicht jede Teilmenge einer Ergebnismenge  $\Omega$  als Ereignis zulassen, sondern nur gewisse geeignete, sogenannte „messbare“ Mengen. Für die meisten praktischen Zwecke ist dies jedoch nicht von Bedeutung, und es soll hier nicht weiter auf diese Problematik eingegangen werden.

**Beispiel 1.1 (Ereignisse beim Würfeln)**

Einfaches Würfeln ist ein Zufallsexperiment mit sechs möglichen Ergebnissen:

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Beispiele für Ereignisse sind

$$A = \{2, 4, 6\} = \{\omega \in \Omega : \omega \text{ ist eine gerade Zahl}\}$$

oder

$$B = \{3, 4, 5, 6\} = \{\omega \in \Omega : \omega \geq 3\}.$$

In Worten ausgedrückt beschreibt  $A$  das Ereignis, dass eine gewürfelte (und damit zufällige) Zahl gerade ist,  $B$  das Ereignis, dass eine gewürfelte Zahl mindestens den Wert 3 hat. Mit Hilfe von Mengenoperationen können neue Ereignisse beschrieben werden: Der Durchschnitt von  $A$  und  $B$  ist

$$A \cap B = \{4, 6\}.$$

Dies bedeutet, dass  $A$  und  $B$  beide eintreten, dass also eine gewürfelte Zahl sowohl gerade als auch größer oder gleich 3 ist. Die Vereinigung

$$A \cup B = \{2, 3, 4, 5, 6\}$$

bedeutet, dass  $A$  oder  $B$  (oder beide) eintreten, d.h. dass die Zahl gerade oder mindestens 3 ist. Das Ereignis

$$A^c := \Omega \setminus A = \{1, 3, 5\}$$

ist das Komplement von Ereignis  $A$ , also das Ereignis, dass  $A$  nicht eintritt, bzw. dass die Zahl ungerade ist.  $A^c$  bezeichnet man auch als das *Gegenereignis* von  $A$ . Manchmal wird gleichbedeutend mit der Notation  $A^c$  auch  $\bar{A}$  geschrieben. Die Schreibweise  $\Omega \setminus A$  bezeichnet die Differenz der Ereignisse, und bedeutet in Worten „ $\Omega$  ohne  $A$ “. Eine weitere Mengenoperation ist die symmetrische Differenz  $A \Delta B := (A \cup B) \setminus (A \cap B)$ . ■

In obigem Beispiel kann man erkennen, dass Ereignisse, also Teilmengen von  $\Omega$ , in Beziehung gesetzt werden können mit geeigneten Aussagen, welche diese Ereignisse beschreiben. Dabei werden die Verknüpfung von Ereignissen in Mengenoperationen übersetzt, und umgekehrt. Die Tabelle 1.1 stellt diesen Zusammenhang dar. Wählt man dabei für die Aussagen auf der rechten Seite  $a = \omega \in A$  und  $b = \omega \in B$ , mit den Mengen  $A$  und  $B$ , so entsprechen sich linke und rechte Seite der Tabelle exakt.

**Beispiel 1.2 (Zufällige Auswahl eines Punktes aus einem Quadrat)**

Einem Quadrat ist eine kompliziertere Figur einbeschrieben, siehe Abbildung 1.1. Wir wählen daraus, beispielsweise durch Werfen eines Pfeils, zufällig einen Punkt aus. Diese zufällige Auswahl ist ein Zufallsexperiment. Wir können das Quadrat selbst, bestehend aus einer Menge von einzelnen Punkten, als Ergebnismenge  $\Omega$  auffassen. Ein Ereignis ist dann z.B.  $A = \{\omega \in \Omega : \omega \text{ liegt in der grauen Fläche}\}$ . ■

Tab. 1.1: Korrespondenz von Mengenoperationen und logischen Verknüpfungen

Mengenoperationen zwischen Mengen $A, B$	logische Verknüpfungen zwischen Aussagen $a, b$
$A \cap B$ Durchschnitt	$a \wedge b$ und
$A \cup B$ Vereinigung	$a \vee b$ oder (einschließend)
$A^c$ Komplement	$\neg a$ Negation
$A \setminus B$ Differenz	$a \wedge (\neg b) = \neg(a \Rightarrow b)$
$A \Delta B$ symmetrische Differenz	$a \oplus b$ oder $a \veebar b$ ausschließendes oder, xor

Die beiden Beispiele 1.1 und 1.2 haben einen großen qualitativen Unterschied: Beim Würfeln gibt es nur endlich viele mögliche Ergebnisse, nämlich genau 6. Im zweiten Beispiel mit dem Quadrat gibt es unendlich (sogar überabzählbar) viele, denn ein Quadrat im  $\mathbb{R}^2$  besteht aus unendlich vielen einzelnen Punkten. Die Wahrscheinlichkeitstheorie ist in beiden Fällen anwendbar, und die Grundbegriffe sind für beide Fälle definiert. Viele Konzepte werden wir jedoch zuerst einmal an Hand von Beispielen mit endlich vielen möglichen Ergebnissen kennen lernen.

Als nächstes wollen wir Ereignissen eine Wahrscheinlichkeit zuordnen. Dies soll mathematisch formalisiert geschehen, jedoch in einer Art und Weise, die unserer Intuition entspricht. Es hat sich eingebürgert, Wahrscheinlichkeiten mit Zahlen zwischen 0 und 1 anzugeben, wobei 1 bedeutet, dass ein entsprechendes Ereignis „sicher eintritt“, und 0 heißt dass das entsprechende Ereignis „sicher nicht“ eintritt. Im Weiteren soll gelten, dass die Wahrscheinlichkeit für ein Ereignis, welches auf verschiedene, unterschiedliche Arten eintreten kann, sich als Summe der Wahrscheinlichkeiten für die unterschiedlichen Möglichkeiten berechnen lässt. Diese Eigenschaften sind in den sogenannten *Axiomen von Kolmogorov* zusammengefasst, welche das Fundament der Wahrscheinlichkeitstheorie bilden. Sie sind in der folgenden Definition für endliche Wahrscheinlichkeitsräume formuliert (siehe Definition 3.1 für eine Verallgemeinerung).

### Definition 1.3 (Endlicher Wahrscheinlichkeitsraum)

Ein **endlicher Wahrscheinlichkeitsraum** ist ein Paar bestehend aus einer endlichen **Ergebnismenge**  $\Omega$ , und einem **Wahrscheinlichkeitsmaß**  $\mathbb{P}$ , welches jedem Ereignis

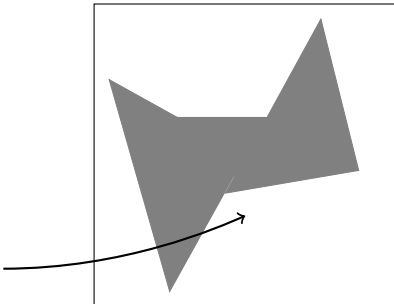


Abb. 1.1: Quadrat mit einbeschriebener Figur

$A \subseteq \Omega$  eine Zahl  $\mathbb{P}(A)$  zuordnet. Diese Zahl ist die **Wahrscheinlichkeit** des Ereignisses  $A$ . Dabei hat  $\mathbb{P}$  die Eigenschaften

**(P0)**  $0 \leq \mathbb{P}(A) \leq 1$  für alle  $A \subseteq \Omega$

**(P1)**  $\mathbb{P}(\Omega) = 1$

**(P2)** Für  $A, B \subseteq \Omega$  mit  $A \cap B = \emptyset$  gilt  $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ .



Ein Wahrscheinlichkeitsmaß ist also eine Funktion auf der Menge aller Teilmengen von  $\Omega$ , welche Werte in  $[0, 1]$  annimmt. Die Ergebnismenge  $\Omega$  hat die Wahrscheinlichkeit 1. Das kann so verstanden werden, dass sie alle möglichen Ergebnisse enthält, also sicher eintritt. Wenn zwei Ereignisse  $A$  und  $B$  *disjunkt* sind ( $A \cap B = \emptyset$  gilt), also nicht gleichzeitig eintreten können, so berechnet sich die Wahrscheinlichkeit für das Eintreten von  $A$  oder  $B$  durch Addition der Einzelwahrscheinlichkeiten.

Es ist wichtig, zwischen Ergebnissen und Ereignissen sauber zu unterscheiden: Ergebnisse sind Elemente der Ergebnismenge  $\Omega$ , Ereignisse sind Teilmengen von  $\Omega$ . Jedoch kann man Ergebnisse  $\omega \in \Omega$  auf einfache Weise als Ereignisse betrachtet, indem man sie als einelementige Menge  $\{\omega\}$  auffasst. Somit ist auch die Wahrscheinlichkeit  $\mathbb{P}(\{\omega\})$  für eine solche einelementige Menge definiert. Der Einfachheit halber lässt man die geschweiften Klammern oft weg, und schreibt  $\mathbb{P}(\omega)$ . Wir sprechen dann auch von der „Wahrscheinlichkeit für das Ergebnis  $\omega$ “, wobei wir formal gesehen die „Wahrscheinlichkeit für das Ereignis  $\{\omega\}$ “ meinen.

### Beispiel 1.3 (Fortsetzung von Beispiel 1.1)

Beim Würfeln mit einem fairen Würfel sind alle Elemente von  $\Omega = \{1, \dots, 6\}$  gleich wahrscheinlich. Nach (P1) und (P2) gilt, da die Ereignisse  $\{1\}, \dots, \{6\}$  paarweise disjunkt sind,

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(\{1, \dots, 6\}) = \mathbb{P}(\{1\} \cup \dots \cup \{6\}) = \mathbb{P}(1) + \dots + \mathbb{P}(6).$$

Dabei wurde die Konvention für Ergebnisse, also  $\mathbb{P}(1) = \mathbb{P}(\{1\})$  etc., verwendet. Da alle Ergebnisse gleich wahrscheinlich sind, muss also

$$\mathbb{P}(1) = \dots = \mathbb{P}(6) = \frac{1}{6}$$

gelten, was sich mit unserer Intuition deckt. Wir können nun Wahrscheinlichkeiten von Ereignissen berechnen, z.B. von  $A = \{\omega \in \Omega : \omega \text{ ist gerade}\}$  :

$$\mathbb{P}(A) = \mathbb{P}(\{2, 4, 6\}) = \mathbb{P}(2) + \mathbb{P}(4) + \mathbb{P}(6) = 3 \cdot \frac{1}{6} = \frac{1}{2}.$$

Ebenso für  $B = \{\omega \in \Omega : \omega \geq 3\}$  und  $A \cup B = \{2, \dots, 6\}$  :

$$\mathbb{P}(B) = \frac{4}{6} = \frac{2}{3}, \quad \text{und} \quad \mathbb{P}(A \cup B) = \frac{5}{6}.$$

Beachte:  $\mathbb{P}(A \cup B) \neq \mathbb{P}(A) + \mathbb{P}(B) = \frac{7}{6}$ . Dies ist kein Widerspruch zu (P2) in Definition 1.3, da  $A \cap B \neq \emptyset$ . Es gilt  $\mathbb{P}(A \cap B) = \frac{2}{6} = \frac{1}{3}$ . Daraus sehen wir

$$\frac{5}{6} = \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = \frac{3}{6} + \frac{4}{6} - \frac{2}{6}.$$

■

Aus der Definition 1.3 können weitere Rechenregeln hergeleitet werden.

**Satz 1.1 (Rechenregeln für Wahrscheinlichkeiten)**

*Sei  $(\Omega, \mathbb{P})$  ein Wahrscheinlichkeitsraum. Für  $A, B \subseteq \Omega$  gelten*

**(P3)**  $\mathbb{P}(A^c) = 1 - \mathbb{P}(A),$

**(P4)**  $\mathbb{P}(\emptyset) = 0.$

**(P5)** Für  $A \subseteq B$  gilt  $\mathbb{P}(A) \leq \mathbb{P}(B).$

**(P6)**  $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$

**Beweis 1.1** Für alle  $A \subseteq \Omega$  gilt  $A \cup A^c = \Omega$ , und  $A \cap A^c = \emptyset$ . Somit folgt  $\mathbb{P}(A^c) + \mathbb{P}(A) = \mathbb{P}(\Omega) = 1$ , und durch Umstellen erhalten wir (P3). Daraus folgt wegen  $\Omega^c = \emptyset$  sofort (P4). Gilt  $A \subseteq B$  so können wir  $B = A \cup (B \setminus A)$  schreiben, wobei  $A \cap (B \setminus A) = \emptyset$  ist. Somit ist nach (P2)  $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$ , und da  $\mathbb{P}(B \setminus A) \geq 0$  ist nach (P0), folgt (P5). Für die letzte Behauptung benutzen wir  $\mathbb{P}(A \cup B) = \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B)$ , und beachten  $\mathbb{P}(A) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B)$  sowie  $\mathbb{P}(B) = \mathbb{P}(B \setminus A) + \mathbb{P}(B \cap A)$ . Somit folgt auch (P6).  $\square$

In Anwendungen besteht ein Grundproblem darin, die Wirklichkeit in geeigneter Weise mathematisch zu formalisieren, beispielsweise durch Angabe eines Wahrscheinlichkeitsraumes. Dafür gibt es sehr oft mehrere Möglichkeiten, welche je nach Situation günstig oder weniger günstig sind. Dies sieht man bereits an folgendem einfachen Beispiel.

**Beispiel 1.4 (Zweifacher Münzwurf)**

Eine faire Münze wird zweimal geworfen. Wir interessieren uns dafür, wie oft „Zahl“ kommt. Was für ein Wahrscheinlichkeitsraum ist zur Untersuchung dieser Frage geeignet? Dazu müssen wir die möglichen Ergebnisse des Zufallsexperiments „zwei Mal Münze werfen“ formalisieren. Eine Möglichkeit besteht darin, ein Ergebnis als Paar  $\omega = (\omega_1, \omega_2)$  zu schreiben, wobei  $\omega_1$  das Ergebnis des ersten Wurfs bezeichnet, also z. B.  $\omega_1 = K$  falls der erste Wurf Kopf zeigt, und  $\omega_1 = Z$  falls der erste Wurf Zahl zeigt, und analog  $\omega_2 = K$  oder  $\omega_2 = Z$  falls der zweite Wurf Kopf bzw. Zahl zeigt. Der Ergebnisraum  $\Omega$  ist dann die Menge aller möglichen Ergebnisse, also

$$\Omega = \{(K, K), (K, Z), (Z, K), (Z, Z)\}.$$

Dabei bedeutet eben  $(K, K)$ , dass der erste Wurf Kopf und der zweite Wurf ebenfalls Kopf zeigt,  $(K, Z)$  dass erst Kopf und dann Zahl geworfen wurde, usw.

Das ist aber nicht die einzige Möglichkeit, die wir haben. Wenn wir uns nur dafür interessieren, wie oft Zahl geworfen wird, könnten wir auch sagen, dass wir den Ergebnisraum

$$\bar{\Omega} = \{0, 1, 2\}$$

betrachten, wobei 0 bedeutet, dass nie Zahl kommt (also zweimal Kopf), 1 dass einmal Kopf (und einmal Zahl) geworfen wurde, und 2 dass zweimal Kopf geworfen wurde.

Wir haben also mindestens zwei Möglichkeiten, einen Ergebnisraum für dieses Experiment zu wählen. Um einen Wahrscheinlichkeitsraum zu erhalten, müssen wir in beiden Fällen noch das zugehörige Wahrscheinlichkeitsmaß angeben. Um dieses zu bestimmen, ist es hilfreich, das Zufallsexperiment als Baum darzustellen, siehe Abbildung 1.2.

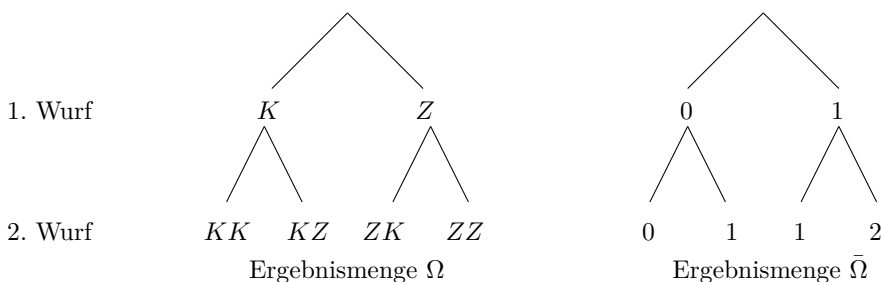


Abb. 1.2: Ergebnisse beim zweifachen Münzwurf.

Da  $K$  und  $Z$  bei beiden Würfeln stets gleich wahrscheinlich sind, müssen im ersten Fall alle vier Ergebnisse gleich wahrscheinlich sein, also

$$\mathbb{P}((K, K)) = \mathbb{P}((K, Z)) = \mathbb{P}((Z, K)) = \mathbb{P}((Z, Z)) = \frac{1}{4}.$$

Den zweiten Fall können wir auf den ersten zurückführen, und erhalten

$$\bar{\mathbb{P}}((0)) = \frac{1}{4}, \quad \bar{\mathbb{P}}((1)) = \frac{1}{2}, \quad \bar{\mathbb{P}}((2)) = \frac{1}{4}.$$

Es ist also wahrscheinlicher, einmal Kopf und einmal Zahl zu bekommen, als beide Male Kopf. ■

Im Beispiel 1.3 waren alle Ergebnisse gleich wahrscheinlich. Wahrscheinlichkeitsräume mit dieser Eigenschaft haben eine eigene Bezeichnung.

#### Definition 1.4 (Laplace-Raum)

Ein **Laplace-Raum** ist ein endlicher Wahrscheinlichkeitsraum, in dem jedes Ergebnis  $\omega \in \Omega$  dieselbe Wahrscheinlichkeit hat. ♦

Im Beispiel 1.4 liegt im ersten beschriebenen Fall ein Laplace-Raum vor, im zweiten aber nicht. Die Fragestellung „wie oft kommt Zahl“ kann aber mit beiden Modellierungen untersucht werden. In dieser Definition muss wiederum zwischen Ergebnissen und Ereignissen unterschieden werden – auch in einem Laplace-Raum hat nicht jedes Ereignis dieselbe Wahrscheinlichkeit, aber jedes Ergebnis. Die Berechnung ist jedoch mit einer einfachen Formel möglich.

**Satz 1.2 (Wahrscheinlichkeiten im Laplace-Raum)**

In einem Laplace-Raum ist die Wahrscheinlichkeit für ein Ereignis  $A \subseteq \Omega$  gegeben durch

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

Dabei bezeichnet  $|A|$  die Kardinalität, also die Anzahl Elemente einer Menge  $A$ .

**Beweis 1.2** Aus der Definition 1.4 folgt, dass  $\mathbb{P}(\omega) = \frac{1}{|\Omega|}$  für alle  $\omega \in \Omega$  gelten muss. Daraus folgt die Aussage des Satzes durch Anwendung von (P2) aus Definition 1.3.  $\square$

**Beispiel 1.5**

Im Beispiel 1.3 hatten wir gesehen, dass alle Ergebnisse beim Würfeln gleich wahrscheinlich sind. Es handelt sich somit um einen Laplace-Raum. Insbesondere hatten wir auch ausgerechnet, dass  $\mathbb{P}(A) = \frac{3}{6} = \frac{|A|}{|\Omega|}$  gilt, und analog für die Ereignisse  $B, A \cap B$ , usw.  $\blacksquare$

**Beispiel 1.6 (Mehrfaches Würfeln)**

Wir würfeln zweimal mit einem fairen Würfel. Als Ergebnisraum wählen wir

$$\Omega_2 = \{(1, 1), (1, 2), (1, 3), \dots, (1, 6), (2, 1), (2, 2), (2, 3), \dots, (6, 5), (6, 6)\},$$

oder, in anderer Schreibweise

$$\Omega_2 = \{\omega = (\omega_1, \omega_2) : \omega_1 \in \{1, \dots, 6\}, \omega_2 \in \{1, \dots, 6\}\}.$$

Das heißt, ein Ergebnis wird als *geordnetes Paar*  $(\omega_1, \omega_2)$  dargestellt, wobei der erste Eintrag das Ergebnis des ersten Wurfs, und der zweite Eintrag das Ergebnis des zweiten Wurfs ist.  $\Omega_2$  hat somit  $6^2 = 36$  Elemente, und da die Ergebnisse alle gleich wahrscheinlich sind, handelt es sich um einen Laplace-Raum, also  $\mathbb{P}(\omega) = \frac{1}{36}$  für jedes  $\omega$  der Form  $(\omega_1, \omega_2) \in \Omega_2$ .

Allgemeiner gilt:  $n$ -faches Würfeln kann durch den Ergebnisraum

$$\Omega_n = \{\omega = (\omega_1, \omega_2, \dots, \omega_n) : \omega_i \in \{1, \dots, 6\}, i = 1, \dots, n\}$$

beschrieben werden. Auch das ist ein Laplace-Raum, und  $|\Omega_n| = 6^n$ .  $\blacksquare$



**Beispiel 1.7 (Wiederholtes Zufallsexperiment)**

Der zweifache Münzwurf und das  $n$ -fache Würfeln sind Beispiele für *wiederholte* Zufallsexperimente. Dieses Prinzip kann man allgemein formulieren. Betrachte ein Zufallsexperiment mit  $k$  möglichen Ergebnissen, und wiederhole dieses  $n$  mal unter gleichbleibenden Bedingungen (wobei die Ergebnisse in der Reihenfolge ihres Auftretens notiert werden). Wir bezeichnen mit  $\Omega_n$  die Menge aller Ergebnisse des  $n$ -fach wiederholten Experiments. Es kann stets

$$\Omega_n = \{\omega = (\omega_1, \omega_2, \dots, \omega_n) : \omega_i \in \{1, \dots, k\}, i = 1, \dots, n\}$$

als Ergebnismenge gewählt werden, mit  $|\Omega_n| = k^n$ . Wenn alle Ergebnisse des einmalig ausgeführten Zufallsexperiments gleich wahrscheinlich sind, das heißt, wenn  $\Omega_1$  ein Laplace-Raum ist, dann ist auch  $\Omega_n$  ein Laplace-Raum. ■

Die Wahl eines geeigneten Wahrscheinlichkeitsraumes ist oft ein wichtiger Teil der mathematischen Beschreibung eines Problems aus der Wirklichkeit. In komplexeren Situationen, beispielsweise in der Modellierung eines Telekommunikationsnetzwerkes, ist es oft nicht möglich oder sinnvoll, einen Wahrscheinlichkeitsraum konkret anzugeben. Dort behilft man sich mit anderen Methoden, zum Beispiel:

- Statistisch über *Häufigkeiten* als Schätzer für Wahrscheinlichkeiten. Dies wird kurz im Abschnit 1.2, sowie im Statistik-Teil diskutiert.
- Durch Definition von geeigneten *Zufallsvariablen* und sinnvollen Annahmen über deren *Verteilung*. Dieser Ansatz wird weitgehend auch in diesem Buch verfolgt, beginnend mit Kapitel 3.
- Durch Modellierung mittels geeigneter einfacher Telexperimente, beispielsweise durch Verknüpfung von Urnenmodellen. Dies findet unter Anderem Anwendung bei stochastischen Algorithmen.

**Beispiel 1.8 (Urnenmodelle)**

Urnenmodelle sind ein wichtiges Hilfsmittel in der Wahrscheinlichkeitstheorie und der stochastischen Modellierung. Ihr Nutzen liegt darin, dass sie als (Teil-)Modell in komplexeren Situationen zum Einsatz kommen. In einer Klasse von einfachen Urnenmodellen betrachtet man eine Urne, welche  $k$  unterscheidbare Kugeln enthält. Die Kugeln seien von 1 bis  $k$  durchnummeriert. Dabei stellen wir uns vor, dass die Kugeln in der Urne gut gemischt werden können, und dass wir ohne hinzusehen eine Kugel aus der Urne ziehen. Das heißt, dass das Experiment „ziehe eine Kugel“ ein Zufallsexperiment ist, und dass jede der insgesamt  $k$  Kugeln dieselbe Wahrscheinlichkeit hat, gezogen zu werden. Es wird nun  $n$  Mal eine Kugel gezogen, je nach Situation mit oder ohne Zurücklegen, und die gezogene Zahl wird notiert. Dabei ist manchmal die Reihenfolge der gezogenen Zahlen von Interesse, und manchmal nicht. Bei solchen Zufallsexperimenten können leicht die entsprechenden Wahrscheinlichkeiten berechnet werden. In praktischen Anwendungen stehen die Urnen und Kugeln oft symbolisch für etwas anderes, z.B. für die Auswahl eines Teilalgorithmus aus mehreren Möglichkeiten, und werden miteinander zu komplizierteren Situationen verknüpft.

1. *Ziehen mit Zurücklegen, mit Beachten der Reihenfolge.* Beachtung der Reihenfolge bedeutet, dass z.B. das Ergebnis  $(1, 2, 2)$  einer dreifachen Ziehung ein anderes ist als  $(2, 1, 2)$ , obwohl dieselben Zahlen gezogen wurden. Dies ist ein einfaches wiederholtes Zufallsexperiment, bei dem, da zurückgelegt wird, das Ergebnis bei jeder Ziehung die gleiche Wahrscheinlichkeit hat. Somit ist ein geeigneter Ergebnisraum  $\Omega = \{(\omega_1, \dots, \omega_n) : \omega_i \in \{1, \dots, k\}, i = 1, \dots, n\}$ . Es handelt sich um einen Laplace-Raum, somit gilt  $\mathbb{P}((\omega_1, \dots, \omega_n)) = \frac{1}{k^n}$ , da man für jede der  $n$  Ziehungen  $k$  Möglichkeiten hat.

2. *Ziehen ohne Zurücklegen, mit Beachten der Reihenfolge.* Zieht man ohne Zurücklegen, so kann jede Zahl nur einmal gezogen werden. Als Ergebnisraum wählen wir deshalb

$$\Omega = \{(\omega_1, \dots, \omega_n) : \omega_i \in \{1, \dots, k\}, i = 1, \dots, n, \omega_i \neq \omega_j \text{ für } i \neq j\}.$$

Ebenso verändern sich die Wahrscheinlichkeiten bei jeder Ziehung, da nun nicht mehr alle Kugeln vorhanden sind. Beim ersten Ziehen gibt es  $k$  Möglichkeiten, beim zweiten  $k - 1$ , etc. Für  $n \leq k$  Ziehungen gibt es also

$$k \cdot (k - 1) \cdot \dots \cdot (k - (n - 1)) = \frac{k!}{(k - n)!}$$

Möglichkeiten. Alle diese Möglichkeiten sind wiederum gleich wahrscheinlich, so dass wir für  $\omega \in \Omega$  erhalten

$$\mathbb{P}(\omega) = \frac{(k - n)!}{k!}.$$

3. *Ziehen ohne Zurücklegen, ohne Beachten der Reihenfolge.* Bei der Ziehung der Fernsehlotterie werden Kugeln in beliebiger Reihenfolge gezogen, die Ergebnisse jedoch stets in aufsteigender Reihenfolge angegeben. Somit entsprechen beispielsweise die Ziehungen  $\{40, 18, 1, 17, 22, 34\}$  und  $\{40, 34, 22, 17, 1, 18\}$  derselben Reihe von Gewinnzahlen, nämlich  $\{1, 17, 18, 22, 34, 40\}$  (was eine weitere mögliche Ziehung ist). Die Reihenfolge der gezogenen Zahlen spielt also *keine* Rolle, da hinterher geordnet wird. Da es  $n!$  Möglichkeiten gibt,  $n$  Zahlen anzuordnen, entsprechen in diesem Fall von  $\frac{k!}{(k - n)!}$  Möglichkeiten der Ziehung ohne Zurücklegen mit Beachtung der Reihenfolge jeweils  $n!$  demselben Ergebnis. Somit gibt es für das wiederholte Ziehen ohne Zurücklegen ohne Beachtung der Reihenfolge

$$\frac{k!}{(k - n)!n!} = \binom{k}{n}$$

Möglichkeiten. Dabei bezeichnet  $\binom{k}{n}$  der aus der Analysis bekannte Binomialkoeffizient. Die Ergebnismenge

$$\Omega = \{(\omega_1, \dots, \omega_n) : \omega_1 < \dots < \omega_n, \omega_i \in \{1, \dots, k\}, i = 1, \dots, n\}$$

wird damit zu einem Laplace-Raum. ■

## 1.2 Häufigkeiten und Wahrscheinlichkeiten

Bis hierher wurden Wahrscheinlichkeiten als rein mathematisches Konzept betrachtet. In der Wirklichkeit können wir Wahrscheinlichkeiten im Sinne eines Wahrscheinlichkeitsmaßes nicht direkt beobachten, wir können sie nur mittels theoretischer Überlegungen berechnen. Werfen wir beispielsweise einen (nicht notwendigerweise fairen) Würfel, so sehen wir ein ganz bestimmtes Ergebnis, eine *Realisierung*, aber wir bekommen durch diese eine Beobachtung keine Information über zugrundeliegende Wahrscheinlichkeiten. Dafür müssen wir die Form des Würfels kennen, und Annahmen über seine Symmetrie treffen, z.B. annehmen, dass er fair ist.

Wir können aber, wenn wir möchten, den Würfel nicht einmal, sondern sehr oft werfen, und uns die jeweiligen Ergebnisse notieren, also ein wiederholtes Zufallsexperiment ausführen. Statt einer Liste der verschiedenen Ergebnisse können wir die *Häufigkeiten* der jeweiligen Ergebnisse betrachten. Wird der Würfel oft genug geworfen, so geben diese – empirisch ermittelten – Häufigkeiten Aufschluss über die Wahrscheinlichkeiten.

Konkret stellen wir uns vor, dass wir den Würfel  $n = 100$  mal werfen. Wir bezeichnen mit  $H_i$  die Häufigkeit, mit der dabei das Ergebnis  $i$  auftritt,  $i = 1, \dots, 6$ . Die Zahl  $H_1$  gibt also an, wie viele unter den 100 Würfeln eine 1 ergeben haben,  $H_2$  ist die Anzahl der Zweien, usw. Sei  $p_i$  die (theoretische, unbekannte, nicht direkt messbare) Wahrscheinlichkeit, dass der Würfel den Wert  $i$  zeigt. Wir erwarten dann, dass

$$p_i \approx \frac{H_i}{n}$$

gilt, d.h. dass die *relative Häufigkeit*  $H_i/n$  eine *Näherung* der unbekannten Wahrscheinlichkeit  $p_i$  ist. Ebenso erwarten wir, dass diese Näherung umso besser wird, je größer  $n$  ist. Umgangssprachlich oder für praktische Zwecke sind viele „Wahrscheinlichkeiten“ in diesem Sinne zu verstehen. Wird beispielsweise gesagt, dass es eine Wahrscheinlichkeit  $p$  gibt, an einer bestimmten Krankheit zu erkranken, so bedeutet das normalerweise, dass in einer (hinreichend großen) Menge von Personen die Krankheit mit einer relativen Häufigkeit  $p$  beobachtet wurde. Dies ist eine „statistische“ Interpretation des Begriffs der Wahrscheinlichkeit, im Gegensatz zur axiomatischen, welche normalerweise in der Mathematik gebräuchlich ist. Entsprechend wird der Zusammenhang zwischen *beobachteten* Häufigkeiten und *theoretischen* Wahrscheinlichkeiten im Statistik-Teil dieses Buches besonders wichtig sein.

### Beispiel 1.9 (Zufällige Zielscheibe; Monte-Carlo-Methode)

Wie in Beispiel 1.2 betrachten wir eine quadratische Zielscheibe, auf der eine komplizierte Figur eingezeichnet ist. Wir möchten wissen, mit welcher Wahrscheinlichkeit ein zufällig ausgewählter Punkt aus dem Quadrat im Innern der gezeichneten Figur liegt. Diese Wahrscheinlichkeit könnte durch Berechnung der Fläche der Figur bestimmt werden. Sollte die Intuition über relative Häufigkeiten stimmen, könnte die Wahrscheinlichkeit auch mit Hilfe eines Dartpfeils bestimmt werden, mit folgendem (wiederholten) Experiment: Ohne zu zielen wird sehr oft ein Pfeil auf das Quadrat geworfen, und gezählt, wie

oft die Figur getroffen wird. Das Verhältnis der Treffer im Vergleich zur Gesamtzahl der geworfenen Pfeile sollte dann ungefähr dieser Wahrscheinlichkeit entsprechen.

Auch wenn man in der Realität selten genau dieses Experiment ausführt, so wird die Grundidee, das wiederholte Ausführen eines Zufallsexperiments, in vielen praktischen Situationen zur näherungsweisen Bestimmung von Wahrscheinlichkeiten verwendet. Dies ist das Grundprinzip für sogenannte *Monte-Carlo-Methoden*, bzw. *Monte-Carlo-Simulationen*. ■

### Zusammenfassung

In diesem Kapitel wurden endliche Wahrscheinlichkeitsräume eingeführt. Diese bestehen aus einer Ergebnismenge  $\Omega$  und einem Wahrscheinlichkeitsmaß  $\mathbb{P}$ . Dabei sind die Elemente von  $\Omega$  die möglichen Ergebnisse eines Zufallsexperiments.  $\mathbb{P}$  ordnet jedem Ereignis, also jeder Teilmenge von  $\Omega$ , eine Wahrscheinlichkeit zu, muss somit insbesondere Werte in  $[0, 1]$  annehmen, und erfüllt eine Reihe weiterer natürlicher Eigenschaften. Eine wichtige Klasse von Wahrscheinlichkeitsräumen sind die Laplace-Räume. Dort ist jedes Ergebnis gleich wahrscheinlich, wie z.B. beim Würfeln mit einem fairen Würfel. Empirisch ermittelt werden können Wahrscheinlichkeiten zum Beispiel über relative Häufigkeiten.

## Aufgaben

**1.1** Betrachte eine Urne mit 8 Kugeln, nummeriert von 1 bis 8. Ziehe zufällig eine Kugel. Sei  $\Omega$  die Menge der möglichen Ergebnisse. Definiere Ereignisse  $A = \{\text{Ergebnis ist mindestens } 5\}$ ,  $B = \{\text{Ergebnis ist höchstens } 7\}$ ,  $C = \{\text{Ergebnis ist ungerade}\}$ . Geben Sie die folgenden Ereignisse durch Auflisten ihrer Elemente an: a)  $\Omega$ , b)  $A \cap B$  c)  $B \cup C$  d)  $(A^c \cup B^c) \cap C$

**1.2** Ein technisches Gerät besteht aus  $n$  verschiedenen Komponenten, nummeriert mit  $i = 1, \dots, n$ . Es bezeichne  $D_i$  das Ereignis, dass die  $i$ -te Komponente defekt ist. Stellen Sie die folgenden Ereignisse mittels elementarer Mengenoperationen dar: a) Mindestens eine Komponente defekt, b) genau eine Komponente defekt, c) höchstens eine Komponente defekt, d) höchstens eine Komponente nicht defekt.

**1.3** In einem Experiment wird die Wellenlänge der elektromagnetischen Strahlung einer Strahlungsquelle gemessen. Die Ergebnismenge wird mit  $\Omega = [0, \infty)$  modelliert, wobei eine Zahl  $\omega \in \Omega$  die Wellenlänge in Nanometer (nm) angibt. Definiere Ereignisse

$$A = \{\text{Die Wellenlänge ist (strikt) kleiner als } 400 \text{ nm}\},$$

$$B = \{\text{Die Quelle strahlt sichtbares Licht ab}\}.$$

(Sichtbares Licht liegt im Wellenlängenbereich zwischen 380 nm und 780 nm.) Welche Intervalle beschreiben die folgenden Ereignisse? *a)*  $A$ , *b)*  $B$  *c)*  $A \cap B$  *d)*  $A^c \cup B$

**1.4** Wie in Aufgaben 1.1 betrachte eine Urne mit 8 Kugeln. Wir ziehen drei Kugeln mit Zurücklegen, und notieren jeweils die Zahlen. *a)* Wie viele Elemente hat der Laplace-Raum zu diesem Zufallsexperiment? *b)* Wie viele Ergebnisse  $\omega$  gehören zum Ereignis {es wird genau zwei Mal die Zahl 2 gezogen}? *c)* Mit welcher Wahrscheinlichkeit wird drei Mal dieselbe Zahl gezogen? *d)* Mit welcher Wahrscheinlichkeit werden drei verschiedene Zahlen gezogen?

**1.5** Es sei  $(\Omega, \mathbb{P})$  ein Wahrscheinlichkeitsraum. Seien  $A, B \subseteq \Omega$ . Beweisen Sie  $\mathbb{P}(A \Delta B) = \mathbb{P}(A) + \mathbb{P}(B) - 2\mathbb{P}(A \cap B)$

**1.6** Es sei  $(\Omega, \mathbb{P})$  ein Wahrscheinlichkeitsraum. Seien  $A, B, C \subseteq \Omega$ . Zeigen Sie:  $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)$ .

**1.7** Verallgemeinern Sie die Formel aus Aufgabe 1.6 auf  $n \geq 3$  Ereignisse.

**1.8** Betrachte eine Urne mit zwei roten und einer schwarzen Kugel. Ziehe daraus 4 Mal mit Zurücklegen. *a)* Mit welcher Wahrscheinlichkeit wird genau eine schwarze Kugel gezogen? *b)* Mit welcher Wahrscheinlichkeit wird mindestens eine schwarze Kugel gezogen?

**1.9** Betrachte eine Urne mit vier Kugeln: Rot, blau, schwarz und weiß. Sei  $\Omega_1$  die Ergebnismenge beim Zufallsexperiment „ziehe 3 Kugeln mit Zurücklegen“, und  $\Omega_2$  die Ergebnismenge beim Zufallsexperiment „ziehe 3 Kugeln ohne Zurücklegen“. *a)*  $|\Omega_1| = ?$  *b)*  $|\Omega_2| = ?$  *c)* Berechne im ersten Experiment die Wahrscheinlichkeit, dass die blaue Kugel (mindestens einmal) gezogen wird *d)* Berechne im zweiten Experiment die Wahrscheinlichkeit, dass die blaue Kugel (mindestens einmal) gezogen wird.



# 2 Bedingte Wahrscheinlichkeiten und Unabhängigkeit

---

## Übersicht

2.1	Bedingte Wahrscheinlichkeiten .....	18
2.2	Mehrstufige Experimente und Formel von der Gesamtwahrscheinlichkeit .....	20
2.3	Unabhängigkeit von Ereignissen .....	23
2.4	Bayes'sche Umkehrformel .....	26

---

In diesem Kapitel werden wir etwas komplexere Situationen als in Kapitel 1 betrachten. Insbesondere interessieren wir uns für Wahrscheinlichkeiten bei wiederholt ausgeführten oder mehrstufigen Zufallsexperimenten. Dabei stellt sich die Frage, inwiefern partielle Informationen über das Ergebnis eines Zufallsexperiments die Wahrscheinlichkeiten beeinflussen. Zentrale Begriffe sind dabei *bedingte Wahrscheinlichkeiten* und die *Unabhängigkeit von Ereignissen*. Außerdem wird die sogenannte *Bayes'sche Umkehrformel* hergeleitet und angewandt.

### Lernziele

- Bedingte Wahrscheinlichkeiten kennen und berechnen können
- Mehrstufige Experimente mathematisch modellieren und behandeln können
- Die Bayes'sche Umkehrformel anwenden können
- Die Unabhängigkeit von Ereignissen bestimmen können und Implikationen davon kennen

### Vorkenntnisse

Die Inhalte von Kapitel 1.

## 2.1 Bedingte Wahrscheinlichkeiten

Zur Berechnung von bedingten Wahrscheinlichkeiten, oder Wahrscheinlichkeiten unter Zusatzinformation, betrachten wir zuerst ein elementares Beispiel.

### Beispiel 2.1 (Würfeln unter Zusatzinformation)

Wir würfeln einmal mit einem fairen Würfel. In Kapitel 1 wurde dies mit Hilfe des Ergebnisraums  $\Omega = \{1, \dots, 6\}$  modelliert. Dabei sind die Ergebnisse alle gleich wahrscheinlich, also

$$\mathbb{P}(1) = \mathbb{P}(2) = \dots = \mathbb{P}(6) = \frac{1}{6}.$$

Nun stellen wir uns vor, dass der Würfel verdeckt geworfen wird, dass wir also das Ergebnis nicht sehen können. Jemand darf sich nun das Ergebnis anschauen, und sagt uns, dass die gewürfelte Zahl gerade ist. Wir erhalten also eine Teilinformation über das Ergebnis. Wie verändern sich die Wahrscheinlichkeiten unter dieser Zusatzinformation? Wir schreiben  $\tilde{\mathbb{P}}$  für die Wahrscheinlichkeit unter Berücksichtigung dieser Zusatzinformation. Dann gelten

$$\tilde{\mathbb{P}}(2) = \tilde{\mathbb{P}}(4) = \tilde{\mathbb{P}}(6) = \frac{1}{3},$$

und

$$\tilde{\mathbb{P}}(1) = \tilde{\mathbb{P}}(3) = \tilde{\mathbb{P}}(5) = 0.$$

Denn ist die Zahl gerade kann sie weder 1 noch 3 noch 5 sein. Die übrigen Zahlen 2, 4 und 6 sind möglich. Da wir keine weiteren Informationen haben, sind diese alle gleich wahrscheinlich, und da Wahrscheinlichkeiten zu 1 aufsummieren müssen, müssen die Wahrscheinlichkeiten alle gleich  $1/3$  sein. ■

$\tilde{\mathbb{P}}$  aus dem obigen Beispiel ist eine sogenannte bedingte Wahrscheinlichkeit. Allgemein wird diese folgendermaßen definiert:

### Definition 2.1 (Bedingte Wahrscheinlichkeit)

Sei  $(\Omega, \mathbb{P})$  ein endlicher Wahrscheinlichkeitsraum, und seien  $A, B \subseteq \Omega$  Ereignisse, mit  $\mathbb{P}(A) > 0$ . Die **bedingte Wahrscheinlichkeit für  $B$  gegeben  $A$**  ist definiert als

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

◆

Für jedes  $A \subseteq \Omega$  ist die Abbildung  $\mathbb{P}(\cdot | A) : B \mapsto \mathbb{P}(B | A)$  ein Wahrscheinlichkeitsmaß gemäß Definition 1.3.  $(\Omega, \mathbb{P}(\cdot | A))$  ist also ein Wahrscheinlichkeitsraum. Auch für dieses Wahrscheinlichkeitsmaß verwenden wir die Notation  $\mathbb{P}(\omega | A) := \mathbb{P}(\{\omega\} | A)$ , wenn wir die bedingte Wahrscheinlichkeit für ein einzelnes Ergebnis berechnen.

### Beispiel 2.2 (Würfeln unter Zusatzinformationen)

Betrachte die Situation aus Beispiel 2.1. Wähle  $A = \{2, 4, 6\}$ , also das Ereignis „das Ergebnis ist gerade“. Dann gilt für einen fairen Würfel

$$\mathbb{P}(A) = \frac{1}{2}.$$



Wir erhalten nun die Zusatzinformation, dass das Ereignis  $A$  eingetreten ist. Das heißt, dass wir uns für die bedingte Wahrscheinlichkeit gegeben  $A$  interessieren. Diese berechnen wir aus der Definition 2.1: Es gilt

$$\mathbb{P}(\{2\} \cap A) = \mathbb{P}(\{2\}) = \frac{1}{6},$$

also

$$\mathbb{P}(\{2\} | A) = \frac{\mathbb{P}(\{2\} \cap A)}{\mathbb{P}(A)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

Andererseits gilt  $\mathbb{P}(\{1\} \cap A) = \mathbb{P}(\emptyset) = 0$ , also

$$\mathbb{P}(\{1\} | A) = \frac{\mathbb{P}(\{1\} \cap A)}{\mathbb{P}(A)} = \frac{0}{1/2} = 0.$$

Analog geht man für 3, 4, 5, 6 vor. Die Wahrscheinlichkeit  $\tilde{\mathbb{P}}$  aus dem Einstiegsbeispiel war also genau die bedingte Wahrscheinlichkeit gegeben  $A$ , also  $\mathbb{P}(\cdot | A)$ . Betrachten wir nun ein weiteres Ereignis,  $B := \{\omega \in \Omega : \omega \geq 3\}$ . Dann gilt nach Definition

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\{4, 6\})}{\mathbb{P}(\{3, 4, 5, 6\})} = \frac{1}{2}.$$

Andererseits ist

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(\{4, 6\})}{\mathbb{P}(\{2, 4, 6\})} = \frac{2}{3}.$$

Die Rollen von  $A$  und  $B$  können also nicht vertauscht werden. ■

### Beispiel 2.3 (Ausfallwahrscheinlichkeit, Sterbewahrscheinlichkeit)

In der Lebensversicherung wird mit statistisch ermittelten Sterbetafeln gearbeitet. Ähnlich kann man auch die Ausfallwahrscheinlichkeit eines technischen Gerätes untersuchen. Wir betrachten ein Zufallsexperiment, welches zufällig eine Person (oder ein Gerät) auswählt. Die Ergebnismenge sei  $\Omega = \mathbb{N}_0$ , wobei  $\omega \in \mathbb{N}_0$  die Lebensdauer der Person auf Jahre gerundet angibt. Die zugehörigen Wahrscheinlichkeiten  $\mathbb{P}(\omega)$  können nun nicht auf Grund mathematischer Überlegungen bestimmt werden, sondern nur statistisch mittels einer großen Anzahl Beobachtungen. Die Werte werden in sogenannten Sterbetafeln zusammengefasst. Eine Größe, welche dabei angegeben wird, ist die Sterbewahrscheinlichkeit im nächsten Jahr, für eine Person im Alter von  $k$  Jahren, bezeichnet mit  $p_k$ . Dabei handelt es sich um die bedingte Wahrscheinlichkeit, dass die Person im nächsten Jahr stirbt, gegeben dass sie aktuell  $k$  Jahre alt ist. Anders ausgedrückt ist das die Wahrscheinlichkeit, genau  $k$  Jahre alt zu werden, unter der Bedingung, mindestens  $k$  Jahre alt zu werden. Mit unserer Wahl von  $\Omega$ ,

$$p_k = \mathbb{P}(k | \{k, k+1, \dots\}).$$

Eine Anmerkung:  $\Omega = \mathbb{N}_0$  ist zwar keine endliche Ergebnismenge, wie in Definition 1.3 gefordert wurde, jedoch kann man damit genauso rechnen, wie in Definition 3.1 gezeigt wird. Alternativ kann man z.B. annehmen, dass niemand älter wird als 120 Jahre. ■

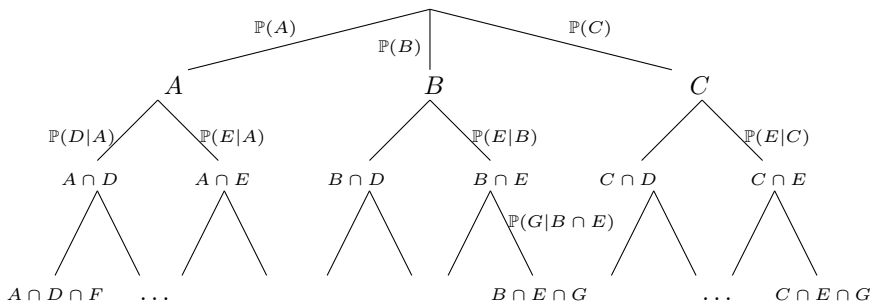


Abb. 2.1: Baumdarstellung eines wiederholten Zufallsexperiments, welches im ersten Schritt die möglichen Ergebnisse  $A, B, C$ , im zweiten Schritt die Ergebnisse  $D, E$  und im dritten Schritt die Ergebnisse  $F, G$  hat.

## 2.2 Mehrstufige Experimente und Formel von der Gesamtwahrscheinlichkeit

Mehrstufige Experimente sind Zufallsexperimente, welche in mehreren Schritten ablaufen, z.B. wiederholtes Würfeln. Es kann hilfreich sein, solche Experimente als *Wahrscheinlichkeitsbaum* darzustellen. Ein „Baum“ besteht dabei aus Kanten (oder Ästen) und Knoten, in denen sich die Äste aufspalten, und welche in Ebenen sortiert sind, beginnend mit einem Knoten. Auf der letzten Ebene enden die Äste in „Blättern“. Die Bäume werden normalerweise von links nach rechts oder von oben nach unten gezeichnet, wie in Abbildung 2.1. In den Baum trägt man nun die Ereignisse und Wahrscheinlichkeiten ein. Dabei geht man nach diesem Schema vor:

- Knoten des Baumes: Diese stellen die Ergebnisse der jeweiligen Stufe des Zufallsexperimentes dar.
- Kanten („Äste“) des Baumes: Hier wird die *bedingte Wahrscheinlichkeit* des entsprechenden Ausgangs eingetragen, gegeben das Ergebnis der vorherigen Stufen, ein.
- „Blätter“ des Baumes: Endergebnis des Experiments entsprechend der Zwischenergebnisse die zu dem jeweiligen Blatt führen.

### Beispiel 2.4 (Test auf Krankheit)

Wir untersuchen eine Krankheit, von der bekannt ist, dass sie bei 2% der Bevölkerung auftritt. Es wird ein Bluttest entwickelt, der die Krankheit mit 99% Wahrscheinlichkeit erkennt, allerdings zeigt er auch bei 3% der gesunden Testpersonen falsch positiv an.

Diese Situation kann man als zweistufiges Zufallsexperiment auffassen: Wir bezeichnen mit  $K$  das Ereignis, dass eine zufällig ausgewählte Person aus der Bevölkerung krank ist, und mit  $T$  die Wahrscheinlichkeit, dass das Testergebnis einer zufällig ausgewählten Person positiv ist. Somit gilt also nach unseren Informationen  $\mathbb{P}(K) = 0.02$ . Außerdem ist die Wahrscheinlichkeit, dass eine Person gesund ist, gegeben durch  $\mathbb{P}(K^c) = 1 - \mathbb{P}(K) =$

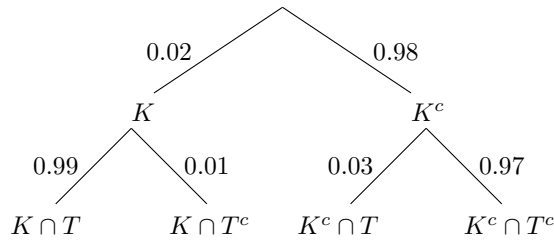


Abb. 2.2: Test auf Krankheit

0.98. Die Information, dass die Krankheit mit 99% Wahrscheinlichkeit erkannt wird, übersetzt sich nun in eine bedingte Wahrscheinlichkeit, nämlich

$$\mathbb{P}(T | K) = 0.99.$$

In Worten ausgedrückt, wenn eine Person krank ist, dann ist der Test mit einer Wahrscheinlichkeit von 0.99 positiv. Andererseits ist er mit 3% Wahrscheinlichkeit falsch positiv, also

$$\mathbb{P}(T | K^c) = 0.03.$$

Natürlich können wir auch wieder die Gegenwahrscheinlichkeiten berechnen:  $\mathbb{P}(T^c | K) = 0.01$ , und  $\mathbb{P}(T^c | K^c) = 0.97$ . Somit können wir die Situation nun wieder als Baum darstellen, siehe Abbildung 2.2. ■

Wenn man die Formel für die bedingte Wahrscheinlichkeit aus Definition 2.1 umformt, erhält man die Gleichung

$$\mathbb{P}(A \cap B) = \mathbb{P}(A | B) \cdot \mathbb{P}(B). \quad (2.1)$$

Da man im Wahrscheinlichkeitsbaum eines mehrstufigen Experiments die jeweiligen bedingten Wahrscheinlichkeiten auf die Äste schreibt, kann man die Definition der bedingten Wahrscheinlichkeit mit Hilfe der Baumdarstellung in eine Rechenregel übersetzen:

**Multiplikationsregel** In einem mehrstufigen Experiment berechnet sich die Wahrscheinlichkeit eines Ergebnisses durch Multiplikation der Wahrscheinlichkeiten entlang der Äste die zum Blatt mit diesem Ergebnis führen.

Aus Abbildung 2.1 liest man z.B.  $\mathbb{P}(B \cap E \cap G) = \mathbb{P}(B) \cdot \mathbb{P}(E|B) \cdot \mathbb{P}(G|B \cap E)$  ab.

### Beispiel 2.5 (Test auf Krankheit, Fortsetzung von Beispiel 2.4)

Im obigen Beispiel können wir die Wahrscheinlichkeit für das Ereignis  $K \cap T$  berechnen als

$$\mathbb{P}(K \cap T) = \mathbb{P}(K) \cdot \mathbb{P}(T|K) = 0.02 \cdot 0.99 = 0.0198.$$

Mit dieser Wahrscheinlichkeit ist also eine zufällig ausgewählte Person krank, und der Test zeigt bei dieser Person die Krankheit an. ■

Wie wir aus Satz 1.1 wissen, berechnet sich die Wahrscheinlichkeit eines Ereignisses durch Summation über die Ergebnisse, die in dem Ereignis enthalten sind. Die Definition der bedingten Wahrscheinlichkeit, oder alternativ die Multiplikationsregel, führt deshalb auf die sogenannte Formel von der *totalen Wahrscheinlichkeit* (oder Formel von der *Gesamtwahrscheinlichkeit*).

**Satz 2.1 (Formel von der Gesamtwahrscheinlichkeit)**

Seien  $A, B \subset \Omega$  Ereignisse, und gelte  $0 < \mathbb{P}(B) < 1$ . Dann gilt

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c).$$

Allgemeiner seien  $A$  ein Ereignis, und  $B_1, \dots, B_n$  eine disjunkte Zerlegung von  $\Omega$ . Dann gilt

$$\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

Disjunkte Zerlegung bedeutet dabei  $\mathbb{P}(B_i) > 0$  für alle  $i = 1, \dots, n$ ,  $B_i \cap B_j = \emptyset$  für  $i \neq j$ , und  $\bigcup_{i=1}^n B_i = \Omega$ .

**Beweis 2.1** Da  $B^c$  das Komplement von  $B$  bezeichnet, gilt  $A = (A \cap B) \cup (A \cap B^c)$ , und  $(A \cap B) \cap (A \cap B^c) = \emptyset$ . Mit der Eigenschaft (P2) aus 1.3 und der Definition der bedingten Wahrscheinlichkeit folgt deshalb

$$\mathbb{P}(A) = \mathbb{P}((A \cap B) \cup (A \cap B^c)) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c).$$

Die allgemeine Formel folgt analog, indem man  $A = (A \cap B_1) \cup \dots \cup (A \cap B_n)$  verwendet.  $\square$

Diese Formel kann man auch in eine Additionsregel für mehrstufige Experimente übersetzen.

**Additionsregel** In einem mehrstufigen Experiment berechnet sich die Wahrscheinlichkeit eines Ereignisses durch Addition der zugehörigen Wahrscheinlichkeiten auf den Blättern des Baumes.

**Beispiel 2.6 (Test auf Krankheit: Fortsetzung von Beispiel 2.4)**

Im Beispiel 2.4 gilt also für die Wahrscheinlichkeit eines positiven Tests

$$\mathbb{P}(T) = \mathbb{P}(T|K)\mathbb{P}(K) + \mathbb{P}(T|K^c)\mathbb{P}(K^c) = 0.99 \cdot 0.02 + 0.03 \cdot 0.98 = 0.0198 + 0.0294 = 0.0492.$$

Dies ist die Wahrscheinlichkeit, dass der Test bei einer zufällig ausgewählten Person positiv ist (ohne zu wissen, ob die Person krank oder gesund ist). Die zu summierenden Ergebnisse sind im Baum in Abbildung 2.3 markiert.  $\blacksquare$

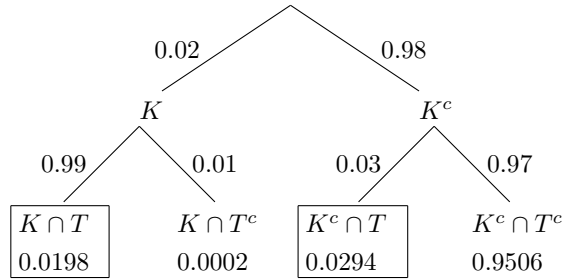


Abb. 2.3: Test auf Krankheit: Additionsregel

**Beispiel 2.7 (Signalübermittlung)**

In einem einfachen Modell der Signalübermittlung wird ein Signal von einem Sender über einen Kanal an einen Empfänger gesendet. Dabei wird das Signal mit einer gewissen Wahrscheinlichkeit gestört. Wir betrachten ein Modell mit mehreren Kanälen, in dem der Sender zufällig einen Kanal zur Übermittlung auswählt. Sei  $a_i$  die Wahrscheinlichkeit, dass Kanal  $i$  gewählt wird,  $i \in \{1, \dots, k\}$ , wobei  $k$  die Anzahl verschiedener Kanäle bezeichnet. Es sei  $0 < a_i < 1$  und  $\sum_{i=1}^k a_i = 1$ , es wird also genau einer der  $k$  Kanäle verwendet. Weiter sei  $p_i$  die bedingte Wahrscheinlichkeit, dass das Signal korrekt ankommt, gegeben dass Kanal  $i$  benutzt wurde. Mit Wahrscheinlichkeit  $1 - p_i$  kommt das Signal bei Verwendung von Kanal  $i$  nicht korrekt an. Was ist die Wahrscheinlichkeit, dass ein versendetes Signal korrekt ankommt? Dies können wir als ein zweistufiges Zufallsexperiment auffassen, wobei in Schritt 1 der Kanal ausgewählt wird, und sich in Schritt 2 entscheidet, ob das Signal korrekt ankommt. Mit der Formel von der Gesamtwahrscheinlichkeit folgt

$$\mathbb{P}(\text{Signal kommt an}) = \sum_{i=1}^k a_i p_i.$$

Dabei haben wir die Ereignisse  $B_i = \{\text{Kanal } i \text{ genutzt}\}$  betrachtet, und  $\mathbb{P}(B_i) = a_i$  sowie  $\mathbb{P}(\text{Signal kommt an} | B_i) = p_i$  verwendet. ■

**2.3 Unabhängigkeit von Ereignissen**

Wir betrachten eine Urne, in der sich vier rote und acht blaue Kugeln befinden, und aus der zwei Mal ohne Zurücklegen gezogen wird. Wir verwenden die Bezeichnungen  $B_i$  für das Ereignis, dass die  $i$ te gezogene Kugel blau ist, und  $R_i = B_i^c$  für das Gegenereignis, dass die  $i$ te Kugel rot ist,  $i = 1, 2$ . Es gilt also  $\mathbb{P}(R_1) = \frac{4}{12} = \frac{1}{3} = 1 - \mathbb{P}(B_1)$ . Beim zweiten Zug müssen wir unterscheiden, ob im ersten Zug rot oder blau gezogen wurde, da wir die Kugeln nicht zurücklegen. Wurde rot gezogen, so befinden sich danach 3 rote und 8 blaue Kugeln in der Urne, also folgt  $\mathbb{P}(R_2 | R_1) = \frac{3}{11}$ . Wurde im ersten Zug blau gezogen, erhalten wir  $\mathbb{P}(R_2 | B_1) = \frac{4}{11}$ , da in diesem Fall noch alle vier roten Kugeln in

der Urne sind, jedoch nur noch sieben blaue. Das Ergebnis des zweiten Zuges hängt also offensichtlich vom Ergebnis des ersten Zuges ab. Was genau „abhängig“ hier bedeuten soll, besagt folgende Definition.

**Definition 2.2 (Unabhängigkeit von Ereignissen)**

Sei  $(\Omega, \mathbb{P})$  ein endlicher Wahrscheinlichkeitsraum, und seien  $A, B \subseteq \Omega$  Ereignisse.  $A$  und  $B$  heißen **unabhängig**, falls

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

gilt. ◆

**Beispiel 2.8 (Urnenmodell)**

Wir verallgemeinern das Beispiel mit der Urne etwas. Betrachte eine Urne, in der sich  $n$  rote und  $m$  blaue Kugeln befinden. Wieder ziehen wir zwei Mal ohne Zurücklegen, und bezeichnen mit  $B_i$  die Wahrscheinlichkeit, im  $i$ ten Zug blau zu ziehen, und  $R_i$  für rot. Es gilt  $\mathbb{P}(R_1) = \frac{n}{n+m}$  und  $\mathbb{P}(B_1) = \frac{m}{n+m}$ . Legen wir die Kugel nicht zurück, so befinden sich nach dem Ziehen einer roten Kugel beim ersten Versuch nur noch  $n-1$  rote Kugeln in der Urne, aber immer noch  $m$  blaue, also insgesamt  $n+m-1$  Kugeln. Es gilt also

$$\mathbb{P}(R_1 \cap B_2) = \mathbb{P}(R_1)\mathbb{P}(B_2 | R_1) = \frac{n}{n+m} \cdot \frac{m}{n+m-1}.$$

Analog folgt

$$\mathbb{P}(B_1 \cap B_2) = \mathbb{P}(B_1)\mathbb{P}(B_2 | R_1) = \frac{m}{n+m} \cdot \frac{m-1}{n+m-1}.$$

Mit der Formel von der Gesamtwahrscheinlichkeit erhalten wir

$$\mathbb{P}(B_2) = \mathbb{P}(R_1 \cap B_2) + \mathbb{P}(B_1 \cap B_2) = \frac{nm + m(m-1)}{(n+m)(n+m-1)} = \frac{m}{n+m}.$$

Somit ist

$$\mathbb{P}(R_1 \cap B_2) = \frac{n}{n+m} \cdot \frac{m}{n+m-1} \neq \frac{n}{n+m} \cdot \frac{m}{n+m} = \mathbb{P}(R_1) \cdot \mathbb{P}(B_2),$$

die Ereignisse sind also nach Definition nicht unabhängig, was unserer Intuition entspricht. Beim Ziehen *mit Zurücklegen* erhalten wir mit unseren Rechenregeln

$$\mathbb{P}(R_1 \cap B_2) = \mathbb{P}(R_1)\mathbb{P}(B_2 | R_1) = \frac{n}{n+m} \cdot \frac{m}{n+m}.$$

Hier sind  $R_1$  und  $R_2$  also unabhängig, denn beim Ziehen mit Zurücklegen bleibt die Wahrscheinlichkeit für das Ergebnis im zweiten Zug unbeeinflusst durch das Ergebnis im ersten Zug. ■

**Satz 2.2**

Zwei Ereignisse  $A$  und  $B$  in einem endlichen Wahrscheinlichkeitsraum  $(\Omega, \mathbb{P})$  mit  $\mathbb{P}(A) > 0$  sind genau dann unabhängig, wenn

$$\mathbb{P}(B | A) = \mathbb{P}(B)$$

gilt.

**Beweis 2.2** Seien  $A$  und  $B$  unabhängig. Somit ist  $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ . Setzen wir das in die Definition der bedingten Wahrscheinlichkeit ein, so folgt

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

Die umgekehrte Beweisrichtung folgt ebenfalls aus dieser Rechnung.  $\square$

Natürlich gilt genauso, dass  $A$  und  $B$  genau dann unabhängig sind, wenn  $\mathbb{P}(A|B) = \mathbb{P}(A)$  erfüllt ist. Aus dieser Darstellung liest man ab, dass  $A$  und  $B$  genau dann unabhängig sind, wenn die Information über das Eintreten von  $B$  die Wahrscheinlichkeit für  $A$  nicht verändert, wenn also die Bedingung, dass  $B$  eingetreten ist, uns nichts über das Eintreten von  $A$  verrät. Im Beispiel 2.8 hatten wir bereits gesehen, dass

$$\mathbb{P}(B_2 | R_1) = \frac{m}{n-1+m} \neq \frac{m}{n+m} = \mathbb{P}(B_2)$$

gilt, dass  $B_2$  und  $R_1$  also nicht unabhängig sind. Vorsicht: Der Begriff der Unabhängigkeit wird manchmal mit der Eigenschaft dass  $A$  und  $B$  disjunkt sind verwechselt. Diese Begriffe sind aber *nicht* gleichbedeutend – eher im Gegenteil: Betrachte in 2.8 das Ereignis  $B_1 \cap R_1$ . Da nicht gleichzeitig eine rote *und* eine blaue Kugel gezogen werden kann, gilt

$$\mathbb{P}(B_1 \cap R_1) = \mathbb{P}(\emptyset) = 0 \neq \mathbb{P}(B_1) \cdot \mathbb{P}(R_1),$$

die Ereignisse sind also nicht unabhängig. Dies ist auch intuitiv klar, denn die Information, dass  $R_1$  eingetreten ist, sagt uns bereits, dass  $B_1$  auf keinen Fall eintreten kann. Es gilt also ganz klar  $\mathbb{P}(B_1|R_1) = 0 \neq \mathbb{P}(B_1)$ .

### Beispiel 2.9 (Vererbung von Farbenblindheit)

Laut Wikipedia sind ca. fünf Prozent der Bevölkerung farbenblind, und zwar etwa neun Prozent der Männer und knapp ein Prozent der Frauen. Die klassische Rot-Grün-Farbenblindheit wird auf dem  $X$ -Chromosom vererbt, wobei Männer nur ein solches Chromosom haben, welches sie von der Mutter erben. Sie sind farbenblind, wenn ihr  $X$ -Chromosom betroffen ist. Frauen haben zwei  $X$ -Chromosome, eines vom Vater und eines von der Mutter. Sie sind farbenblind, wenn beide dieser Chromosome betroffen sind, und Trägerin des Gens wenn eins der beiden Chromosome betroffen ist. In einer Familie mit zwei Brüdern bezeichnen wir mit  $A$  das Ereignis, dass der erste Bruder farbenblind ist, und mit  $B$  das Ereignis, dass der zweite Bruder farbenblind ist. Nach den Angaben von oben gilt  $\mathbb{P}(A) = \mathbb{P}(B) = 0.09$ . Da die Brüder jedoch die selbe Mutter haben, sind die Ereignisse nicht unabhängig. Die bedingte Wahrscheinlichkeit  $\mathbb{P}(A|B)$  ist nämlich nicht gleich 0.09, da die Information, dass  $B$  eingetreten ist, uns sagt, dass die Mutter entweder Trägerin oder selbst farbenblind ist. Somit gilt  $\mathbb{P}(A|B) > 0.5$ , denn der zweite Bruder hat mindestens eine 50% Chance, das betroffene Chromosom von der Mutter zu bekommen, da diese mindestens eins, vielleicht sogar zwei davon hat.  $\blacksquare$

**Beispiel 2.10 (Leitungssystem)**

In einem Leitungssystem mit drei Knotenpunkten gibt es jeweils zwischen zwei Knoten verbindende Leitungen. Die Leitungen können unabhängig voneinander geöffnet oder geschlossen werden. Wir bezeichnen mit  $A_{i,j}$ ,  $1 \leq i < j \leq 3$ , das Ereignis, dass die Leitung zwischen Knoten  $i$  und  $j$  geöffnet ist,  $i = 1, 2, 3$ . Die Wahrscheinlichkeit, dass die Leitung zwischen 1 und 2 sowie die Leitung zwischen 2 und 3 offen ist, wird also gegeben durch

$$\mathbb{P}(A_{1,2} \cap A_{2,3}) = \mathbb{P}(A_{1,2}) \cdot \mathbb{P}(A_{2,3})$$

wegen der Unabhängigkeit. Analog für die beiden anderen Leitungspaare. Da es drei Leitungen gibt, muss aber auch

$$\mathbb{P}(A_{1,2} \cap A_{2,3} \cap A_{1,3}) = \mathbb{P}(A_{1,2}) \cdot \mathbb{P}(A_{2,3}) \cdot \mathbb{P}(A_{1,3})$$

gelten. Wir sprechen also hier von einer Situation mit mehr als drei Ereignissen, für welche die Definition der Unabhängigkeit erweitert werden muss. ■

**Definition 2.3**

Seien  $A_1, \dots, A_n \subseteq \Omega$  Ereignisse. Die  $n$  Ereignisse heißen **unabhängig**, falls für alle  $2 \leq k \leq n$  und Indizes  $i_1, \dots, i_k \in \{1, \dots, n\}$ ,  $i_l \neq i_j$  für  $l \neq j$  gilt

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \mathbb{P}(A_{i_2}) \cdot \dots \cdot \mathbb{P}(A_{i_k}).$$



Drei Ereignissen  $A, B, C$  sind also genau dann unabhängig, wenn *alle* folgenden Gleichungen erfüllt sind:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B), \quad \mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C), \quad \mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C),$$

und

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C).$$

Man kann Beispiele finden, in denen die Gleichungen

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B), \quad \mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C), \quad \mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C)$$

alle erfüllt sind, jedoch  $A, B$  und  $C$  trotzdem nicht unabhängig sind. Ebenso gibt es Beispiele, in denen  $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$  gilt, aber die drei Ereignisse nicht unabhängig sind.

## 2.4 Bayes'sche Umkehrformel

Nicht jede bedingte Wahrscheinlichkeit ist einfach zu berechnen. In praktischen Anwendungen steht man oft vor der Situation, dass man z.B. die bedingte Wahrscheinlichkeit  $\mathbb{P}(A|B)$  (leicht) berechnen kann (z.B. aus einer geeigneten Baumdarstellung), sich jedoch eigentlich für die umgekehrte Wahrscheinlichkeit  $\mathbb{P}(B|A)$  interessiert. In solchen Fällen ist die Bayes'sche Umkehrformel nützlich.



**Satz 2.3 (Bayes-Formel)**

Seien  $A, B$  Ereignisse mit  $0 < \mathbb{P}(A), \mathbb{P}(B) < 1$ . Dann gilt

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)}$$

**Beweis 2.3** Wir können die Definition der bedingten Wahrscheinlichkeit verwenden, und erhalten so

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

Dabei haben wir im letzten Schritt die Definition der bedingten Wahrscheinlichkeit noch ein zweites Mal verwendet. Mit der Formel von der Gesamtwahrscheinlichkeit folgt

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c).$$

Dies setzen wir in der obigen Gleichung auf der rechten Seite ein, und erhalten so das gewünschte Ergebnis.  $\square$

**Beispiel 2.11 (Test auf Krankheit, Fortsetzung von Beispiel 2.4)**

In Beispiel 2.4 können wir uns nun folgende Frage stellen: Wenn eine Person positiv auf besagte Krankheit getestet wird, wie wahrscheinlich ist es dann, dass sie diese Krankheit tatsächlich hat?

Wir interessieren uns also für die bedingte Wahrscheinlichkeit  $\mathbb{P}(K|P)$ . Diese können wir nun aus den bekannten Werten  $\mathbb{P}(K) = 0.02$ ,  $\mathbb{P}(P|K) = 0.99$  und  $\mathbb{P}(T|K^c) = 0.03$  mit Hilfe der Bayes-Formel berechnen. Es gilt nämlich

$$\mathbb{P}(K|T) = \frac{\mathbb{P}(T|K)\mathbb{P}(K)}{\mathbb{P}(K)\mathbb{P}(T|K) + \mathbb{P}(K^c)\mathbb{P}(T|K^c)} = \frac{0.99 \cdot 0.02}{0.99 \cdot 0.02 + 0.98 \cdot 0.03} \approx 0.402$$

Mit anderen Worten, die Wahrscheinlichkeit dass jemand bei positivem Test wirklich krank ist, beträgt ungefähr 0.4. Das bedeutet aber auch, dass die Wahrscheinlichkeit, tatsächlich krank zu sein, auch bei einem positiven Test gerade mal 40% beträgt, also keineswegs nahe an 1 liegt! Das Ergebnis wirkt vielleicht zuerst einmal nicht intuitiv, aber kann dadurch erklärt werden, dass eben nur 1% der Bevölkerung tatsächlich krank ist, wenn man also nach dem Zufallsprinzip testet, man meistens gesunde Personen testet, bei denen wiederum der Test mit einer gewissen (kleinen, aber nicht vernachlässigbaren) Wahrscheinlichkeit falsch positiv ist.

Dieses Beispiel illustriert, dass bedingte Wahrscheinlichkeiten nicht immer rein intuitiv gut geschätzt werden. Gerade wenn man sich für relativ wenig wahrscheinliche Ereignisse interessiert, können bedingte Wahrscheinlichkeiten leicht um Größenordnungen von einer ersten intuitiven Vermutung abweichen. Es ist deshalb wichtig, die Rechenregeln für bedingte Wahrscheinlichkeiten und insbesondere die Bayes-Formel korrekt anzuwenden!

Versprechen wie „der Test hat eine Genauigkeit von 99.999%“ sollten ebenso mit Vorsicht betrachtet werden. Dabei ist zu hinterfragen, was für eine Genauigkeit genau gemeint ist. Im Beispiel sind sowohl  $\mathbb{P}(T|K)$  als auch  $\mathbb{P}(T^c|K^c)$  relativ groß, dennoch ist  $\mathbb{P}(K|T)$  weit entfernt von 1.  $\blacksquare$

Mit Hilfe der allgemeinen Formel von der Gesamtwahrscheinlichkeit erhält man eine allgemeinere Variante der Bayes-Formel:

**Satz 2.4 (Allgemeine Bayes-Formel)**

Sei  $A$  ein Ereignis und  $B_1, \dots, B_n$  eine disjunkte Zerlegung von  $\Omega$ . Dann gilt

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{j=1}^n \mathbb{P}(A|B_j)\mathbb{P}(B_j)}$$

**Beispiel 2.12 (Signalübertragung)**

Ein Sender hat überträgt ein Signal an einen Empfänger, wobei zufällig einer von  $k$  Kanälen ausgewählt wird. Sei  $a_i$  die Wahrscheinlichkeit, dass Kanal  $i$  benutzt wird, wobei  $a_1 + \dots + a_k = 1$  gilt. Falls Kanal  $i$  benutzt wird, kommt das Signal mit Wahrscheinlichkeit  $p_i$  richtig an, und mit Wahrscheinlichkeit  $1 - p_i$  kommt das Signal nicht oder falsch an. Sei  $A$  das Ereignis, dass das Signal richtig ankommt. Nach der Formel für die Gesamtwahrscheinlichkeit gilt dann

$$\mathbb{P}(S) = \sum_{i=1}^k a_i p_i.$$

Falls das Signal angekommen ist, was ist dann die Wahrscheinlichkeit, dass Kanal  $j$  benutzt wurde? Bezeichne mit  $K_j$  dieses Ereignis. Dann gilt also  $\mathbb{P}(K_j) = a_j$  und  $\mathbb{P}(S|K_j) = p_j$ . Nach der Formel von Bayes erhalten wir

$$\mathbb{P}(K_j|S) = \frac{\mathbb{P}(S|K_j)\mathbb{P}(K_j)}{\mathbb{P}(S)} = \frac{p_j a_j}{\sum_{i=1}^k p_i a_i}.$$

■

**Beispiel 2.13 (Markov-Ketten)**

Betrachte ein Netzwerk, bestehend aus  $n$  Knoten. Es wird zufällig ein Startknoten ausgewählt, mit Wahrscheinlichkeit  $a_i$  wählt man Knoten  $i \in \{1, \dots, n\}$ . Nun springt man zu einem anderen Knoten, wobei  $p_{ij}$  die Wahrscheinlichkeit bezeichnet, von Knoten  $i$  zu Knoten  $j$  zu springen. Somit ist man nach einem Sprung mit der Wahrscheinlichkeit  $\sum_{i=1}^n a_i p_{ij}$  in Knoten  $j$ , da man von jedem der Knoten aus starten kann. Was ist die Wahrscheinlichkeit, dass man in Knoten  $l$  gestartet ist, wenn man nach dem ersten Sprung in Knoten  $j$  angekommen ist? Nach der Bayes-Formel ist diese Wahrscheinlichkeit

$$\frac{a_l p_{lj}}{\sum_{i=1}^n a_i p_{ij}}.$$

Dieses Beispiel beschreibt eine sogenannte *Markov-Kette*, welche in Kapitel 11 ausführlicher behandelt werden.

■

### Zusammenfassung

Bedingte Wahrscheinlichkeiten beschreiben Wahrscheinlichkeiten unter Zusatzinformation. Mit ihnen lassen sich mehrstufige Zufallsexperimente beschreiben und untersuchen. Dabei ist die Baumdarstellung nützlich. Es gelten Multiplikations- und Additionsregel, letztere ist die sogenannte Formel von der Gesamtwahrscheinlichkeit.

Beeinflussen zwei Ereignisse sich wechselseitig nicht, so sind sie unabhängig, und die Wahrscheinlichkeit, dass beide eintreten, berechnet sich durch Multiplikation der Einzelwahrscheinlichkeiten.

Mit Hilfe der Bayes-Formel können bedingte Wahrscheinlichkeiten berechnet werden, die in die aus dem ursprünglichen mehrstufigen Experiment nicht direkt bestimmbar sind.

Beim Einschätzen von Größenordnungen von bedingten Wahrscheinlichkeiten kann die Intuition oft weit daneben liegen. Deshalb ist eine genaue Modellierung und Rechnung wichtig.

## Aufgaben

**2.1** Beim Würfeln mit einem fairen Würfel betrachte folgende Ereignisse:  $A = \{\text{Ergebnis mindestens } 4\}$ , und  $B = \{\text{Ergebnis ist ungerade}\}$ . Berechne folgende Wahrscheinlichkeiten: a)  $\mathbb{P}(A)$ , b)  $\mathbb{P}(B)$ , c)  $\mathbb{P}(A \cap B)$ , d)  $\mathbb{P}(B|A)$ .

**2.2** Aus einer Urne mit einem roten, einer blauen und einer weißen Kugel wird drei Mal mit Zurücklegen gezogen. Betrachte die Ereignisse

$A = \{\text{Alle gezogenen Kugeln sind rot}\}$ ,

$B = \{\text{mindestens zwei der gezogenen Kugeln haben dieselbe Farbe}\}$ ,

$C = \{\text{die erste gezogene Kugel ist weiß}\}$ ,  $D = \{\text{die letzte gezogene Kugel ist blau}\}$ .

Berechne die folgenden bedingte Wahrscheinlichkeiten. a)  $\mathbb{P}(A|B)$ , b)  $\mathbb{P}(B|C)$ , c)  $\mathbb{P}(C|D)$ , d)  $\mathbb{P}(D|A)$

**2.3** In einer Gruppe von 640 Frauen und 360 Männern rauchen 15 Prozent der Frauen und 20 Prozent der Männer. a) Mit welcher Wahrscheinlichkeit ist eine zufällig ausgewählte Person dieser Gruppe Frau und Raucherin? b) Mit welcher Wahrscheinlichkeit raucht eine zufällig ausgewählte Person? c) Wenn eine zufällig ausgewählte Person raucht, mit welcher Wahrscheinlichkeit handelt es sich dann um eine Frau?

**2.4** (Situation von Aufgabe 2.2) Aus einer Urne mit einem roten, einer blauen und einer weißen Kugel wird drei Mal mit Zurücklegen gezogen. Betrachte die Ereignisse

$A = \{\text{Alle gezogenen Kugeln sind rot}\}$ ,

$B = \{\text{mindestens zwei der gezogenen Kugeln haben dieselbe Farbe}\}$ ,

$C = \{\text{die erste gezogene Kugel ist weiß}\}$ ,  $D = \{\text{die letzte gezogene Kugel ist blau}\}$ .

Sind die folgenden Ereignisse unabhängig? a)  $A$  und  $B$ , b)  $B$  und  $C$ , c)  $C$  und  $D$ , d)  $A$  und  $D$

**2.5** In einer Fabrik wird ein Produkt von drei Maschinen,  $A$ ,  $B$  und  $C$  produziert. Maschine  $A$  produziert 50 Prozent,  $B$  und  $C$  jeweils 25 Prozent der Gesamtproduktion. Ein Produkt von Maschine  $A$  ist mit Wahrscheinlichkeit 0.02 defekt, von  $B$  mit Wahrscheinlichkeit 0.04 und von  $C$  mit Wahrscheinlichkeit 0.08. a) Mit welcher Wahrscheinlichkeit ist ein Produkt dieser Fabrik defekt? b) Falls ein Produkt defekt ist, mit welcher Wahrscheinlichkeit wurde es von Maschine  $B$  produziert? c) Wenn ein Produkt nicht defekt ist, mit welcher Wahrscheinlichkeit wurde es dann von  $C$  produziert?

**2.6** Zeigen Sie: Sind die Ereignisse  $A$  und  $B$  unabhängig, so sind auch  $A$  und  $B^c$ , sowie  $A^c$  und  $B$ , sowie  $A^c$  und  $B^c$  unabhängig.

**2.7** Es sei  $\Omega = \{a, b, c, d, e, f\}$  mit  $\mathbb{P}(a) = 1/27$ ,  $\mathbb{P}(b) = 14/27$ ,  $\mathbb{P}(c) = P(d) = P(e) = 4/27$  sowie  $A = \{d, e, a\}$ ,  $B = \{c, e, a\}$ ,  $C = \{c, d, a\}$ . a) Zeigen Sie, dass  $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$  gilt. b) Zeigen Sie, dass  $A, B, C$  nicht unabhängig sind.

**2.8** In der Situation von Beispiel 2.12 sei  $k = 4$ ,  $a_1 = a_2 = 1/4$ ,  $a_3 = 1/8$ ,  $a_4 = 3/8$ . Weiter seien  $p_1 = p_2 = p_3 = 0.95$  und  $p_4 = 0.9$ . a) Berechne die Wahrscheinlichkeit, dass ein Signal korrekt übertragen wird. b) Falls das Signal nicht korrekt übertragen wurde, mit welcher Wahrscheinlichkeit wurde es über Kanal 3 oder 4 gesendet?

# 3 Diskrete Zufallsvariablen und Verteilungen

---

## Übersicht

3.1	Zufallsvariablen, Verteilung, Verteilungsfunktion .....	32
3.2	Bernoulli-Verteilung, Binomialverteilung, geometrische Verteilung .....	39
3.3	Poisson-Verteilung und Approximation .....	45
3.4	Zipf-Verteilung und skalenfreie Netzwerke .....	47

---

In Anwendungen kommen zufällige Ereignisse oder Abläufe und deren Wahrscheinlichkeiten oft in komplexen Situationen vor, beispielsweise bei der Modellierung von Position, Intensität und Kapazität von sich zufällig bewegenden Sendern. In solchen Situationen ist es meist nicht möglich oder praktikabel, einen Wahrscheinlichkeitsraum explizit anzugeben. Man verwendet zur Modellierung stattdessen geeignete *Zufallsvariablen*. Oft kann man unter sinnvollen Modellannahmen Aussagen über deren *Verteilung* treffen, welche es erlauben, wichtige Aspekte des Problems zu beschreiben. Diese beiden zentralen Begriffe werden wir in diesem Kapitel einführen und diskutieren.

### Lernziele

- Mit den Begriffen Zufallsvariable, Verteilung und Verteilungsfunktion umgehen können
- Wichtige diskrete Verteilungen und ihr Auftreten kennen
- Mit wichtigen diskreten Verteilungen rechnen können

### Vorkenntnisse

In diesem Kapitel werden Kenntnisse über Funktionen und Abbildungen benötigt. Insbesondere mit den Begriffen Definitionsbereich, Wertebereich, Bild und Urbild sollte souverän umgegangen werden können. Außerdem werden einige aus der Analysis bekannte unendliche Reihen benutzt.

### 3.1 Zufallsvariablen, Verteilung, Verteilungsfunktion

Vor der Einführung der zentralen Begriffe des Kapitels wird der Begriff des Wahrscheinlichkeitsraums verallgemeinert, vergleiche Definition 1.3.

**Definition 3.1 (Diskreter Wahrscheinlichkeitsraum)**

Ein **diskreter Wahrscheinlichkeitsraum** ist ein Paar bestehend aus einer endlichen oder abzählbar unendlichen **Ergebnismenge**  $\Omega$ , und einem **Wahrscheinlichkeitsmaß**  $\mathbb{P}$ , welches jedem Ereignis  $A \subseteq \Omega$  eine Zahl  $\mathbb{P}(A) \in [0, 1]$  zuordnet. Diese Zahl ist die Wahrscheinlichkeit des Ereignisses  $A$ . Dabei ist ein Wahrscheinlichkeitsmaß eine Abbildung aus der Menge aller Teilmengen von  $\Omega$  nach  $[0, 1]$  mit den Eigenschaften

**(P0)**  $0 \leq \mathbb{P}(A) \leq 1$  für alle  $A \subseteq \Omega$

**(P1)**  $\mathbb{P}(\Omega) = 1$

**(P2')** Für  $(A_n)_{n \in \mathbb{N}}$  mit  $A_n \subset \Omega$  für alle  $n \in \mathbb{N}$ , so dass  $A_n \cap A_m = \emptyset$  für alle Paare  $n \neq m$  ist, gilt  $\mathbb{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$ .

◆

Diese Definition unterscheidet sich von Definition 1.3 nur dadurch, dass  $\Omega$  nun nicht mehr endlich sein muss, sondern abzählbar unendlich sein darf, sowie durch die Eigenschaft  $(P2')$ , welche eine Verallgemeinerung von  $(P2)$  ist. Wenn  $(P2')$  erfüllt ist, gilt automatisch auch die Eigenschaft  $(P2)$ .

Ein *allgemeiner Wahrscheinlichkeitsraum* wird definiert wie ein diskreter Wahrscheinlichkeitsraum, wobei die Ergebnismenge  $\Omega$  nicht endlich oder abzählbar unendlich sein muss. Zulässig sind beliebige Mengen, z.B.  $\Omega = \mathbb{R}$ . Der einzige Unterschied besteht darin, dass streng gesehen im Allgemeinen nicht jede Teilmenge von  $\Omega$  ein Ereignis sein kann, sondern nur bestimmte, „geeignete“ Teilmengen von  $\Omega$ . Wir gehen hier nicht auf diese Problematik ein. Für die meisten praktischen Anwendungen spielt sie keine Rolle. Der Einfachheit halber nehmen wir auch im allgemeinen Fall an, alle Teilmengen von  $\Omega$  seine geeignete Ereignisse, und verwenden die Notation  $(\Omega, \mathbb{P})$  auch für beliebige Wahrscheinlichkeitsräume. Für Weiteres sei auf die mathematische Literatur zur maßtheoretischen Stochastik verwiesen.

Alle Resultate aus Kapitel 2 gelten unverändert auch für allgemeine Wahrscheinlichkeitsräume.

**Definition 3.2 (Zufallsvariable)**

Sei  $(\Omega, \mathbb{P})$  ein Wahrscheinlichkeitsraum. Eine (eindimensionale) **Zufallsvariable** auf  $(\Omega, \mathbb{P})$  ist eine Abbildung

$$X : \Omega \rightarrow \mathbb{R}.$$

$X(\Omega) := \{x : \exists \omega \in \Omega : X(\omega) = x\} \subseteq \mathbb{R}$  ist der **Wertebereich** der Zufallsvariablen  $X$ .

Notation: Zufallsvariablen werden normalerweise mit Großbuchstaben vom Ende des Alphabets bezeichnet, also  $X, Y, Z, V, \dots$  oder auch mit  $X_1, X_2, \dots$

◆

Eine Zufallsvariable ist also a priori einfach eine Abbildung von einem Wahrscheinlichkeitsraum in die reellen Zahlen oder in eine Teilmenge davon. In konkreten Situationen beschreiben Zufallsvariablen Größen die im Zufallsexperiment von Bedeutung sind. Wir werden im Verlauf dieses Kapitels verschiedene Beispiele sehen. Zuerst jedoch einige weitere Begriffe.

**Definition 3.3 (Diskrete Zufallsvariablen, Zufallsvektor)**

- Sei  $X : \Omega \rightarrow \mathbb{R}$  eine Zufallsvariable. Wenn der Wertebereich  $X(\Omega)$  endlich oder abzählbar ist, so heißt  $X$  **diskrete** Zufallsvariable.
- Eine Abbildung  $X : \Omega \rightarrow \mathbb{R}^d$  heißt  **$d$ -dimensionale Zufallsvariable** oder **Zufallsvektor**. Wir können dann auch  $X = (X_1, \dots, X_d)$  schreiben, wobei  $X_1, \dots, X_d$  (eindimensionale) Zufallsvariablen sind.



Falls  $(\Omega, \mathbb{P})$  ein diskreter Wahrscheinlichkeitsraum ist, so ist jede Zufallsvariable auf  $\Omega$  diskret. Beispielsweise ist  $\mathbb{N}_0$  ein möglicher Wertebereich einer diskreten Zufallsvariablen.

**Beispiel 3.1 (Würfeln)**

Wir nehmen einen fairen Würfel, und werfen ihn zwei Mal hintereinander. Wir wissen bereits aus Beispiel 1.6, dass ein geeigneter Laplace-Wahrscheinlichkeitsraum gegeben ist durch

$$\Omega = \{\omega = (\omega_1, \omega_2) : \omega_1, \omega_2 \in \{1, \dots, 6\}\}.$$

Auf diesem Wahrscheinlichkeitsraum können wir verschiedene Zufallsvariablen definieren, zum Beispiel:

- $X_1 : \Omega \rightarrow \mathbb{R} : X_1(\omega) = \omega_1$ . In Worten ausgedrückt beschreibt  $X_1$  das Ergebnis des ersten Wurfs.
- Analog können wir als  $X_2 : \Omega \rightarrow \mathbb{R} : X_2(\omega) = \omega_2$  das Ergebnis des zweiten Wurfs definieren.
- Sei  $Z : \Omega \rightarrow \mathbb{R}$  definiert als  $Z(\omega) = \omega_1 + \omega_2$ . Somit bezeichnet  $Z$  die Summe der beiden Ergebnisse.
- Mit  $Y$  wollen wir das Minimum der beiden Ergebnisse bezeichnen, also  $Y(\omega) = \min\{\omega_1, \omega_2\}$ .

Wie sehen die Wertebereiche dieser Zufallsvariablen aus? Jede dieser Zufallsvariablen kann nur bestimmte Werte annehmen. Für  $X_1$  sind das die Werte von  $\omega_1$ , da ja  $X(\omega) = \omega_1$  ist. Nach Definition ist  $\omega_1 \in \{1, \dots, 6\}$ . Somit ist

$$X_1(\Omega) = \{1, \dots, 6\}.$$

Natürlich kann  $X_2$  die gleichen Werte annehmen wie  $X_1$ . Betrachten wir also die Zufallsvariable  $Z$ . Hier haben wir  $Z(\omega) = \omega_1 + \omega_2$ , wobei  $\omega_1$  und  $\omega_2$  jeweils aus der Menge  $\{1, \dots, 6\}$  stammen. Die Summe kann also die Werte zwischen  $1 + 1 = 2$  und  $6 + 6 = 12$  annehmen, also

$$Z(\Omega) = \{2, 3, \dots, 12\}.$$

Ähnliche Überlegungen führen auf  $Y(\Omega) = \{1, \dots, 6\}$ .



**Beispiel 3.2 (Wartezeit)**

Die Zeit bis zum Zerfall eines radioaktiven Nukleids oder die Laufzeit eines zufälligen Algorithmus können mittels Zufallsvariablen beschrieben werden. Diese können Werten in den positiven reellen Zahlen  $[0, \infty[$  annehmen, sind also keine diskreten Zufallsvariablen. Sie können jedoch durch Runden z.B. auf ganze Sekunden diskretisiert werden. Typischerweise sind Ereignisse der Form  $\{\omega \in \Omega : X(\omega) \leq 100\}$  oder  $\{\omega \in \Omega : X(\omega) \in [10, 20]\}$  von Interesse. ■

**Beispiel 3.3 (Indikatorfunktion)**

Betrachten wir ein Ereignis, das heißt, eine Teilmenge  $A \subset \Omega$ . Die **Indikatorfunktion** von  $A$  ist definiert als die Abbildung  $1_A : \Omega \rightarrow \{0, 1\}$ ,

$$1_A(\omega) = \begin{cases} 1 & \text{falls } \omega \in A, \\ 0 & \text{sonst.} \end{cases}$$

$1_A$  bezeichnet also die Funktion auf  $\Omega$ , welche genau dann den Wert 1 annimmt, wenn  $\omega$  in  $A$  liegt, und andernfalls gleich 0 ist. Das bedeutet, dass  $1_A$  anzeigt, ob ein Ergebnis zum Ereignis  $A$  gehört oder nicht. Daher der Name Indikatorfunktion. Indikatorfunktionen sind nützlich zur Formalisierung vieler Sachverhalte. Als Beispiel betrachten wir wieder den zweifachen Münzwurf. Ein Ereignis ist beispielsweise

$$A = \{\omega \in \Omega : \omega_1 + \omega_2 \leq 3\}.$$

Für die Zufallsvariable  $1_A$  gilt also  $1_A((1, 2)) = 1$ , da für  $\omega = (1, 2)$  gilt  $\omega_1 + \omega_2 = 1 + 2 = 3$ , aber  $1_A((4, 1)) = 0$ , da hier die Summe strikt größer als 3 ist. ■

**Definition 3.4 (Funktionen von Zufallsvariablen)**

Seien  $X_1, \dots, X_n, n \geq 1$ , Zufallsvariablen, und sei  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  eine Funktion von  $n$  Variablen. Dann ist  $f(X_1, \dots, X_n) : \Omega \rightarrow \mathbb{R}$  definiert durch  $f(X_1, \dots, X_n)(\omega) := f(X_1(\omega), \dots, X_n(\omega))$  wieder eine Zufallsvariable. ♦

**Beispiel 3.4 (Funktionen von Zufallsvariablen, Fortsetzung von Beispiel 3.1)**

Betrachte die Zufallsvariablen  $X_1$  und  $X_2$  aus Beispiel 3.1, d.h.  $X_1$  ist das Ergebnis des ersten Wurfes und  $X_2$  das Ergebnis des zweiten Wurfes. Als Funktion können wir  $f : \mathbb{R}^2 \rightarrow \mathbb{R} : f(x_1, x_2) = x_1 + x_2$  betrachten. Somit ist  $f(X_1, X_2) = X_1 + X_2$  als Abbildung von  $\Omega \rightarrow \mathbb{R}$  wiederum eine Zufallsvariable, mit  $(X_1 + X_2)(\omega) = X_1(\omega) + X_2(\omega) = \omega_1 + \omega_2$ . Damit ist  $X_1 + X_2$  dieselbe Zufallsvariable wie  $Z$  im Beispiel 3.1, nämlich die Summe der beiden Würfe. Ebenso sind Differenzen, Produkte etc. von Zufallsvariablen wieder Zufallsvariablen.

Ist beispielsweise  $g(x) = x^2$ , so ist nach dieser Definition  $g(X) = X_1^2$  wieder eine Zufallsvariable. Sie nimmt genau dann den Wert  $x^2$  an, wenn  $X_1$  den Wert  $x$  hat. Da  $X_1$  den Wertebereich  $X_1(\Omega) = \{1, \dots, 6\}$  hat, ist der Wertebereich von  $X_1^2$  gegeben durch  $X_1^2 = \{1, 4, 9, 16, 25, 36\}$ . ■



**Definition 3.5 (Kurzschreibweise für Ereignisse und Wahrscheinlichkeiten)**

Sei  $X$  eine Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathbb{P})$ . Für jede Menge  $E \subseteq X(\Omega)$  ist das Urbild  $X^{-1}(E) = \{\omega \in \Omega : X(\omega) \in E\}$  ein Ereignis. Für Ereignisse dieser Form führen wir eine oft verwendete Kurzschreibweise ein: Für  $E \subseteq X(\Omega)$  bzw. für  $x \in X(\Omega)$  schreiben wir

- $\{X \in E\} := \{\omega \in \Omega : X(\omega) \in E\} = X^{-1}(E)$
- $\{X = x\} := \{\omega \in \Omega : X(\omega) = x\} = X^{-1}(\{x\})$
- $\{X \leq x\} := \{\omega \in \Omega : X(\omega) \leq x\} = X^{-1}((-\infty, x])$

etc. Außerdem schreiben wir für die entsprechenden Wahrscheinlichkeiten kurz

- $\mathbb{P}(X \in E) := \mathbb{P}(\{\omega \in \Omega : X(\omega) \in E\})$ ,
- $\mathbb{P}(X \leq x) := \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\})$  etc.

◆

Wir kommen nun zum zweiten zentralen Begriff dieses Kapitels.

**Definition 3.6 (Verteilungsfunktion)**

Sei  $X : \Omega \rightarrow \mathbb{R}$  eine Zufallsvariable. Die **(kumulative) Verteilungsfunktion**  $F_X$  von  $X$  ist definiert als

$$F_X : \mathbb{R} \rightarrow [0, 1], \quad F_X(x) := \mathbb{P}(X \leq x).$$

◆

Wenn wir die Verteilungsfunktion einer Zufallsvariablen kennen, so können wir damit also (im Prinzip) die Wahrscheinlichkeiten aller Ereignissen der Form  $\{X \leq x\}, x \in \mathbb{R}$ , bestimmen. Die Verteilungsfunktion sollte nicht verwechselt werden mit einem ähnlich klingenden Begriff:

**Definition 3.7 (Diskrete Verteilung)**

Sei  $X : \Omega \rightarrow \mathbb{R}$  eine diskrete Zufallsvariable. Die **Verteilung** (oder Wahrscheinlichkeitsfunktion) von  $X$  ist definiert als

$$p_X : \mathbb{R} \rightarrow [0, 1], \quad p_X(k) := \mathbb{P}(X = k).$$

◆

Im Fall von nicht diskreten Zufallsvariablen ist die Verteilung nicht definiert, aus Gründen die in Kapitel 6 diskutiert werden. Die Verteilungsfunktion ist hingegen für jede Zufallsvariable definiert.

Falls  $k \notin X(\Omega)$  ist, so gilt  $p_X(k) = 0$ . Die Verteilung von  $X$  wird also durch Angabe von  $X(\Omega)$  sowie von  $p_X(k)$  für alle  $k \in X(\Omega)$  festgelegt.

**Satz 3.1**

Sei  $X$  eine Zufallsvariable mit Verteilungsfunktion  $F_X$ . Dann gelten

1.  $\mathbb{P}(X > x) = 1 - F_X(x)$

$k$	2	3	4	5	6	7	8	9	10	11	12
	(1,1)	(1,2) (2,1)	(1,3) (3,1) (2,2)	(1,4) (4,1) (2,3) (1,4)	(1,5) (5,1) (2,4) (4,2) (3,3)	(1,6) (6,1) (2,5) (5,2) (3,4) (4,3)	(2,6) (6,2) (3,5) (5,3) (4,4)	(3,6) (6,3) (4,5) (5,4)	(4,6) (6,4) (5,5)	(5,6) (6,5)	(6,6)
$\mathbb{P}(Z = k)$	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

Tab. 3.1: Verteilung der Summe von zwei Würfeln

2. Für  $a < b, a, b \in \mathbb{R}$ , gilt  $\mathbb{P}(X \in (a, b]) = F_X(b) - F_X(a)$ .

Ist  $X$  diskret mit Verteilung  $p_X$ , so gelten außerdem

3.  $\sum_{k \in X(\Omega)} p_X(k) = 1$

4.  $F_X(k) = \sum_{l \leq k} p_X(l)$ .

**Beweis 3.1** 1. Es gilt  $\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x) = 1 - F_X(x)$  nach Eigenschaft **(P1)** aus Definition 3.1 und Definition 3.6.

2. Somit folgt  $\mathbb{P}(X \in (a, b]) = \mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a)$ .

3. Dies folgt aus **(P2')** aus Definition 3.1.

4.  $F_X(k) = \mathbb{P}(X \leq k) = \sum_{l \leq k} \mathbb{P}(X = l) = \sum_{l \leq k} p_X(l)$ .

□

### Beispiel 3.5 (Summe beim zweifachen Würfeln, Fortsetzung von Beispiel 3.4)

Wir werfen zwei faire Würfel, und bezeichnen mit  $Z$  die Summe der beiden gewürfelten Augenzahlen, siehe auch Beispiel 3.1. Wie dort wählen wir als Wahrscheinlichkeitsraum  $\Omega = \{(\omega_1, \omega_2) : \omega_i \in \{1, \dots, 6\}, i = 1, 2\}$ . Das ist nützlich, da es sich um einen Laplace-Raum handelt, und wir wissen, dass  $\mathbb{P}(\omega) = \frac{1}{36}$  für jedes  $\omega$  gilt.  $Z$  ist also eine Zufallsvariable auf  $\Omega$ , und es gilt  $Z(\Omega) = \{2, 3, \dots, 12\}$ . Um für  $x \in Z(\Omega)$  die jeweiligen Wahrscheinlichkeitsgewichte  $p_Z(x) = \mathbb{P}(Z = k)$  zu bestimmen, überlegen wir uns, wie jeweils die Menge  $\{\omega \in \Omega : Z(\omega) = x\}$  aussieht. Betrachten wir den Fall  $x = 2$ . Die Summe zweier Würfe kann genau dann 2 sein, wenn beide Würfel 1 gezeigt haben. Somit ist  $\{\omega \in \Omega : Z(\omega) = 2\} = \{(1, 1)\}$ , und deshalb  $\mathbb{P}(Z = 2) = \mathbb{P}(\{(1, 1)\}) = \frac{1}{36}$ . Das Ergebnis  $x = 3$  kann jedoch auf zwei Arten erreicht werden: Erster Wurf 1 und zweiter Wurf 2, oder erster Wurf 2 und zweiter Wurf 1. Somit ist  $\mathbb{P}(Z = 3) = \mathbb{P}(\{(1, 2), (2, 1)\}) = \frac{2}{36} = \frac{1}{18}$ . In der Tabelle 3.1 sind alle Ergebnisse und die zugehörigen Wahrscheinlichkeiten aufgelistet.

Wir sehen also, dass  $Z = 7$  die höchste Wahrscheinlichkeit hat – eine Tatsache, die aus Gesellschaftsspielen bekannt sein dürfte. Der Grund ist, dass  $Z = 7$  aus der höchsten Zahl von möglichen Ergebnissen zusammengesetzt werden kann. ■

### Beispiel 3.6 (Diskrete Gleichverteilung)

Ein wichtiges Beispiel für eine diskrete Verteilung ist gegeben, wenn  $X$  mit gleicher Wahrscheinlichkeit jeden der Werte in einer Menge mit  $n \in \mathbb{N}$  Elementen annehmen

kann. Ohne Beschränkung der Allgemeinheit können wir dann  $X(\Omega) = \{1, \dots, n\}$  wählen, und es gilt

$$\mathbb{P}(X = k) = \frac{1}{n}, \quad k \in \{1, \dots, n\}.$$

Da jedes  $k$  dieselbe Wahrscheinlichkeit hat, entspricht das einem Laplace-Raum. Wird umgangssprachlich gesagt, dass etwas „zufällig“ geschieht, dann ist oft implizit eine Gleichverteilung gemeint. ■

### Beispiel 3.7 (Funktionen von Zufallsvariablen)

Sei  $X$  eine Zufallsvariable, welche einer diskreten Gleichverteilung auf  $\{-2, -1, 0, 1, 2, 3\}$  folgt, es gilt also  $\mathbb{P}(X = k) = \frac{1}{6}$  für jeden dieser Werte. Wie sieht die Verteilung von  $X^2$  aus? Da  $X$  die Werte  $-2, -1, 0, 1, 2, 3$  annimmt, kann  $X^2$  die Werte  $0, 1, 4, 9$  annehmen. Es gilt

$$\mathbb{P}(X^2 = 0) = \mathbb{P}(X = 0) = \frac{1}{6}, \quad \mathbb{P}(X^2 = 1) = \mathbb{P}(X = -1) + \mathbb{P}(X = 1) = \frac{1}{3},$$

$$\mathbb{P}(X^2 = 4) = \mathbb{P}(X = -2) + \mathbb{P}(X = 2) = \frac{1}{3}, \quad \mathbb{P}(X^2 = 9) = \mathbb{P}(X = 3) = \frac{1}{6}.$$

■

### Beispiel 3.8 (Gleichverteilung auf $[0, 1]$ )

Das Intervall  $[0, 1]$  enthält unendlich (sogar überabzählbar) viele Werte. Wie definiert man darauf eine Zufallsvariable, die jeden Wert mit gleicher Wahrscheinlichkeit annimmt? Diese Wahrscheinlichkeit für einen einzelnen Wert  $x$  müsste dann gleich 0 sein. Wir verwenden das Prinzip der „Zielscheibe“ aus Beispiel 1.2. Statt eines einzelnen Werts  $x \in [0, 1]$  betrachten wir das Intervall  $[0, x]$ , und fordern, dass die Wahrscheinlichkeit, dass  $X$  einen Wert in  $[0, x]$  annimmt, proportional zur Länge dieses Intervalls, also zu  $x$ , ist. Es soll also für alle  $x \in [0, 1]$  gelten

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

Für  $0 \leq a \leq b \leq 1$  gilt dann  $\mathbb{P}(X \in [a, b]) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = b - a$ . Eine solche Zufallsvariable ist *gleichverteilt* auf  $[0, 1]$ . Wir werden dieses Beispiel später auf beliebige Intervalle verallgemeinern. Hier handelt es sich nicht um eine diskrete Zufallsvariable. ■

### Beispiel 3.9 (Simulation von Zufallsvariablen mit endlichem Wertebereich)

Das Prinzip der Zielscheibe kann auch zur Simulation von Zufallsvariablen verwendet werden. Sei  $X$  eine Zufallsvariable mit endlichem Wertebereich  $X(\Omega) = \{1, \dots, n\}$ , und sei  $p_X(k) = \mathbb{P}(X = k)$  für alle  $k$  bekannt. Wir können das Intervall  $[0, 1]$  in  $n$  disjunkte Teilintervalle zerlegen, wobei das  $k$ -te Teilintervall  $I_k$  die Länge  $p_X(k)$  haben soll, siehe Abbildung 3.1.

Nun sei  $Z$  eine auf  $[0, 1]$  gleichverteilte Zufallsvariable. Wir definieren in Abhängigkeit von  $Z$  eine neue Zufallsvariable  $Y$ , wobei  $Y = k$  sein soll, falls  $Z$  einen Wert im Intervall  $I_k$  annimmt. Die Zufallsvariable  $Z$  spielt also sozusagen die Rolle des Dartspeils, der den Wert von  $Y$  bestimmt, da das Intervall  $I_k$  mit Wahrscheinlichkeit  $p_X(k)$  getroffen

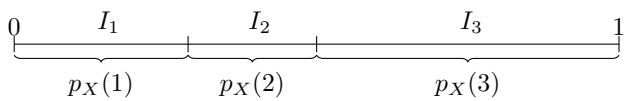


Abb. 3.1: Intervallmethode für diskrete Zufallsvariablen

wird. Damit hat  $Y$  dieselbe Verteilung wie  $X$ , wir können die beschriebene Konstruktion also als Simulation für  $X$  verwenden. Diesem Prinzip werden wir später noch einmal begegnen. ■

**Beispiel 3.10 (Häufigkeiten, Histogramme und Verteilungen)**

Genau wie Wahrscheinlichkeiten sind auch Zufallsvariablen (und Verteilungen) ein mathematisches Konzept, und können in der Wirklichkeit nicht direkt beobachtet werden. Messen können wir dagegen einzelne konkrete Ergebnisse von Zufallsexperimenten, bzw. deren jeweiligen Häufigkeiten. Beobachtet man ein Ergebnis eines Zufallsexperiments und bestimmt daraus den Wert einer dadurch bestimmten Zufallsvariablen, so nennt man das Ergebnis daraus eine *Realisierung* einer Zufallsvariablen. Führt man das Experiment mehrfach aus, kann man so mehrere Realisierungen bestimmen, und somit wiederum *Häufigkeiten* für die auftretenden Werte bestimmen. Häufigkeiten werden gern in *Histogrammen* dargestellt. In Abbildung 3.2 sind gemessene und theoretische Häufigkeiten aus Beispiel 3.5 dargestellt.

Werden eine hinreichend große Anzahl an (unabhängigen) Realisierungen derselben diskreten Zufallsvariablen beobachtet, so sollte die Form des Histogramms der Verteilung der Zufallsvariablen sehr ähnlich sein – denn wir hatten in Kapitel 1.2 gesehen, dass die Häufigkeiten jeweils den Wahrscheinlichkeiten entsprechen. ■

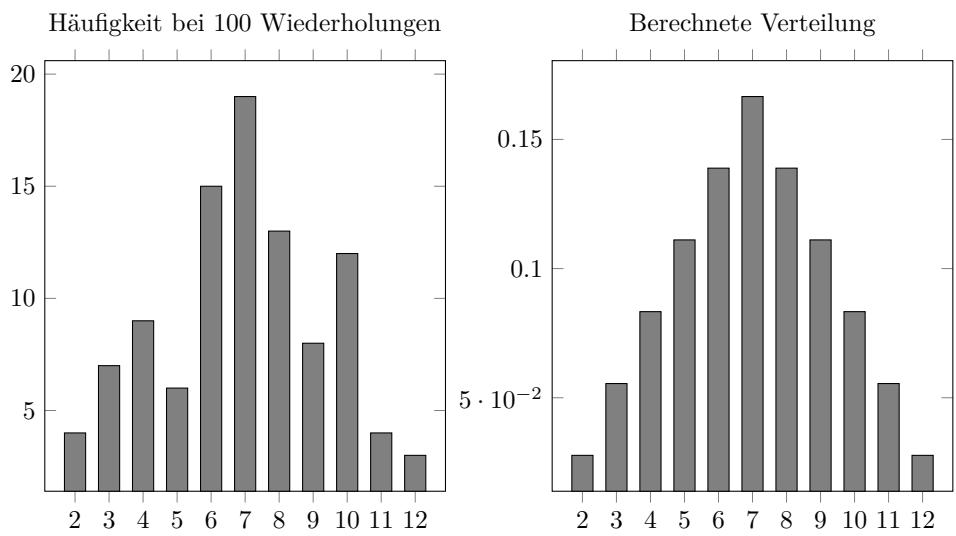


Abb. 3.2: Summe der Augenzahlen beim zweifachen Würfeln

## 3.2 Bernoulli-Verteilung, Binomialverteilung, geometrische Verteilung

Mehrere wichtige Verteilungen sind über ein besonders einfaches Zufallsexperiment definiert.

### Definition 3.8 (Bernoulli-Experiment)

Ein **Bernoulli-Experiment** ist ein Zufallsexperiment, welches genau zwei mögliche Ausgänge hat. Die beiden möglichen Ausgänge werden mit **Erfolg** und **Misserfolg** bezeichnet. Ein **wiederholtes Bernoulli-Experiment** ist ein mehrfach ausgeführtes Bernoulli-Experiment, bei dem die Wahrscheinlichkeiten für Erfolg und Misserfolg in jeder Ausführung gleich ist, und bei dem die Ergebnisse der einzelnen Ausführungen unabhängig voneinander sind. ♦

Die beiden möglichen Ausgänge bezeichnet man manchmal auch mit 0 und 1. Da es nur zwei Möglichkeiten gibt, muss nur die Wahrscheinlichkeit für eins der beiden möglichen Ergebnisse bekannt sein, das andere berechnet sich dann über die Gegenwahrscheinlichkeit. Normalerweise wählt man die Bezeichnung  $p = \mathbb{P}(1) = \mathbb{P}(\text{“Erfolg“})$  als die sogenannte **Erfolgswahrscheinlichkeit**. Ein wiederholtes Bernoulli-Experiment ist also eine mehrfache Ausführung unter gleichbleibenden Bedingung und mit unabhängigen Wiederholungen.

### Definition 3.9 (Bernoulli-Verteilung)

Eine Zufallsvariable  $X : \Omega \rightarrow \mathbb{R}$  heißt **Bernoulli-verteilt zum Parameter**  $p \in [0, 1]$  (bzw. hat die Bernoulli-Verteilung zum Parameter  $p$ ), falls

$$X(\Omega) = \{0, 1\} \quad \text{und} \quad \mathbb{P}(X = 1) = p$$

gelten. ♦

Da  $X$  nur zwei Werte annimmt, folgt sofort  $\mathbb{P}(X = 0) = 1 - \mathbb{P}(X = 1) = 1 - p$ .

### Beispiel 3.11 (Münzwurf)

Das einmalige Werfen einer Münze hatten wir mit dem Ereignisraum  $\Omega = \{K, Z\}$  modelliert. Setzen wir nun als Zufallsvariable  $X := 1_K$ , so ist  $X(\omega) = 1$  falls  $\omega = K$  ist, also falls Kopf geworfen wird, und  $X(\omega) = 0$  für  $\omega = Z$ . Bei einer fairen Münze gilt

$$\mathbb{P}(X = 1) = \mathbb{P}(X = 0) = \frac{1}{2}.$$

Ist die Münze nicht (notwendigerweise) fair, können wir  $p := \mathbb{P}(X = 1) \in [0, 1]$  setzen, und somit  $\mathbb{P}(X = 0) = 1 - p$ . In beiden Fällen handelt es sich also um eine Bernoulli-Verteilung, die faire Münze ist ein Spezialfall. Werfen wir die Münze  $k$  Mal hintereinander, so bekommen wir auf diese Weise die Zufallsvariablen  $X_1, \dots, X_k$ , welche jeweils einzeln Bernoulli-verteilt mit demselben Parameter  $p \in [0, 1]$  ist. ■

**Beispiel 3.12 (Qualitätskontrolle)**

Aus einer großen Zahl  $N$  von gleichartigen Produkten werden  $n$  zufällig ausgewählte getestet. Einige davon weisen Fehler auf. Schreibe  $A_k, k = 1, \dots, n$  für das Ereignis, dass der  $k$ te getestete Artikel fehlerhaft ist, und  $X_k = 1_{A_k}$ . Nach Beispiel 3.3 sind die  $X_k$  Zufallsvariablen, und falls alle Produkte gleichartig sind, existiert eine Wahrscheinlichkeit  $p \in [0, 1]$  so dass

$$\mathbb{P}(X_1 = 1) = \mathbb{P}(X_2 = 1) = \dots = \mathbb{P}(X_n = 1) = p$$

ist (und  $\mathbb{P}(X_k = 0) = 1 - p, k = 1, \dots, n$ ). Jede Zufallsvariable  $X_k$  ist also Bernoulli-verteilt mit einem unbekannten Parameter  $p$ . Die beiden Ausgänge des Bernoulli-Experiments sind „fehlerhaft“ bzw. „nicht fehlerhaft“. ■

Die Bernoulli-Verteilung ist ihrer Einfachheit wegen geeignet, als „Baustein“ für kompliziertere Situationen zu dienen.

**Definition 3.10 (Binomialverteilung)**

Seien  $n \in \mathbb{N}$  und  $p \in [0, 1]$ . Eine Zufallsvariable  $X$  heißt **binomialverteilt mit Parametern  $n$  und  $p$**  falls gilt:  $X(\Omega) = \{0, \dots, n\}$ , und

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



Hierbei ist  $\binom{n}{k}$  der *Binomialkoeffizient*:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{1 \cdot 2 \cdot \dots \cdot n}{1 \cdot \dots \cdot k \cdot 1 \cdot \dots \cdot (n-k)}.$$

Diese Verteilung sieht bereits etwas komplizierter aus als die Bernoulli-Verteilung. Bevor wir sehen, welche Bedeutung die Binomialverteilung hat, rechnen wir erst nach, dass es sich tatsächlich um eine Wahrscheinlichkeitsverteilung handelt.

**Satz 3.2**

Für die *Binomialverteilung* gelten  $\mathbb{P}(X = k) \geq 0$  für alle  $k = 0, \dots, n$ , und  $\sum_{k=0}^n \mathbb{P}(X = k) = 1$ .

**Beweis 3.2** Dass  $\mathbb{P}(X = k) \geq 0$  ist, folgt direkt aus der Definition. Die zweite Behauptung beweisen wir mittels des binomischen Lehrsatz, der besagt, dass für zwei Zahlen  $a, b \in \mathbb{R}$  gilt  $(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$ . Mit  $a = p$  und  $b = 1 - p$  folgt also

$$\sum_{k=0}^n \mathbb{P}(X = k) = \sum_{k=0}^n \binom{n}{k} p^k (1 - p)^{n-k} = (p + (1 - p))^n = 1^n = 1.$$

□

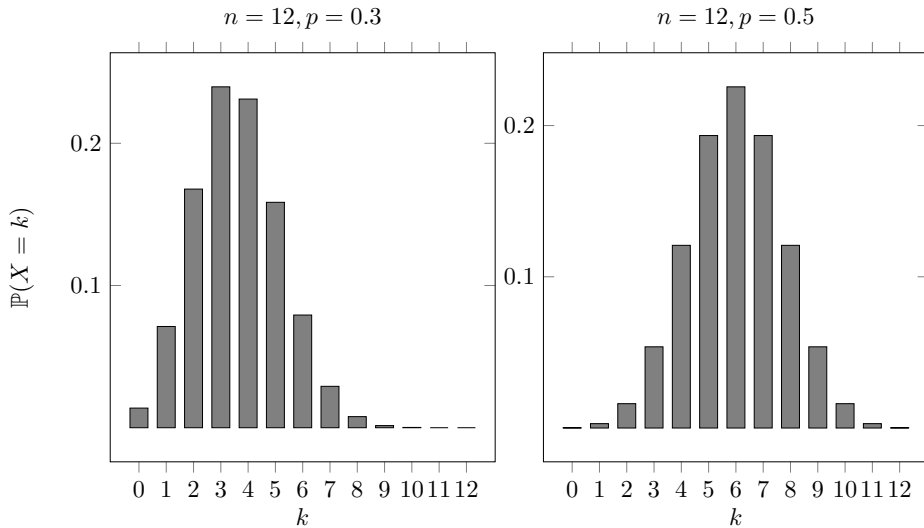


Abb. 3.3: Binomialverteilung

**Satz 3.3 (Binomialverteilung im wiederholten Bernoulli-Experiment)**

Ein Bernoulli-Experiment mit Erfolgswahrscheinlichkeit  $p$  wird  $n$  mal wiederholt. Die Ergebnisse der einzelnen Wiederholungen seien dabei unabhängig. Sei  $X$  die Anzahl Erfolge. Dann ist  $X$  binomialverteilt mit Parameter  $n$  und  $p$ .

**Beweis 3.3** Wir geben nur eine Skizze des Beweises. Nach der Multiplikationsregel hat beim Bernoulli-Experiment jedes Ergebnis, bei dem  $k$  Erfolge und  $n - k$  Misserfolge auftreten, die Wahrscheinlichkeit  $p^k(1-p)^{n-k}$ . Nun gibt es bei  $n$  Wiederholungen genau  $\binom{n}{k}$  Ergebnisse, bei denen  $k$  Misserfolge auftreten (zur Veranschaulichung: Ist  $k = 1$ , so dann dieser eine Erfolg im ersten Versuch, im zweiten Versuch, im dritten Versuch etc. auftreten, also  $n = \binom{n}{1}$  Möglichkeiten). Dies erklärt den Vorfaktor.  $\square$

**Beispiel 3.13 (Urnenmodell, Binomialverteilung)**

Wir betrachten eine Urne, welche 20 rote und 30 weiße Kugeln enthält. Daraus ziehen wir 4 Kugeln mit Zurücklegen, d.h. wir ziehen eine Kugel, notieren ob sie rot oder weiß ist, legen sie zurück, mischen gut, ziehen wieder eine Kugel, usw. bis wir 4 Kugeln gezogen haben. Wie groß ist die Wahrscheinlichkeit, dabei mindestens drei Mal eine weiße Kugel zu ziehen?

Wir bezeichnen die Anzahl weißer Kugeln mit  $X$ . Bei diesem Experiment handelt es sich um ein wiederholtes Bernoulli-Experiment, denn wir haben in jedem Versuch genau zwei Möglichkeiten, nämlich eine rote oder eine weiße Kugel zu ziehen. Nennen wir das Ergebnis „weiß“ den Erfolg. Was ist dann die Erfolgswahrscheinlichkeit  $p$ ? Dies ist nach Definition die Wahrscheinlichkeit, mit der in *einem* Versuch, also einmal Ziehen, Erfolg, also weiß kommt. Somit ist

$$p = \frac{\text{Anzahl weißer Kugeln in der Urne}}{\text{Gesamtzahl der Kugeln in der Urne}} = \frac{30}{50} = \frac{3}{5}.$$

Da wir viermal Ziehen, ist die Anzahl der Wiederholungen  $n = 4$ . Somit ist also nach Satz 3.3  $X$  binomialverteilt mit Parametern  $p = 3/5$  und  $n = 4$ , also gilt,

$$\mathbb{P}(X = k) = \binom{4}{k} \left(\frac{3}{5}\right)^k \left(1 - \frac{3}{5}\right)^{4-k}, \quad k = 0, \dots, 4.$$

Da wir die Wahrscheinlichkeit für  $X \geq 3$  berechnen wollen, erhalten wir

$$\mathbb{P}(X \geq 3) = \mathbb{P}(X = 3) + \mathbb{P}(X = 4) = \binom{4}{3} \left(\frac{3}{5}\right)^3 \left(\frac{2}{5}\right)^1 + \binom{4}{4} \left(\frac{3}{5}\right)^4 \left(\frac{2}{5}\right)^0 = \frac{297}{625}$$

■

### Beispiel 3.14 (Formalisierung mittels Indikatorfunktion)

Wiederholte Bernoulli-Experimente können stets auf einfache Weise mittels der Indikatorfunktion formalisiert werden. Dabei definiert man Zufallsvariablen  $X_k$  als die Indikatorfunktion des Ereignisses, dass im  $k$ -ten Versuch ein Erfolg beobachtet wurde. In Beispiel 3.13 beispielsweise wäre  $X_k := 1_{\{\text{im } k\text{-ten Zug eine weiße Kugel}\}}$ . Dann ist nach Satz 3.3 die Zufallsvariable

$$Y := \sum_{k=0}^n X_k = \sum_{k=0}^n 1_{\{\text{im } k\text{-ten Versuch Erfolg}\}}$$

binomialverteilt mit Parameter  $n$  und  $p$ , wenn  $p$  die Erfolgswahrscheinlichkeit ist. Diese Formalisierung erleichtert oft die Rechnungen. ■

### Beispiel 3.15 (Terminierung eines Algorithmus)

Von einem zufälligen Algorithmus ist bekannt, dass er mit Wahrscheinlichkeit 0.8 terminiert. Wird der Algorithmus 40 Mal aufgerufen, so ist die Anzahl Läufe, bei denen der Algorithmus terminiert, binomialverteilt mit  $n = 40$  und  $p = 0.8$ . ■

### Beispiel 3.16 (Qualitätskontrolle)

Falls in der Situation von Beispiel 3.12 die überprüften Produkte unabhängig voneinander fehlerhaft sind, handelt es sich um ein wiederholtes Bernoulli-Experiment, und somit ist die Anzahl fehlerhafter Produkte gegeben durch eine binomialverteilte Zufallsvariable mit den Parametern  $n$  und  $p$ . ■

### Beispiel 3.17 (Erdős-Rényi Zufallsgraph)

Ein Erdős-Rényi Zufallsgraph **referenz einfügen** besteht aus  $k$  Knoten, sowie einer zufälligen Anzahl von Kanten. Die Knoten sind dabei fest vorgegeben, und für jede mögliche Kante, also für jedes (ungeordnete) Paar von Knoten wird mit einem unabhängigen Bernoulli-Experiment mit Erfolgswahrscheinlichkeit  $p$  entschieden, ob sie im Graph tatsächlich vorhanden ist oder nicht. Bei  $k$  Knoten gibt es  $\binom{k}{2}$  ungeordnete Paare. Für jedes dieser Paare werfen wir nun (hypothetisch) eine Münze, welche mit Wahrscheinlichkeit  $p$  Kopf zeigt. Fällt die Münze auf Kopf, zeichnen wir die Kante zwischen den beiden Knoten ein, fällt sie auf Zahl, fehlt die Kante, siehe Abbildung 3.4.

Wir bezeichnen mit  $X^{(k)}$  die Anzahl vorhandener Kanten in einem Erdős-Rényi Zufallsgraph mit  $k$  Knoten. Das ist also eine Zufallsvariable, mit Wertebereich

$$X^{(k)}(\Omega) = \{0, 1, \dots, \binom{k}{2}\}.$$



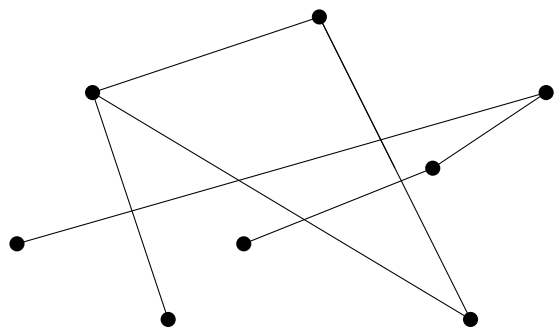


Abb. 3.4: Realisierung eines Erdős-Rényi Zufallsgraphen mit  $n = 8$  Knoten,  $p = 0.3$

Weiter ist  $X^{(k)}$  die Anzahl Erfolge im wiederholten Bernoulli-Experiment mit Erfolgswahrscheinlichkeit  $p$ , somit ist  $X^{(k)}$  binomialverteilt mit Parametern  $p$  und  $n = \binom{k}{2}$ . ■

Die Binomialverteilung tritt also als Verteilung der Anzahl Erfolge im wiederholten Bernoulli-Experiment auf, wenn die Anzahl Wiederholungen vorgegeben ist. Wir können uns aber auch eine andere Frage stellen, nämlich „wie viele Versuche muss ich machen, bis ich zum ersten Mal einen Erfolg sehe?“, oder im Urnenmodell von Beispiel 3.13, „wie viele Kugeln muss ich (mit Zurücklegen, also ohne gegenseitige Beeinflussung der Teilexperimente) aus der Urne ziehen, bis ich zum ersten Mal eine weiße Kugel ziehe?“

**Definition 3.11 (Geometrische Verteilung)**

Sei  $p \in [0, 1]$  Eine Zufallsvariable  $X$  heißt **geometrisch verteilt mit Parameter  $p$**  falls gilt:  $X(\Omega) = \mathbb{N}$ , und

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

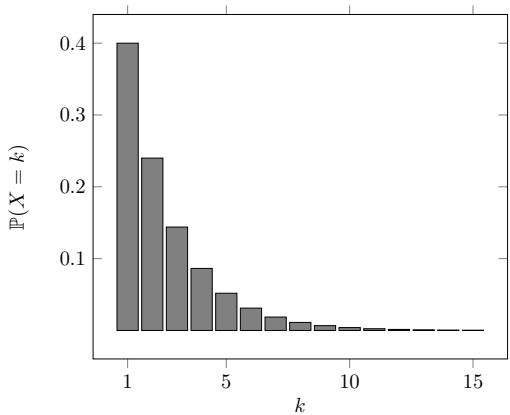


Abb. 3.5: Geometrische Verteilung mit  $p = 0.4$

Eine geometrische Zufallsvariable hat also einen abzählbaren, keinen endlichen Wertebereich. Es handelt sich tatsächlich um eine Verteilung, denn

$$\sum_{k=1}^{\infty} \mathbb{P}(X = k) = \sum_{k=1}^{\infty} p(1-p)^{k-1} = p \sum_{k=0}^{\infty} (1-p)^k = p \cdot \frac{1}{1-(1-p)} = p \cdot \frac{1}{p} = 1.$$

Dabei haben wir beim zweiten Gleichheitszeichen den Faktor  $p$  vor die Summe gezogen und den Summationsindex um eins verschoben, und beim dritten Gleichheitszeichen die aus der Analysis bekannte Formel für eine *geometrische Reihe* benutzt.

### Satz 3.4

*Wir betrachten ein wiederholtes Bernoulli-Experiment mit Erfolgswahrscheinlichkeit  $p$ . Die Ergebnisse der einzelnen Wiederholungen seien dabei unabhängig. Sei  $X$  die Anzahl Versuche, die gemacht werden müssen, bis zum ersten Mal „Erfolg“ eintritt. Dann ist  $X$  geometrisch verteilt mit Parameter  $p$ .*

**Beweis 3.4** Da wir es mit einem wiederholten Bernoulli-Experiment zu tun haben, ist bei jedem Versuch die Wahrscheinlichkeit Erfolg zu haben gleich  $p$ , und die Wahrscheinlichkeit für Misserfolg gleich  $1-p$ . Um genau beim  $k$ -ten Versuch zum ersten Mal Erfolg zu haben, müssen also zuerst  $k-1$  Misserfolge und dann ein Erfolg eintreten. Da die Ereignisse „1. Zug Misserfolg“, „2. Zug Misserfolg“ ... „ $k$ -ter Zug Erfolg“ unabhängig sind, erhalten wir

$$\mathbb{P}(X = k) = \underbrace{(1-p) \cdot \dots \cdot (1-p)}_{k-1 \text{ Mal}} p = (1-p)^{k-1} p.$$

□

In einigen Lehrbüchern wird die geometrische Verteilung leicht anders definiert, indem  $X(\Omega) = \mathbb{N}_0$  gewählt wird, und  $p_X(k) = p(1-p)^k$ . Dies beschreibt die Anzahl Misserfolge vor dem ersten Erfolg. Die beiden Definitionen sind äquivalent, und gehen durch Verschiebung ineinander über.

### Beispiel 3.18 (Terminierung eines Algorithmus)

Bei einem stochastischen Algorithmus, der mit Wahrscheinlichkeit 0.8 terminiert, möchten wir wissen, wie viele Versuche wir machen müssen, bis der Algorithmus zum ersten Mal terminiert. Gemäß Satz 3.4 ist dies eine geometrisch verteilte Zufallsvariable  $X$  mit Parameter  $p = 0.8$ , also ist

$$\mathbb{P}(X = 1) = 0.8, \quad \mathbb{P}(X = 2) = 0.2 \cdot 0.8 = 0.16, \quad \mathbb{P}(X = 3) = (0.2)^2 \cdot 0.8 \approx 0.032.$$

■

**Beispiel 3.19 (Coupon collector)**

Das sogenannte coupon collector problem oder Bildchensammelproblem untersucht folgende Situation: Es gibt  $n$  verschiedene Sammelbilder. Diese werden in Verpackungen verkauft, die jeweils ein zufällig ausgewähltes dieser Bilder enthält. Von Interesse ist zum Beispiel, wie viele Packungen gekauft werden müssen, um jedes Bild mindestens einmal zu haben. Bei der mathematischen Modellierung gehen wir wie folgt vor. Sei  $Y_k$  die Anzahl Packungen, die man kaufen muss, um ein neues Bild zu erhalten, wenn man aktuell  $k$  verschiedene Bilder besitzt. Die  $Y_k$  sind geometrisch verteilt, denn es liegt ein Bernoulli-Experiment zu Grunde: In jeder gekauften Packung liegt entweder ein neues Bild, oder eines, welches man schon hat. Der Parameter hängt dabei von  $k$  ab: Hat man aktuell  $k$  verschiedene Bilder, so ist die Wahrscheinlichkeit, in der nächsten Packung ein neues zu finden, genau

$$p_k := \frac{n - k}{n}.$$

Wir werden diesem Beispiel in Kapitel 5 noch einmal begegnen. ■

### 3.3 Poisson-Verteilung und Approximation

Die Poisson-Verteilung tritt oft in Situationen auf, in denen die Anzahl (seltener) Ereignisse in einem bestimmten Zeitraum untersucht werden. Beispiele dafür sind

- Anzahl Anrufe, die in einer Telefonzentrale pro Minute eingehen
- Anzahl Fahrzeuge, die pro Stunde eine bestimmte Brücke befahren
- Anzahl Mutationen auf einem Stück DNA
- Anzahl Schäden pro Tag in einem Versicherungsunternehmen

**Definition 3.12 (Poisson-Verteilung)**

Sei  $\lambda > 0$ . Eine Zufallsvariable  $X$  heißt **Poisson-verteilt** mit Parameter  $\lambda$ , falls  $X(\Omega) = \mathbb{N}_0$ , und

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

◆

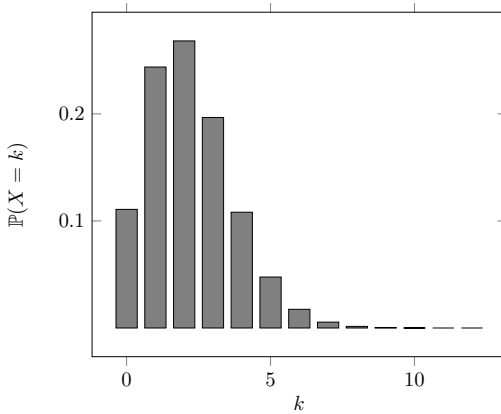
Der Parameter  $\lambda$  bezeichnet die *Rate* mit der die Ereignisse eintreten – je größer  $\lambda$ , desto kürzer die Abstände, d.h. desto mehr Ereignisse pro Zeiteinheit. Der Parameter ist also keine Wahrscheinlichkeit, im Gegensatz zum Parameter  $p$  der geometrischen Verteilung.

Ein Grund für die Bedeutung der Poisson-Verteilung liegt in folgendem Satz:

**Satz 3.5 (Poisson-Grenzwertsatz)**

Sei  $n \in \mathbb{N}$ , und sei  $(p_n)_{n \in \mathbb{N}}$  eine Folge von Zahlen aus  $[0, 1]$  mit  $\lim_{n \rightarrow \infty} n \cdot p_n = \lambda \in ]0, \infty[$ . Sei  $X_n$  binomial verteilt mit Parametern  $n$  und  $p_n$ , und sei  $X$  Poisson-verteilt mit Parameter  $\lambda$ . Dann gilt

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

Abb. 3.6: Poisson-Verteilung mit  $\lambda = 2.2$ 

Für einen Beweis siehe z.B. [?], Satz 3.7. Den Inhalt des Poisson-Grenzwertsatzes kann man folgendermaßen zusammenfassen: Wenn die Anzahl Versuche groß und die Erfolgswahrscheinlichkeit klein ist, so ist die Anzahl Erfolge ungefähr Poisson-verteilt. Deshalb beschreibt die Poisson-Verteilung auch die Anzahl seltener Ereignisse. Für praktische Anwendungen verwendet man statt des Grenzwertsatzes eine Approximation:

**Satz 3.6 (Poisson-Approximation)**

Sei  $\lambda > 0$ . Sei  $X$  binomialverteilte Zufallsvariable mit Parametern  $n$  und  $p = \lambda/n$ . Dann gilt

$$\mathbb{P}(X = k) \approx \frac{e^{-\lambda} \lambda^k}{k!}, \quad k \in \{0, \dots, n\},$$

d.h.  $X$  ist annähernd Poisson-verteilt mit Parameter  $\lambda = np$ . Die Approximation wird besser, je größer  $n$  bzw. je kleiner  $p$  ist.

**Beispiel 3.20 (Erdős-Rényi Zufallsgraph, Fortsetzung von Beispiel 3.17)**

Wir wählen nun im Zufallsgraphen aus Beispiel 3.17 die Wahrscheinlichkeit für die Existenz einer Kante abhängig von der Anzahl der Knoten im Graph. Sei  $k \in \mathbb{N}$  die Anzahl Knoten, und

$$p_k := \frac{2\lambda}{k^2}$$

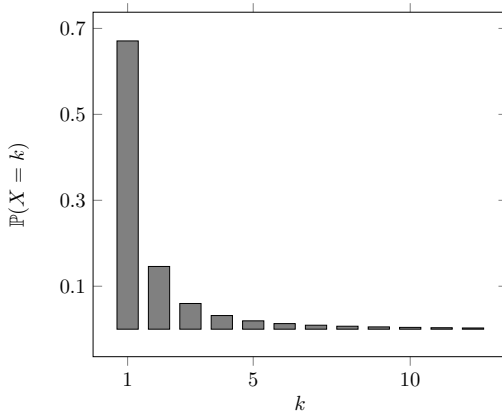
für die Erfolgswahrscheinlichkeit, wobei  $\lambda > 0$  sein soll. Nach Beispiel 3.17 ist die Anzahl Kanten binomialverteilt mit den Parametern  $\binom{k}{2}$  und  $p_k$ . Es gilt

$$\lim_{k \rightarrow \infty} \binom{k}{2} p_k = \lim_{k \rightarrow \infty} \frac{k(k-1)}{2} \cdot \frac{2\lambda}{k^2} = \lambda.$$

Somit ist die Voraussetzung aus dem Poisson-Grenzwertsatz 3.5 erfüllt, und für große  $k$  ist die Zufallsvariable  $X^{(k)}$  ungefähr Poisson-verteilt mit Parameter  $\lambda$ . Es gilt also für alle  $m \in \mathbb{N}$

$$\mathbb{P}(X^{(k)} = m) \approx \frac{e^{-\lambda} \lambda^m}{m!}.$$

■

Abb. 3.7: Zipf-Verteilung mit  $a = 2.2$ 

### 3.4 Zipf-Verteilung und skalenfreie Netzwerke

Als letzte Verteilung in diesem Kapitel präsentieren wir die Zipf-Verteilung, die eine wichtige Rolle unter anderem in der Theorie der skalenfreien Netzwerke spielt.

**Definition 3.13 (Zipf-Verteilung)**

Sei  $a > 1$ . Eine Zufallsvariable  $X$  heißt **Zipf-verteilt (oder Zeta-verteilt)** mit Parameter  $a$  falls gilt:  $X(\Omega) = \mathbb{N}$ , und

$$\mathbb{P}(X = k) = \frac{k^{-a}}{Z(a)},$$

wobei  $Z(a) := \sum_{l=1}^{\infty} l^{-a}$  ist. ◆

Aus der Analysis ist bekannt, dass für  $a > 1$  die unendliche Reihe  $Z(a) = \sum_{k=1}^{\infty} k^{-a}$  konvergiert, also endlich ist. Somit gilt

$$\sum_{k \in X(\Omega)} \mathbb{P}(X = k) = \sum_{k=1}^{\infty} \mathbb{P}(X = k) = \frac{1}{Z(a)} \sum_{k=1}^{\infty} k^{-a} = 1,$$

es handelt sich also tatsächlich um eine Wahrscheinlichkeitsverteilung. Manchmal wird die Zipf-Verteilung auch für  $a \leq 1$  (insbesondere für  $a = 1$ ) definiert. In diesem Fall muss ein maximaler Wert  $N$  spezifiziert werden, also  $X(\Omega) = \{1, \dots, N\}$ , und

$$\mathbb{P}(X = k) = \frac{k^{-a}}{Z_N(a)},$$

mit  $Z_N(a) = \sum_{k=1}^N k^{-a}$ .

**Beispiel 3.21**

Bei einer Zipf-Verteilung mit  $a = 1$  sind die Wahrscheinlichkeitsgewichte gegeben durch

$$\mathbb{P}(X = k) = \frac{1}{k} \cdot \frac{1}{Z_N(1)},$$

wobei  $Z_N(1) = \sum_{k=1}^N \frac{1}{k}$  ist. Somit gilt für eine Zipf(1)-verteilte Zufallsvariable  $X$

$$\mathbb{P}(X = 1) = 2\mathbb{P}(X = 2) = 3\mathbb{P}(X = 3) = \dots = N\mathbb{P}(X = N).$$

Allgemein gilt für eine Zipf-Verteilung mit Parameter  $a$

$$\mathbb{P}(X = 1) = 2^a \mathbb{P}(X = 2) = \dots = k^a \mathbb{P}(X = k), \quad k \in \mathbb{N}.$$

Umgeformt ergibt das

$$\mathbb{P}(X = k) = \frac{1}{k^a} \mathbb{P}(X = 1).$$

Als Funktion von  $k$  fällt die Verteilung von  $X$  also *polynomiell* mit Exponent  $a$ . Beispielsweise wird empirisch beobachtet, dass die Verteilung der Worthäufigkeiten in Sprachen ein solches Verhalten zeigen, welches entsprechend auch „Zipf’sches Gesetz“ genannt wird **referenz einfügen**. ■

### Beispiel 3.22 (Preferential attachment Modell)

Die Zipf-Verteilung tritt unter anderem im sogenannten preferential attachment Modell auf, welches einen zufälligen Graphen erzeugt, der sich in vielerlei Hinsicht vom Erdős-Rényi Zufallsgraph aus Beispiel 3.17 unterscheidet. Eine mögliche Konstruktion läuft nach folgendem Algorithmus ab: Beginne mit  $n = 1$  Knoten. In jedem Schritt wird ein Knoten hinzugefügt, und mit genau einem bereits vorhandenen Knoten verbunden. Dieser Knoten wird mit einer Wahrscheinlichkeit ausgewählt, welche proportional zur Anzahl seiner bereits vorhandenen Verbindungen ist. Sind also aktuell  $n$  Knoten vorhanden, so bezeichnen wir mit  $z_i^{(n)}$  die Anzahl Kanten, die von Knoten Nummer  $i$  zu einem anderen Knoten führen, mit  $i \in \{1, \dots, n\}$ . Sei  $A_i^{(n+1)}$  das Ereignis, dass der  $n + 1$ te Knoten mit Knoten Nummer  $j \in \{1, \dots, n\}$  verbunden wird (und sonst mit keinem weiteren Knoten). Dann muss

$$\mathbb{P}(A_i^{(n+1)}) = \frac{z_j^{(n)}}{z_1^{(n)} + \dots + z_n^{(n)}}$$

gelten. Insbesondere bedeutet diese Konstruktionsvorschrift, dass Knoten mit vielen Verbindungen tendentiell noch mehr Verbindungen hinzubekommen. Bezeichnet man mit  $Z$  die Anzahl Verbindungen eines zufällig ausgewählten Knoten im Netzwerk, so kann bewiesen werden, dass für  $n \rightarrow \infty$  die Zufallsvariable  $X$  annähernd Zipf-verteilt ist. **referenz einfügen** ■

Das beschriebene preferential attachment Modell führt auf ein sogenanntes *skalenfreies Netzwerk*, charakterisiert durch die Eigenschaft, dass es relativ wenige Knoten mit sehr vielen Verbindungen und viele Knoten mit nur wenigen Verbindungen besitzt. Solche Netzwerke spielen eine wichtige Rolle in praktischen Anwendungen. Von vielen echten Netzwerken wird zumindest vermutet, dass sie skalenfrei sind, so z.B. das www, Internet, Soziale Netzwerke, Netzwerke wissenschaftlicher Zusammenarbeit, Stromnetze, und andere. **referenz einfügen**

Tab. 3.2: Wichtige diskrete Verteilungen

Name	Parameter	$X(\Omega)$	$\mathbb{P}(X = k)$	Notation
Gleichverteilung	$n \in \mathbb{N}$	$\{1, \dots, n\}$	$1/n$	$X \sim \text{Unif}(n)$
Bernoulli	$p \in [0, 1]$	$\{0, 1\}$	$\mathbb{P}(X = 1) = p$	$X \sim \text{Ber}(p)$
Binomial	$p \in [0, 1], n \in \mathbb{N}$	$\{0, \dots, n\}$	$\binom{n}{k} p^k (1-p)^{n-k}$	$X \sim \text{Bin}(n, p)$
Geometrisch	$p \in [0, 1]$	$\mathbb{N}$	$(1-p)^{k-1} p$	$X \sim \text{Geo}(p)$
Poisson	$\lambda > 0$	$\mathbb{N}_0$	$\frac{e^{-\lambda} \lambda^k}{k!}$	$X \sim \text{Poi}(\lambda)$
Zipf	$a > 1$	$\mathbb{N}$	$\sum_{k=1}^{\infty} \frac{k^{-a}}{k^{-a}}$	$X \sim \text{Zipf}(a)$

**Zusammenfassung**

Zufallsvariablen sind Abbildungen, welche interessante Größen in Zufallsexperimenten beschreiben. Die relevanten Informationen über Zufallsvariablen sind durch ihre Verteilungsfunktion gegeben, im diskreten Fall auch über die Verteilung. Es gibt eine Reihe von wichtigen Verteilungen, welche in verschiedenen spezifischen Situationen auftreten. Tabelle 3.2 gibt einen Überblick über die speziellen diskreten Verteilungen aus diesem Kapitel.

**Aufgaben**

- 3.1** Beim zweifachen Würfeln wie in Beispiel 3.5 berechne man die Wahrscheinlichkeiten dafür, dass das Minimum der beiden geworfenen Zahlen a) gleich 3, b) mindestens 5, c) ungerade ist.
- 3.2** Wir würfeln mit zwei fairen Würfeln. Sei  $X$  die Augenzahl des ersten und  $Y$  die Augenzahl des zweiten Würfels. Sei  $W := \min\{X, Y\}$  und  $Z := 2^Y$ . Geben Sie die Wertebereiche  $X(\Omega)$ ,  $W(\Omega)$ , und  $Z(\Omega)$  an und berechnen Sie die Verteilungen von  $X$ ,  $W$  und  $Z$ .
- 3.3** In einer Urne liegen 6 rote, 3 weiße und eine schwarze Kugel. Ziehe drei Kugeln mit Zurücklegen. Bezeichne mit  $X$  die Anzahl rote und mit  $Y$  die Anzahl weiße Kugeln, die dabei gezogen wurden. Berechnen Sie die Wahrscheinlichkeiten für die folgenden Ereignisse: a)  $\{X = 2\}$  b)  $\{Y > X\}$  c)  $\{X + Y \geq 4\}$  d)  $\{X \cdot Y = 2\}$
- 3.4** Eine bestimmte Krankheit tritt bei 1.1 Prozent einer Population auf. Was ist die Wahrscheinlichkeit, dass bei 25 zufällig ausgewählten Individuen a) niemand, b) genau zwei, c) mindestens zwei diese Krankheit haben?
- 3.5** Fluggesellschaften rechnen damit, dass 5 Prozent der Passagiere ihren gebuchten Flug nicht antreten. Deshalb überbuchen sie normalerweise die Flüge. Eine Fluggesell-

schaft verkauft 52 Tickets für einen Flug, der 50 Passagiere befördern kann. Wie groß ist die Wahrscheinlichkeit, dass alle Passagiere die zum Flughafen kommen, Platz auf dem Flug bekommen?

**3.6** Ein Kandidat in einer Quiz-Show bekommt zu jeder Frage vier Antwortmöglichkeiten, wovon eine richtig ist. Bei einer falschen Antwort fliegt er raus. Wenn der Kandidat jedesmal zufällig eine Antwort rät, wie wahrscheinlich ist es dann, dass er mindestens bis zur sechsten Frage kommt?

**3.7** Wie in Aufgabe 3.4 nehmen wir an, dass eine Krankheit bei 1.1 Prozent der Bevölkerung auftritt. Benutze den Poisson'schen Grenzwertsatz, um näherungsweise die Wahrscheinlichkeit zu bestimmen, dass höchstens drei von 500 Individuen die Krankheit haben.

**3.8** Auf einem Drucker kommen Druckaufträge Poisson-verteilt mit einer Rate von 2.2 Jobs pro Minute an. Was ist die Wahrscheinlichkeit, dass in einer Minute a) kein Auftrag, b) genau 2 Aufträge, c) mehr als 10 Aufträge ankommen?

**3.9** Es sei  $X$  geometrisch verteilt mit Parameter  $p$ . Zeigen Sie: a) Für  $k \in \mathbb{N}$  gilt  $\mathbb{P}(X \geq k) = \mathbb{P}(X > k-1) = (1-p)^{k-1}$ , b) Für  $k, n \in \mathbb{N}$  gilt  $\mathbb{P}(X > k+n \mid X > k) = \mathbb{P}(X > n)$ .

**3.10** Seien  $a, b > 0$ . Für  $n \in \mathbb{N}$  sei  $X_n$  geometrisch verteilt mit Parameter  $p_n = a/n$ . Beweisen Sie:  $\lim_{n \rightarrow \infty} \mathbb{P}(X_n \geq bn) = e^{-ab}$ .



# 4 Gemeinsame Verteilung, Unabhängigkeit von Zufallsvariablen

---

## Übersicht

4.1	Gemeinsame Verteilung, bedingte Verteilung .....	52
4.2	Unabhängigkeit von Zufallsvariablen .....	55

---

In komplexeren Situationen wird es oft nicht möglich sein, einen ganzen Sachverhalt mit nur einer Zufallsvariablen zu beschreiben. Es werden möglicherweise zwei oder mehr Zufallsvariablen benötigt, die sich unter Umständen gegenseitig beeinflussen. In diesem Kapitel wird gezeigt, wie zwei oder mehr Zufallsvariablen mit Hilfe ihrer gemeinsamen Verteilung beschrieben werden, und wie man “gegenseitigen stochastischen Einfluss” von zwei Zufallsvariablen quantifizieren kann.

### Lernziele

- Gemeinsame Verteilungen kennen und berechnen können
- Bedingte Verteilungen und Randverteilungen berechnen können
- Zufallsvariablen auf Unabhängigkeit überprüfen können und die Implikationen kennen.

### Vorkenntnisse

Inhalt der vorhergehenden Kapitel, insbesondere Verteilungen.

## 4.1 Gemeinsame Verteilung, bedingte Verteilung

In Kapitel 3 wurden Zufallsvariablen und deren Verteilung definiert. Betrachtet man mehrere Zufallsvariablen gleichzeitig, so muss dieser Begriff erweitert werden.

### Definition 4.1 (Gemeinsame Verteilung)

Seien  $X$  und  $Y$  zwei diskrete Zufallsvariablen, welche auf einem gemeinsamen Wahrscheinlichkeitsraum  $(\Omega, \mathbb{P})$  definiert sind. Die **gemeinsame Verteilung** von  $X$  und  $Y$  ist gegeben durch die Kollektion der Wahrscheinlichkeiten

$$\mathbb{P}(\{X = x\} \cap \{Y = y\}), \quad x \in X(\Omega), y \in Y(\Omega).$$

Wir schreiben auch kurz  $\mathbb{P}(X = x, Y = y) := \mathbb{P}(\{X = x\} \cap \{Y = y\}), x \in X(\Omega), y \in Y(\Omega)$ , das Komma steht also für den Durchschnitt.  $\blacklozenge$

Eine gemeinsame Verteilung von zwei Zufallsvariablen kann somit auch als eine Funktion von zwei Variablen aufgefasst werden, welche für jedes Paar von Werten  $x \in X(\Omega), y \in Y(\Omega)$  die Wahrscheinlichkeit dafür angibt, dass die Zufallsvariable  $X$  den Wert  $x$  und gleichzeitig die Zufallsvariable  $Y$  den Wert  $y$  annimmt. Beispiele von Zufallsvariablen, bei denen die gemeinsame Verteilung von Interesse sein kann, sind Größe und Gewicht eines Menschen, Anzahl Bestellungen und Bearbeitungszeit in einem Logistikzentrum, oder Anzahl und Höhe von Schäden in einem Versicherungsunternehmen, usw.

Die gemeinsame Verteilung von mehr als zwei Zufallsvariablen wird analog definiert: Seien  $X_1, \dots, X_n$  diskrete Zufallsvariablen. Ihre gemeinsame Verteilung ist gegeben durch

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) \quad x_1 \in X_1(\Omega), \dots, x_n \in X_n(\Omega).$$

Wiederum steht hier das Komma zwischen den einzelnen Ereignisse für den Durchschnitt, beziehungsweise für das Wort „und“.

### Beispiel 4.1 (Urnenmodell)

In einer Urne befinden sich 4 Kugeln, numeriert von 1 bis 4. Es werden zwei Kugeln mit Zurücklegen gezogen. Sei  $X$  die Summe der beiden Zahlen, und  $Y$  das Minimum der beiden gezogenen Zahlen. Was ist die gemeinsame Verteilung von  $X$  und  $Y$ ?

Zuerst müssen wir die Wertebereiche bestimmen.  $X(\Omega)$  ist die Menge aller Zahlen, die als Summe zweier Zahlen zwischen 1 und 4 dargestellt werden können, also  $X(\Omega) = \{2, \dots, 8\}$ . Für  $Y(\Omega)$  finden wir  $Y(\Omega) = \{1, \dots, 4\}$ . Dabei ist es sinnvoll, den Wahrscheinlichkeitsraum

$$\Omega = \{1, 2, 3, 4\}^2$$

zugrunde zu legen, d.h.  $\omega \in \Omega$  hat die Form  $\omega = (\omega_1, \omega_2)$ , mit  $\omega_1, \omega_2 \in \{1, \dots, 4\}$ . Dies ist ein Laplace-Raum mit  $\mathbb{P}(\omega) = \frac{1}{16}$  für jedes  $\omega \in \Omega$ . Wir können formal schreiben

$$X(\omega) = \omega_1 + \omega_2, \text{ und } Y(\omega) = \min\{\omega_1, \omega_2\}.$$

Nun interessieren wir uns für die Wahrscheinlichkeiten  $\mathbb{P}(X = x, Y = y)$ . Diese müssen wir nun für *alle* möglichen Werte  $x \in X(\Omega)$  und  $y \in Y(\Omega)$  bestimmen, z.B. also für  $x = 2$  und  $y = 1$ . Wie kann dieses Ereignis realisiert werden? Die Summe ist genau dann 2, wenn zweimal die Kugel 1 gezogen wurde, d.h. wenn  $\omega = (1, 1)$  ist. In diesem Fall ist automatisch  $Y = 1$ , also ist  $\{X = 2, Y = 1\} = \{(1, 1)\}$ . Somit erhalten wir

$$\mathbb{P}(X = 2, Y = 1) = \mathbb{P}((1, 1)) = \frac{1}{16}.$$

Für  $x = 3, y = 1$  gibt es zwei Möglichkeiten: Erste Kugel 1 und zweite 2, oder umgekehrt, also  $\{X = 3, Y = 1\} = \{(1, 2), (2, 1)\}$ .

$$\mathbb{P}(X = 3, Y = 1) = \mathbb{P}((1, 2), (2, 1)) = \frac{2}{16} = \frac{1}{8}.$$

Andererseits ist  $x = 3, y = 3$  unmöglich zu erreichen, denn wenn die Summe gleich 3 ist, kann keine gezogene Zahl größer als 2 sein, also insbesondere kann das Minimum nicht gleich 3 sein. Somit ist

$$\mathbb{P}(X = 3, Y = 3) = 0.$$

Analog kann man die anderen Wahrscheinlichkeiten ausrechnen. Eine kompakte Darstellung der gemeinsamen Verteilung ist in Tabelle 4.1 gegeben. ■

### Definition 4.2 (Randverteilung)

Sei die gemeinsame Verteilung von  $X$  und  $Y$  gegeben. Die **Randverteilungen** von  $X$  bzw. von  $Y$  sind die Verteilungen der einzelnen Zufallsvariablen  $X$  bzw.  $Y$ , also

$$\mathbb{P}(X = x) = \sum_{y \in Y(\Omega)} \mathbb{P}(X = x, Y = y),$$

$$\mathbb{P}(Y = y) = \sum_{x \in X(\Omega)} \mathbb{P}(X = x, Y = y).$$

◆

### Beispiel 4.2 (Urnenmodell, Fortsetzung von Beispiel 4.1)

Die Tabelle der gemeinsamen Verteilung wird um die Randverteilungen ergänzt, indem man unten eine Zeile und rechts eine Spalte einfügt, und dort die jeweiligen Randverteilungen einträgt, siehe Tabelle 4.1. Die Zahlen ergeben sich als Summe der Werte oberhalb bzw. rechts davon. Die 1 in der unteren linken Ecke kann als Kontrollwert verstanden werden, sie ergibt sich durch Summation sowohl der Spalte ganz rechts als auch der Zeile ganz unten. ■

### Definition 4.3 (Bedingte Verteilung)

Seien  $X, Y$  zwei Zufallsvariablen welche auf demselben Wahrscheinlichkeitsraum  $(\Omega, \mathbb{P})$  definiert sind. Sei  $y \in Y(\Omega)$ . Die **bedingte Verteilung** von  $X$  gegeben  $Y$  ist definiert als

$$\mathbb{P}(X = x | Y = y), \quad x \in X(\Omega), y \in Y(\Omega).$$

$\begin{array}{c} X \\ \backslash \\ Y \end{array}$	2	3	4	5	6	7	8	$\mathbb{P}(Y = \cdot)$
1	1/16	1/8	1/8	1/8	0	0	0	7/16
2	0	0	1/16	1/8	1/8	0	0	5/16
3	0	0	0	0	1/16	1/8	0	3/16
4	0	0	0	0	0	0	1/16	1/16
$\mathbb{P}(X = \cdot)$	1/16	1/8	3/16	1/4	3/16	1/8	1/16	1

Tab. 4.1: Gemeinsame Verteilung mit Randverteilung

Für einen *festen* Wert von  $y \in Y(\Omega)$  ist die *bedingte Verteilung von  $X$  gegeben  $Y = y$*  definiert als

$$\mathbb{P}(X = x | Y = y), \quad x \in X(\Omega).$$



Nach Definition der bedingten Wahrscheinlichkeit und der gemeinsamen Verteilung kann man die bedingte Verteilung folgendermaßen berechnen:

$$\mathbb{P}(X = x | Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}.$$

Man beachte dabei, dass die Rollen von  $X$  und  $Y$  verschieden sind, der Ausdruck ist nicht symmetrisch, d.h. Vertauschen von  $X$  und  $Y$  führt auf unterschiedliche Ergebnisse!

#### Beispiel 4.3 (Urnenmodell, Fortsetzung von Beispiel 4.1)

Betrachte eine Urne mit 4 Kugeln, numeriert von 1 bis 4. Wir ziehen zwei Kugeln mit Zurücklegen. Sei wieder  $X$  die Summe der beiden Zahlen, und  $Y$  das Minimum der beiden gezogenen Zahlen. Gesucht ist die bedingte Verteilung von  $Y$  gegeben  $X$ .

Betrachte erst die bedingte Verteilung von  $Y$  gegeben  $X = 2$ . Aus der Tabelle in 4.2 sehen wir:  $\mathbb{P}(X = 2) = \frac{1}{16}$ , und  $\mathbb{P}(X = 2, Y = 1) = \frac{1}{16}$ ,  $\mathbb{P}(X = 2, Y = 2) = \mathbb{P}(X = 2, Y = 3) = \mathbb{P}(X = 2, Y = 4) = 0$ . Somit erhalten wir

$$\mathbb{P}(Y = 1 | X = 2) = \frac{1/16}{1/16} = 1,$$

$$\mathbb{P}(Y = 2 | X = 2) = \mathbb{P}(Y = 3 | X = 2) = \mathbb{P}(Y = 4 | X = 2) = \frac{0}{1/16} = 0.$$

Analog berechnet sich  $\mathbb{P}(Y = y | X = 3), \mathbb{P}(Y = y | X = 4), \dots, \mathbb{P}(Y = y | X = 8), y \in \{1, \dots, 4\}$ . Wiederum bietet sich eine tabellarische Zusammenfassung der Werte an, siehe Tabelle 4.2. Die bedingte Verteilung von  $Y$  gegeben  $X = 4$  ist dabei z.B. die Spalte zum Wert  $X = 4$ .



$Y \backslash X$	2	3	4	5	6	7	8
1	1	1	$2/3$	$1/2$	0	0	0
2	0	0	$1/3$	$1/2$	$2/3$	0	0
3	0	0	0	0	$1/3$	1	0
4	0	0	0	0	0	0	1

Tab. 4.2: Bedingte Verteilung  $\mathbb{P}(Y|X)$  von  $Y$  gegeben  $X$ 

## 4.2 Unabhängigkeit von Zufallsvariablen

In Kapitel 2 wurden zwei *Ereignisse*  $A$  und  $B$  *unabhängig* genannt, falls  $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$  gilt. Äquivalent dazu gelten  $\mathbb{P}(A|B) = \mathbb{P}(A)$  bzw.  $\mathbb{P}(B|A) = \mathbb{P}(B)$ , die Wahrscheinlichkeit für das eine Ereignis wird also nicht durch das Eintreten (oder Nicht-Eintreten) des anderen beeinflusst.

Nun definieren wir den analogen Begriff für *Zufallsvariablen*. Obwohl die beiden Begriffe viel miteinander zu tun haben, ist es wichtig, sich die Unterschiede vor Augen zu führen

### Definition 4.4 (Unabhängigkeit von Zufallsvariablen)

Zwei Zufallsvariablen  $X$  und  $Y$  definiert auf demselben Wahrscheinlichkeitsraum heißen **unabhängig**, falls für alle  $x \in X(\Omega), y \in Y(\Omega)$  die *Ereignisse*  $\{X = x\}$  und  $\{Y = y\}$  unabhängig sind.  $\blacklozenge$

Daraus folgt unmittelbar mit der Definition der Unabhängigkeit von Ereignissen:

### Satz 4.1

Zwei Zufallsvariablen  $X$  und  $Y$  sind genau dann unabhängig, wenn für alle  $x \in X(\Omega), y \in Y(\Omega)$  gilt

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y).$$

### Beispiel 4.4 (Urnenmodell, Fortsetzung von Beispiel 4.2)

Sind  $X$  und  $Y$  aus dem Urnen-Beispiel 4.1 unabhängig? Aus Tabelle 4.1 sehen wir beispielsweise gleich in der ersten Zeile, dass  $\mathbb{P}(X = 2, Y = 1) \neq \mathbb{P}(X = 2)\mathbb{P}(Y = 1)$ . Also können die Zufallsvariablen *nicht* unabhängig sein. Bei unabhängigen Zufallsvariablen müsste *jeder Eintrag im Inneren der Tabelle* gleich dem Produkt der beiden zugehörigen Randwahrscheinlichkeiten sein.

Man kann sich auch anschaulich klar machen, dass  $X$  und  $Y$  nicht unabhängig sein können, da beispielsweise die Information  $X = 8$  zwingend festlegt, dass  $Y = 4$  sein muss. Aus der Kenntnis von  $X$  lassen sich also Rückschlüsse auf  $Y$  ziehen, und umgekehrt.  $\blacksquare$

$\begin{array}{c} X \\ \backslash \\ Y \end{array}$	0	1	2	$\mathbb{P}(Y = \cdot)$
0	0.12	0.18	0.3	0.6
1	0.08	0.12	0.2	0.4
$\mathbb{P}(X = \cdot)$	0.2	0.3	0.5	1

Tab. 4.3: Verteilung aus Beispiel 4.5

**Beispiel 4.5 (Unabhängigkeit)**

Wir betrachten Zufallsvariablen  $X$  und  $Y$ , deren gemeinsame Verteilung in Tabelle 4.3 gegeben ist. Wie man direkt nachrechnen kann, gilt für *jede* Wahrscheinlichkeit im Inneren der Tabelle, dass sie das Produkt der zugehörigen Randwahrscheinlichkeiten ist, also

$$\mathbb{P}(X = 0, Y = 0) = 0.12 = 0.2 \cdot 0.6 = \mathbb{P}(X = 0)\mathbb{P}(Y = 0),$$

$$\mathbb{P}(X = 1, Y = 0) = 0.18 = 0.3 \cdot 0.6 = \mathbb{P}(X = 1)\mathbb{P}(Y = 0),$$

usw. Somit sind die Zufallsvariablen  $X$  und  $Y$  in diesem Beispiel unabhängig. ■

**Beispiel 4.6 (Erdős-Rényi-Zufallsgraph)**

Der Zufallsgraphen aus Beispiel 3.17 kann nun etwas formaler konstruiert werden. Sei  $k$  die Anzahl Knoten, und seien  $Y_1, \dots, Y_{\binom{k}{2}}$  unabhängige Bernoulli-verteilte Zufallsvariablen mit Erfolgswahrscheinlichkeit  $p$ . Das Ereignis  $\{Y_l = 1\}$  wird dann als “Kante Nr.  $l$  ist vorhanden” interpretiert. Sei nun  $X^{(k)}$  wie bisher die (zufällige) Anzahl der im Graph vorhandenen Kanten. Somit können wir nun schreiben

$$X^{(k)} = \sum_{l=1}^{\binom{k}{2}} Y_l.$$

Wir wissen bereits, dass  $X^{(k)}$  binomialverteilt ist. Betrachten wir zusätzlich die Zufallsvariable  $Z$  definiert als die Anzahl Knoten im Graphen, welche mit keinem anderen Knoten verbunden sind. Sind  $X^{(k)}$  und  $Z$  unabhängig? Die Intuition sagt nein, denn weiß man beispielsweise, dass  $X^{(k)}$  groß ist, so sind viele Kanten vorhanden, und entsprechend dürfte es nur wenige Knoten geben, die keine Verbindung haben. Im Extremfall ist  $X^{(k)} = \binom{k}{2}$ . Dieses Ereignis hat Wahrscheinlichkeit  $p^{\binom{k}{2}}$ . In diesem Fall sind also alle Kanten vorhanden, und somit muss  $Z = 0$  sein, denn es kann keinen Knoten ohne Verbindung geben. Somit gilt

$$\mathbb{P}\left(X^{(k)} = \binom{k}{2}, Z = 0\right) = p^{\binom{k}{2}}.$$

Jedoch gilt

$$\mathbb{P}\left(X^{(k)} = \binom{k}{2}\right) \cdot \mathbb{P}(Z = 0) \neq p^{\binom{k}{2}},$$

denn wir haben  $\mathbb{P}\left(X^{(k)} = \binom{k}{2}\right) = p^{\binom{k}{2}}$ , aber  $\mathbb{P}(Z = 0) < 1$ , da es durchaus Situationen gibt, in denen ein Knoten ohne Verbindung vorkommt. Somit gilt

$$\mathbb{P}\left(X^{(k)} = \binom{k}{2}, Z = 0\right) \neq \mathbb{P}\left(X^{(k)} = \binom{k}{2}\right) \cdot \mathbb{P}(Z = 0).$$

Wir haben also ein Ereignis gefunden, für welches die gemeinsame Wahrscheinlichkeit nicht gleich dem Produkt der Einzelwahrscheinlichkeit ist, und somit können  $X^{(k)}$  und  $Z$  nicht unabhängig sein. ■

Die Darstellung von  $X^{(k)}$  als Summe von unabhängigen Bernoulli-verteilten Zufallsvariablen im Beispiel 4.6 gilt allgemein für binomialverteilte Zufallsvariablen. Dies folgt aus Satz 3.3, welcher nun noch etwas knapper und präziser formuliert werden kann.

### Satz 4.2

Seien  $Y_1, \dots, Y_n$  unabhängige Bernoulli-verteilte Zufallsvariablen mit Erfolgswahrscheinlichkeit  $p$ . Dann ist  $X := \sum_{i=1}^n Y_i$  binomialverteilt mit Parametern  $n$  und  $p$ .

Wir wissen aus Beispiel 3.4, dass Funktionen von Zufallsvariablen wieder Zufallsvariablen sind. Die Unabhängigkeit bleibt dabei erhalten.

### Satz 4.3 (Funktionen von unabhängigen Zufallsvariablen)

Seien  $X$  und  $Y$  unabhängige (diskrete) Zufallsvariablen, und seien  $f : X(\Omega) \rightarrow \mathbb{R}$  und  $g : Y(\Omega) \rightarrow \mathbb{R}$  zwei Funktionen. Dann sind die Zufallsvariablen  $f(X)$  und  $g(Y)$  ebenfalls unabhängig.

Wie zu erwarten, kann man Unabhängigkeit auch von drei oder mehr Zufallsvariablen untersuchen. Dabei muss man jedoch etwas Vorsicht walten lassen.

### Definition 4.5 (Unabhängigkeit von mehr als zwei Zufallsvariablen)

Seien  $X_1, X_2, \dots, X_n$  diskrete Zufallsvariablen auf  $(\Omega, \mathbb{P})$ . Diese heißen **unabhängig**, falls

$$\mathbb{P}(X_{i_1} = a_{i_1}, \dots, X_{i_k} = a_{i_k}) = \mathbb{P}(X_{i_1} = a_{i_1}) \cdot \dots \cdot \mathbb{P}(X_{i_k} = a_{i_k})$$

für alle  $k \leq n$ , für alle paarweise verschiedenen  $i_1, \dots, i_k \in \{1, \dots, n\}$ , und für alle  $a_{i_1} \in X_{i_1}(\Omega), \dots, a_{i_k} \in X_{i_k}(\Omega)$  gilt. ♦

Eine analoge Definition gilt für Folgen  $(X_n)_{n \in \mathbb{N}}$  von Zufallsvariablen.

Wenn zwei Zufallsvariablen unabhängig sind, kann man eine Formel für die Verteilung ihrer Summe herleiten.

### Satz 4.4 (Faltungsformel)

Seien  $X$  und  $Y$  zwei unabhängige diskrete Zufallsvariablen auf  $(\Omega, \mathbb{P})$ . Dann hat die Zufallsvariable  $X + Y$  die Verteilung

$$\mathbb{P}(X + Y = k) = \sum_{m \in X(\Omega)} \mathbb{P}(X = m) \cdot \mathbb{P}(Y = k - m).$$

Man kann sich relativ leicht davon überzeugen, dass der Satz ohne die Voraussetzung der Unabhängigkeit nicht stimmt, vgl. z.B. Aufgabe 4.6.

**Beweis 4.1** Beachte dass  $(X + Y)(\Omega) = \{m + l : m \in X(\Omega), l \in Y(\Omega)\}$  gilt. Wir haben für  $k \in (X + Y)(\Omega)$ ,

$$\begin{aligned}\mathbb{P}(X + Y = k) &= \sum_{m \in X(\Omega)} \mathbb{P}(X = m, X + Y = k) \\ &= \sum_{m \in X(\Omega)} \mathbb{P}(X = m, Y = k - m) \\ &= \sum_{m \in X(\Omega)} \mathbb{P}(X = m) \mathbb{P}(Y = k - m).\end{aligned}$$

Dabei haben wir im letzten Schritt die Unabhängigkeit verwendet. □

#### Beispiel 4.7 (Summe von Poisson-Zufallsvariablen)

Seien  $X$  und  $Y$  unabhängige Zufallsvariablen,  $X$  Poisson-verteilt mit Parameter  $\lambda$ ,  $Y$  Poisson-verteilt mit Parameter  $\mu$ . Was ist die Verteilung der Summe? Wir benutzen die Faltungsformel, und erhalten

$$\begin{aligned}\mathbb{P}(X + Y = k) &= \sum_{m=0}^{\infty} \mathbb{P}(X = m) \mathbb{P}(Y = k - m) \\ &= \sum_{m=0}^k \frac{\lambda^m e^{-\lambda}}{m!} \cdot \frac{\mu^{k-m} e^{-\mu}}{(k-m)!} \\ &= e^{-(\lambda+\mu)} \sum_{m=0}^k \lambda^m \mu^{k-m} \frac{\binom{k}{m}}{k!} \\ &= e^{-(\lambda+\mu)} \frac{(\lambda + \mu)^k}{k!},\end{aligned}$$

wobei wir im zweitletzten Schritt  $\binom{k}{m} = \frac{k!}{m!(k-m)!}$  und im letzten Schritt den binomischen Lehrsatz  $(\lambda + \mu)^k = \sum_{m=0}^k \binom{k}{m} \lambda^m \mu^{k-m}$  benutzt haben.

Die rechte Seite in obiger Rechnung ist nun wieder die Formel für die Poisson-Verteilung, mit Parameter  $\lambda + \mu$ . Wir haben somit eine wichtige Eigenschaft der Poisson-Verteilung hergeleitet: Die Summe zweier unabhängiger Poisson-verteilter Zufallsvariablen ist wiederum Poisson-verteilt, und der Parameter ist durch die Summe der Parameter gegeben. ■

#### Zusammenfassung

Die gemeinsame Verteilung von Zufallsvariablen gibt die Wahrscheinlichkeiten für sämtliche Paare von Werten an. Sie kann in Tabellen dargestellt werden. Mit Hilfe der Tabelle kann unter anderem überprüft werden, ob die Zufallsvariablen unabhängig sind. Für unabhängige Zufallsvariablen kann man beispielsweise die Verteilung der Summe mit Hilfe der Faltungsformel berechnen.



$\begin{array}{c} X \\ \backslash \\ Y \end{array}$	-1	0	1	$\mathbb{P}(Y = \cdot)$
-1	1/20	3/20	0	1/5
0	1/4	1/4	1/10	3/5
1	1/20	0	3/20	1/5
$\mathbb{P}(X = \cdot)$	7/20	2/5	1/4	

Tab. 4.4: Zu Aufgabe 4.3

## Aufgaben

**4.1** Beim Würfeln mit einem fairen Würfel bezeichne  $X$  die geworfene Zahl. Man werfe anschließend  $X$  faire Münzen, und bezeichnen mit  $Y$  die Anzahl dabei geworfener Köpfe.

a) Man gebe die gemeinsame Verteilung von  $X$  und  $Y$  an, b) Berechne  $\mathbb{P}(X = 2)$ , c) Berechne  $\mathbb{P}(Y = 5)$ .

**4.2** In der Situation von Aufgabe 4.1 bestimme man die bedingte Verteilung von  $X$  gegeben  $Y$ .

**4.3** Die gemeinsame Verteilung von  $X$  und  $Y$  sowie die Randverteilungen sind in der Tabelle 4.4 gegeben. a) Man bestimme die bedingte Verteilung von  $Y$  gegeben  $X = -1$  b) Bestimme die gemeinsame Verteilung von  $X^2$  und  $Y^2$  c) sind  $X^2$  und  $Y^2$  unabhängig?

**4.4** Es sei  $X$  eine Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathbb{P})$ , welche das Ergebnis beim einmaligen Werfen eines fairen Würfels beschreibt. Es sei  $Y$  die Zufallsvariable auf demselben Wahrscheinlichkeitsraum, welche gegeben ist durch

$$Y(\omega) = \begin{cases} X(\omega) & \text{falls } X(\omega) \text{ gerade ist,} \\ 1 & \text{sonst.} \end{cases}$$

Sei außerdem  $Z := 1_{\{X \leq 4\}}$ . Man gebe die gemeinsame Verteilung von  $Y$  und  $Z$  an.

**4.5** Es sei  $X$  eine Zufallsvariable mit  $\mathbb{P}(X = -1) = \mathbb{P}(X = 1) = 1/4$ ,  $\mathbb{P}(X = 0) = 1/2$ . Man berechne die bedingte Verteilung von  $X$  gegeben  $X^2$ .

**4.6** a) Es seien  $X, Y$  zwei unabhängige Zufallsvariablen mit Wertebereich  $X(\Omega) = Y(\Omega)$ . Zeige: Für alle  $k \in X(\Omega)$  gilt  $\mathbb{P}(X+Y = 2k) \geq \mathbb{P}(X = k)\mathbb{P}(Y = k)$ . b) Es sei nun  $Z = -X$ . Man zeige, dass in diesem Fall  $\mathbb{P}(X + Z = 2k) \geq \mathbb{P}(X = k)\mathbb{P}(Z = k)$  nicht gilt. c) Man zeige in der Situation von b), dass auch die Aussage von Satz 4.4 nicht gilt.

**4.7** Es bezeichne  $W$  das Ergebnis beim einmaligen Würfeln. Für  $W = k$  werde eine unfaire Münze mit Wahrscheinlichkeit  $1/k$  für Kopf zweimal geworfen.  $X$  bezeichne bei diesem Münzwurf das Ergebnis des ersten Wurfs,  $Y$  das Ergebnis des zweiten. Man zeige:  $X$  und  $Y$  sind abhängig, jedoch unabhängig gegeben  $W$ , d.h.  $\mathbb{P}(X = x, Y = y | W = k) = \mathbb{P}(X = x | W = k)\mathbb{P}(Y = y | W = k)$  für alle  $x \in X(\Omega)$ ,  $Y = Y(\Omega)$ ,  $k \in W(\Omega)$ .



# 5 Kenngrößen für Zufallsvariablen

---

## Übersicht

5.1	Erwartungswert .....	62
5.2	Varianz .....	67
5.3	Markov- und Chebyshev-Ungleichung .....	70
5.4	Kovarianz und Korrelation .....	71

---

Im letzten Kapitel wurde gezeigt, dass die *Verteilung* im Prinzip alle Informationen über die zugehörigen Zufallsvariablen enthält. Jedoch ist es nicht immer einfach und teilweise aufwändig, die Verteilung komplett zu bestimmen. Nicht immer sind auf den ersten Blick die für ein konkretes Problem relevanten Angaben erkennbar. Ziel dieses Kapitels ist es deshalb, einige *Kenngrößen* zu identifizieren, die bereits zentrale Informationen über eine Zufallsvariable enthalten, und die in vielen Fällen einfach zu berechnen sind. Konkret wird der Erwartungswert und die Varianz von Zufallsvariablen eingeführt, sowie Kovarianz und Korrelation als Kenngrößen für den Zusammenhang zweier Zufallsvariablen.

### Lernziele

- Die Begriffe Erwartungswert, Varianz, Standardabweichung kennen
- Erwartungswert und Varianz für diskrete Verteilungen berechnen können
- Informationen über die Verteilung daraus herleiten können
- Kovarianz und Korrelation berechnen können, und wissen, welche Informationen daraus erhalten werden können.

### Vorkenntnisse

Neben dem bisher gelernten Stoff sowie den dabei benötigten Grundlagen werden in diesem Kapitel insbesondere mehrfach Grenzwerte von unendlichen Reihen berechnet.

## 5.1 Erwartungswert

Wir werfen einen fairen Würfel 100 Mal, und notieren die Ergebnisse. Sei  $y_i$  das Ergebnis des  $i$ -ten Wurfs. Wie groß ist dann der Mittelwert über alle Würfe, d.h.

$$\frac{1}{100} \sum_{i=1}^{100} y_i?$$

Da die Ergebnisse  $\{1, \dots, 6\}$  gleich wahrscheinlich sind, erwarten wir im Durchschnitt jedes der 6 möglichen Ergebnisse gleich oft. Das Mittel über viele (hier 100) Versuche sollte damit ungefähr gleich dem arithmetischen Mittel (Durchschnitt) der möglichen Werte sein, also

$$\frac{1}{100} \sum_{i=1}^{100} y_i \approx \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = \frac{21}{6} = 3.5.$$

Führt man dieses Experiment mit einem echten Würfel oder mit einem geeigneten Programm am Rechner durchführen, wird man feststellen, dass tatsächlich eine Zahl in der Nähe von 3.5 dabei herauskommt.

Diese Beobachtung ist ein Beispiel für das sogenannte „Gesetz der großen Zahlen“, welches wir später in der Vorlesung kennen lernen werden, und welches besagt, dass das Mittel über viele Ergebnisse eines wiederholten Zufallsexperiments ungefähr gleich dem *Erwartungswert* ist. Im obigen Beispiel ist der Erwartungswert genau das arithmetische Mittel der möglichen Ergebnisse. Allgemein ist er folgendermaßen definiert.

### Definition 5.1 (Erwartungswert einer diskreten Zufallsvariablen)

Sei  $X$  eine diskrete Zufallsvariable. Der **Erwartungswert** von  $X$  ist definiert als

$$\mathbb{E}[X] := \sum_{k \in X(\Omega)} k \cdot \mathbb{P}(X = k),$$

sofern  $\sum_{k \in X(\Omega)} |k| \cdot \mathbb{P}(X = k) < \infty$  ist. Wenn letzters erfüllt ist, so sagen wir, dass der **Erwartungswert existiert**.  $\blacklozenge$

Falls der Wertebereich  $X(\Omega)$  endlich ist, d.h. falls die Zufallsvariable  $X$  nur endlich viele Werte annehmen kann (z.B. beim Würfeln), ist die Bedingung  $\sum_{k \in X(\Omega)} |k| \cdot \mathbb{P}(X = k) < \infty$  stets erfüllt. Ist jedoch der Wertebereich unendlich, so kann es vorkommen, dass  $\sum_{k \in X(\Omega)} |k| \cdot \mathbb{P}(X = k) = \infty$  ist. In diesem Fall ist der Erwartungswert nicht definiert, man sagt *der Erwartungswert existiert nicht*. Diese Situation kann tatsächlich vorkommen, wie wir noch sehen werden.

### Beispiel 5.1 (Würfeln)

$X$  bezeichne die Augenzahl beim Werfen eines fairen Würfels. Es gilt also  $X(\Omega) = \{1, 2, \dots, 6\}$ , und  $\mathbb{P}(X = x) = \frac{1}{6}$  für  $x \in \{1, \dots, 6\}$ . Dann erhalten wir aus der Definition des Erwartungswerts

$$\mathbb{E}[X] = \sum_{k \in X(\Omega)} k \mathbb{P}(X = k) = \sum_{k=1}^6 k \cdot \mathbb{P}(X = k) = \sum_{k=1}^6 k \cdot \frac{1}{6} = \frac{1}{6} (1 + 2 + \dots + 6) = 3.5.$$

Das ist genau der Wert, den wir zu Beginn dieses Abschnitts bereits erraten hatten. Man beachte, dass in einem einzelnen Wurf dieser Wert überhaupt nicht angenommen werden kann. Er ist lediglich das, was im Mittel über viele Würfe erwartet wird. ■

Im Beispiel mit dem Würfel sind alle Ergebnisse gleich wahrscheinlich. Dies ist im Allgemeinen nicht so.

### Beispiel 5.2

Sei  $X$  eine Zufallsvariable mit  $X(\Omega) = \{0, \dots, 3\}$  und Verteilung gegeben durch die Tabelle

$k$	0	1	2	3
$\mathbb{P}(X = k)$	0.2	0.1	0.5	0.2

So erhalten wir aus der Definition

$$\mathbb{E}[X] = \sum_{k \in X(\Omega)} k \cdot \mathbb{P}(X = k) = \sum_{k=0}^3 k \cdot \mathbb{P}(X = k) = 0 \cdot 0.2 + 1 \cdot 0.1 + 2 \cdot 0.5 + 3 \cdot 0.2 = 1.7.$$

■

### Beispiel 5.3 (Geometrische Verteilung)

Sei  $X$  eine Zufallsvariable mit geometrischer Verteilung zum Parameter  $p$ . Wie kann der Erwartungswert  $\mathbb{E}[X]$  berechnet werden? Da  $X(\Omega) = \mathbb{N}$  und  $\mathbb{P}(X = k) = p(1-p)^{k-1}$  ist, erhalten wir aus der Definition des Erwartungswerts

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k \cdot p \cdot (1-p)^{k-1} = p \sum_{k=0}^{\infty} k(1-p)^{k-1}.$$

Um die unendliche Summe zu berechnen, gibt es einen Trick. Aus der Analysis (geometrische Reihe) ist bekannt, dass

$$\sum_{k=0}^{\infty} (1-p)^k = \frac{1}{1-(1-p)} = \frac{1}{p}$$

ist. Leitet man diese Gleichung links und rechts nach  $p$  ab und setzt gleich, so erhält man

$$\sum_{k=0}^{\infty} k(1-p)^{k-1}(-1) = -\frac{1}{p^2}, \quad (*)$$

wobei einerseits die Kettenregel verwendet wurde, und andererseits die Tatsache, dass bei einer Potenzreihe innerhalb des Konvergenzradius die Ableitung mit der Summation vertauscht werden können. Die linke Seite dieser Gleichung  $(*)$  ist fast (bis auf das Vorzeichen und den konstanten Vorfaktor  $p$ ) die rechte Seite der obigen Gleichung für  $\mathbb{E}[X]$ . Zusammengefasst erhält man

$$\mathbb{E}[X] = p \sum_{k=0}^{\infty} k(1-p)^{k-1} = -p \sum_{k=0}^{\infty} k(1-p)^{k-1}(-1) \stackrel{(*)}{=} (-p) \cdot \left(-\frac{1}{p^2}\right) = \frac{1}{p}.$$

Also ist der Erwartungswert der geometrischen Verteilung mit Parameter  $p$  genau  $\frac{1}{p}$ . Ist also etwa  $p = 1/2$ , ist der entsprechende Erwartungswert 2. Für kleine  $p$  muss man also im Mittel länger auf den ersten Erfolg warten, was nicht überrascht. ■

#### Beispiel 5.4 (Zipf-Verteilung)

Bei der Zipf-Verteilung gilt  $X(\Omega) = \mathbb{N}$  und  $\mathbb{P}(X = k) = \frac{k^{-a}}{Z(a)}$ , wobei  $Z(a)$  so gewählt wird, dass es sich um eine Wahrscheinlichkeitsverteilung handelt, also  $Z(a) = \sum_{k=1}^{\infty} k^{-a}$ , was für  $a > 1$  definiert ist. Für den Erwartungswert einer Zipf-verteilten Zufallsvariablen  $X$  erhält man

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k \cdot \frac{k^{-a}}{Z(a)} = \frac{1}{Z(a)} \sum_{k=1}^{\infty} k^{1-a}.$$

Aus der Analysis ist bekannt, dass diese Reihe genau dann konvergiert, wenn  $1 - a < -1$  ist, also wenn  $a > 2$  ist. In diesem Fall gilt

$$\mathbb{E}[X] = \frac{1}{Z(a)} \sum_{k=1}^{\infty} k^{1-a} = \frac{Z(a-1)}{Z(a)}.$$

In anderen Worten, der Erwartungswert der Zipf-Verteilung existiert nur, wenn der Parameter  $a > 2$  ist, und nicht für  $1 < a < 2$ . ■

#### Beispiel 5.5 (Bedingter Erwartungswert)

Es sei  $X$  eine Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathbb{P})$ , und  $A$  ein Ereignis mit  $\mathbb{P}(A) > 0$ . Wir wissen, dass dann die bedingte Wahrscheinlichkeit  $\mathbb{P}(\cdot | A)$  auch ein Wahrscheinlichkeitsmaß ist, und somit können wir den *bedingten Erwartungswert* von  $X$  gegeben  $A$  definieren als Erwartungswert bezüglich  $\mathbb{P}(\cdot | A)$ , also

$$\mathbb{E}[X | A] := \sum_{k \in X(\Omega)} k \cdot \mathbb{P}(X = k | A).$$

Ist beispielsweise  $X$  das Ergebnis eines fairen Würfels, und  $A$  das Ereignis, dass das Ergebnis gerade ist, so gilt  $\mathbb{P}(X = 2 | A) = \mathbb{P}(X = 4 | A) = \mathbb{P}(X = 6 | A) = \frac{1}{3}$ , und somit

$$\mathbb{E}[X | A] = \frac{1}{3}(2 + 4 + 6) = 4.$$

■

Der Erwartungswert hat eine Reihe wichtiger Eigenschaften.

#### Satz 5.1 (Eigenschaften des Erwartungswertes)

Seien  $X$  und  $Y$  diskrete Zufallsvariablen auf  $(\Omega, \mathbb{P})$ . Dann gelten

- (a)  $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- (b)  $\mathbb{E}[a \cdot X] = a \cdot \mathbb{E}[X]$  für jedes  $a \in \mathbb{R}$
- (c) Falls  $X(\omega) \geq 0$  für alle  $\omega$ , so ist  $\mathbb{E}[X] \geq 0$ .
- (d) Falls  $X$  konstant ist, d.h.  $X(\omega) = c \in \mathbb{R}$  für alle  $\omega$ , dann ist  $\mathbb{E}[X] = c$ .

(e) Sei  $f : \mathbb{R} \rightarrow \mathbb{R}$  eine Funktion. Dann gilt

$$\mathbb{E}[f(X)] = \sum_{x \in X(\Omega)} f(x) \mathbb{P}(X = x).$$

(f) Falls  $X$  und  $Y$  unabhängig sind, so gilt

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

**Beweis 5.1** (a) Mit Hilfe der Faltungsformel aus Satz 4.4 folgt

$$\begin{aligned} \mathbb{E}[X + Y] &= \sum_{z \in (X+Y)(\Omega)} z \cdot \mathbb{P}(X + Y = z) = \sum_{z \in (X+Y)(\Omega)} z \sum_{x \in X(\Omega)} \mathbb{P}(X = x) \mathbb{P}(Y = z - x) \\ &= \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} (x + y) \mathbb{P}(X = x) \mathbb{P}(Y = y) = \sum_{x \in X(\Omega)} x \mathbb{P}(X = x) \sum_{y \in Y(\Omega)} y \mathbb{P}(Y = y) \\ &= \mathbb{E}[X] + \mathbb{E}[Y]. \end{aligned}$$

(b)-(e) folgen direkt aus der Definition.

(f) Für den Wertebereich von  $XY$  gilt  $(XY)(\Omega) = \{xy : x \in X(\Omega), y \in Y(\Omega)\}$ . Man erhält damit

$$\mathbb{E}[XY] = \sum_{z \in (XY)(\Omega)} z \cdot \mathbb{P}(XY = z) = \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} xy \cdot \mathbb{P}(X = x, Y = y).$$

Wegen der Unabhängigkeit gilt  $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \mathbb{P}(Y = y)$ , und somit

$$\mathbb{E}[XY] = \sum_{x \in X(\Omega)} x \mathbb{P}(X = x) \sum_{y \in Y(\Omega)} y \mathbb{P}(Y = y) = \mathbb{E}[X] \mathbb{E}[Y].$$

□

Mit Hilfe dieses Satzes kann beispielsweise der Erwartungswert der Binomialverteilung einfach berechnet werden.

### Beispiel 5.6 (Erwartungswert von Bernoulli- und Binomialverteilung)

Wir beginnen mit dem Erwartungswert der Bernoulli-Verteilung, und verwenden dann die Beziehung zwischen Binomialverteilung und Bernoulli-Verteilung. Sei  $Y$  Bernoulli-verteilt zum Parameter  $p$ . Dann gilt

$$\mathbb{E}[Y] = 0 \cdot \mathbb{P}(Y = 0) + 1 \cdot \mathbb{P}(Y = 1) = 0 + 1 \cdot p = p.$$

Sei nun  $X$  binomialverteilt mit Parametern  $n$  und  $p$ . Dank Satz 3.3 können wir  $X$  schreiben als

$$X = \sum_{i=1}^n Y_i,$$

wenn  $Y_i, i = 1, \dots, n$  unabhängige, Bernoulli-verteilte Zufallsvariable mit Parameter  $p$  sind. Nach Satz 5.1 (a) gilt dann

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^n Y_i\right] = \sum_{i=1}^n \mathbb{E}[Y_i] = \sum_{i=1}^n p = np.$$

Der Erwartungswert einer binomialverteilten Zufallsvariablen mit Parametern  $n$  und  $p$  ist also genau  $np$ , d.h. beim  $n$ -fach wiederholten Bernoulli-Experiment mit Erfolgswahrscheinlichkeit  $p$  sehen wir *im Mittel*  $np$  Erfolge. ■

### Beispiel 5.7 (Erdős-Renyi-Zufallsgraph)

Im Erdős-Renyi Zufallsgraph aus Beispiel 3.17 wissen wir, dass die Anzahl vorhandener Kanten binomialverteilt mit Parametern  $\binom{k}{2}, p$  ist. Somit ist die erwartete Anzahl vorhandener Kanten gleich  $p \cdot \binom{k}{2}$ . ■

### Beispiel 5.8 (Coupon collector)

Betrachte die Situation aus Beispiel 3.19: Es gibt  $n$  verschiedene Sammelbilder. Diese werden in Verpackungen verkauft, die jeweils ein zufällig ausgewähltes dieser Bilder enthält. Wie viele Packungen muss man im Mittel kaufen, um jedes Bild mindestens einmal zu haben?

Bezeichne mit  $X$  die zufällige Anzahl Packungen, die gekauft werden müssen. Um den Erwartungswert von  $X$  zu bestimmen, nutzen wir die Hilfsvariablen aus Beispiel 3.19. Sei  $Y_k$  die Anzahl Packungen, die man kaufen muss, um ein neues Bild zu erhalten, wenn man aktuell  $k$  verschiedene Bilder besitzt. Damit gilt offensichtlich

$$X = \sum_{k=0}^{n-1} Y_k.$$

In Beispiel 3.19 wurde gezeigt, dass die  $Y_k$  sind geometrisch verteilt mit Parameter  $p_k := \frac{n-k}{n}$  abhängig von  $k$  sind. Nach Beispiel 5.3 ist

$$\mathbb{E}[Y_k] = \frac{1}{p_k} = \frac{n}{n-k}.$$

Mit der Linearität des Erwartungswerts aus Satz 5.1 und erhält man

$$\mathbb{E}[X] = \sum_{k=0}^{n-1} \mathbb{E}[Y_k] = \sum_{k=0}^{n-1} \frac{n}{n-k} = n \sum_{l=1}^n \frac{1}{l}.$$

Aus der Analysis ist bekannt, dass  $\sum_{l=1}^n \frac{1}{l} \approx \int_1^n \frac{1}{t} dt = \log n$  ist. Somit ist

$$\mathbb{E}[X] \approx n \log n,$$

es müssen also normalerweise, insbesondere für große  $n$ , deutlich mehr als  $n$  Packungen gekauft werden, um alle Bilder zu erhalten. ■

Die Definition des Erwartungswerts kann leicht umgeschrieben werden, um das folgende Ergebnis zu erhalten:

### Satz 5.2

*Falls  $\Omega$  höchstens abzählbar ist und der Erwartungswert von  $X$  existiert, so gilt*

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \mathbb{P}(\omega).$$



**Beweis 5.2** Dies folgt durch Einsetzen in die Definition des Erwartungswerts und einige Umformungen.  $\square$

In den meisten Situationen ist es einfacher, mit der Definition des Erwartungswertes zu arbeiten als mit diesem Satz, da man üblicherweise die Verteilung kennt. Für Beweise ist aber manchmal Satz 5.2 nützlich.

## 5.2 Varianz

Wir betrachten ein einfaches Glücksspiel, wobei eine Münze geworfen wird. Fällt “Kopf”, so bekommt Spieler A einen Euro, fällt “Zahl”, so muss er einen bezahlen. In jeder Spielrunde ist also die Wahrscheinlichkeit für einen Gewinn gleich  $\frac{1}{2}$ . Sei  $X$  die Gewinnsumme eines Spielers nach 5 Spielen. Offensichtlich gilt für den Erwartungswert  $\mathbb{E}[X] = 0$ .

Nun wird der Einsatz auf 1000 Euro pro Spielrunde erhöht. Natürlich bleibt der Erwartungswert dabei gleich. Trotzdem ändert sich etwas – viele Menschen würden möglicherweise beim ersten Spiel mitspielen, bei der zweiten Variante aber dankend verzichten. Was ist der Grund? Im zweiten Fall kann ein Spieler bis zu 5000 Euro verlieren (allerdings auch gewinnen). In anderen Worten, es gibt eine höhere *Streuung*.

Dieses Beispiel zeigt, dass der Erwartungswert durchaus nicht alle wichtigen Informationen über eine Zufallsvariable enthält. Für eine genauere Beschreibung benötigen wir ein Maß für die Streuung.

### Definition 5.2 (Varianz)

Sei  $X$  eine Zufallsvariable. Die **Varianz** von  $X$  ist definiert als

$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2],$$

falls dieser Erwartungswert existiert, in diesem Fall sagen wir dass die **Varianz existiert**.  $\blacklozenge$

Diese Größe kann man gemäß Satz 5.1 ausrechnen.

### Beispiel 5.9 (Varianz der Bernoulli-Verteilung)

Sei  $X$  Bernoulli-verteilt mit Parameter  $p$ . Dann gilt  $\mathbb{E}[X] = p$ , und mit Satz 5.1 (e) folgt

$$\mathbb{V}(X) = \mathbb{E}[(X - p)^2] = (0 - p)^2(1 - p) + (1 - p)^2p = p(1 - p).$$

■

### Satz 5.3

Sei  $X$  eine Zufallsvariable, deren Varianz existiert. Dann gilt

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

**Beweis 5.3** Wir formen die Definition der Varianz durch Ausmultiplizieren um, und verwenden die Linearität des Erwartungswerts. Damit erhalten wir

$$\begin{aligned}\mathbb{V}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2.\end{aligned}$$

□

Je nach Situation kann diese Formel einfacher anzuwenden sein als die Definition der Varianz. Dabei ist es wichtig, die Reihenfolge von Quadrieren und Erwartungswertbildung zu beachten. Man kann außerdem zeigen, dass die Varianz genau dann existiert, wenn  $\mathbb{E}[X^2]$  existiert.

**Definition 5.3 (Standardabweichung)**

Sei  $X$  eine Zufallsvariable, deren Varianz existiert. Dann heißt

$$\sigma(X) := \sqrt{\mathbb{V}(X)}$$

**Standardabweichung** von  $X$ . Dabei ist die positive Quadratwurzel gemeint. ♦

Die Standardabweichung enthält also im Prinzip dieselbe Information wie die Varianz, es wird lediglich die Wurzel gezogen. In manchen Situationen, insbesondere in der Statistik, ist die Standardabweichung eine etwas natürlichere Größe zur Quantifizierung der Streuung.

**Beispiel 5.10**

Sei  $X$  eine Zufallsvariable mit  $X(\Omega) = \{-1, 1\}$ ,  $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{2}$ . Dann gilt  $\mathbb{E}[X] = 0$ , und

$$\mathbb{E}[X^2] = (-1)^2 \cdot \mathbb{P}(X = -1) + 1^2 \cdot \mathbb{P}(X = 1) = 1.$$

Somit ist  $\mathbb{V}(X) = \sigma(X) = 1$ . Diese Rechnung werden wir später noch benutzen. ■

Auch die Varianz hat einige wichtige theoretische Eigenschaften.

**Satz 5.4 (Eigenschaften der Varianz)**

Seien  $X$  und  $Y$  zwei Zufallsvariablen deren Varianzen existieren.

*](a)] Für  $a, b \in \mathbb{R}$  gilt*

$$\mathbb{V}(aX + b) = a^2\mathbb{V}(X),$$

*unabhängig von  $b$ .*

**(b)** Falls  $X$  und  $Y$  unabhängig sind, so gilt

$$\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y).$$

**(c)** Falls  $X$  konstant ist, so gilt  $\mathbb{V}(X) = 0$ .

**Beweis 5.4 (a)** Mit Satz 5.3, der ersten binomischen Formel und der Linearität des Erwartungswerts (Satz 5.1) folgt

$$\begin{aligned}\mathbb{V}(aX + b) &= \mathbb{E}[(aX + b)^2] - \mathbb{E}[aX + b]^2 = \mathbb{E}[a^2X^2 + 2abX + b^2] - (a\mathbb{E}[X] + b)^2 \\ &= a^2\mathbb{E}[X^2] + 2ab\mathbb{E}[X] + b^2 - a^2\mathbb{E}[X]^2 - 2ab\mathbb{E}[X] - b^2 = a^2\mathbb{E}[X^2] - a^2\mathbb{E}[X]^2 \\ &= a^2\mathbb{V}(X).\end{aligned}$$

**(b)** kann man durch Einsetzen in die Definition und Verwendung von Satz 5.1 beweisen, die Aussage folgt aber auch als Spezialfall von Satz 5.9, den wir später beweisen werden.

**(c)** folgt direkt durch Einsetzen in die Definition der Varianz.  $\square$

Beachte: Falls  $X$  und  $Y$  nicht unabhängig sind, so gilt im allgemeinen *nicht*  $\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y)$ . Im Abschnitt über Kovarianzen wird eine allgemeinere Aussage bewiesen, die stets gilt.

### Beispiel 5.11 (Varianz der Binomialverteilung)

Um die Varianz einer binomialverteilten Zufallsvariablen  $X$  mit Parametern  $n$  und  $p$  zu berechnen, ist es wiederum nützlich, diese als  $X = \sum_{i=1}^n Y_i$  zu schreiben, wobei die  $Y_i$  unabhängig und Bernoulli-verteilt sind. Dann gilt wegen der Unabhängigkeit, Satz 5.4 und Beispiel 5.9

$$\mathbb{V}(X) = \mathbb{V}\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n \mathbb{V}(Y_i) = \sum_{i=1}^n p(1-p) = np(1-p).$$

■

### Beispiel 5.12 (Glücksspiel)

Wir berechnen nun die Varianz im Glücksspiel aus dem Beispiel am Anfang des Abschnittes, für die beiden Fälle mit einem Euro bzw. mit 1000 Euro Einsatz. Zur Vereinfachung der Berechnung verwenden wir Zufallsvariablen  $Y_i$ , um den Gewinn im  $i$ ten Spiel zu bezeichnen, also  $Y_i = 1$  falls das  $i$ te Spiel gewonnen wurde, und  $Y_i = -1$  falls das  $i$ te Spiel verloren wurde. Somit ist bei einem Einsatz von einem Euro pro Spiel der Gewinn nach 5 Spielen gleich  $X = Y_1 + \dots + Y_5$ , und da die  $Y_i$  unabhängig sind, gilt  $\mathbb{V}(X) = \mathbb{V}(Y_1) + \dots + \mathbb{V}(Y_5)$  nach Satz 5.4 (b). Aus Beispiel 5.10 wissen wir, dass  $\mathbb{V}(Y_1) = \dots = \mathbb{V}(Y_5) = 1$  ist. Somit folgt  $\mathbb{V}(X) = 5$ .

Beträgt der Einsatz pro Spiel hingegen 1000 Euro, so ist der Gewinn  $\tilde{X}$  nach 5 Spielen gegeben durch  $\tilde{X} = 1000 \cdot X_1 + \dots + 1000 \cdot X_5 = 1000 \cdot X$ . Also ist

$$\mathbb{V}(\tilde{X}) = \mathbb{V}(1000 \cdot X) = 1000^2 \cdot \mathbb{V}(X) = 5 \cdot 10^6,$$

unter Verwendung von Satz 5.4 (a). Die Varianz ist also extrem viel höher im zweiten Fall, weil die Streuung des Gewinns viel höher ist.  $\blacksquare$

Tabelle 5.1 fasst die Erwartungswerte und Varianzen einiger wichtiger diskreter Verteilungen zusammen. Einige dieser Formeln wurden in Beispielen hergeleitet, andere finden sich in den Übungsaufgaben.

Verteilung	Parameter	Erwartungswert	Varianz
Bernoulli	$p$	$p$	$p(1 - p)$
Binomial	$n, p$	$np$	$np(1 - p)$
Geometrisch	$p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	$\lambda$	$\lambda$	$\lambda$
Zipf	$a > 1$	$\frac{Z(a-1)}{Z(a)}, a > 2, Z(a) = \sum_{k=1}^{\infty} k^{-a}$	$\frac{Z(a)Z(a-2)-Z(a-1)^2}{Z(a)^2}, a > 3$

Tab. 5.1: Erwartungswert und Varianz einiger wichtiger Verteilungen

### 5.3 Markov- und Chebyshev-Ungleichung

Wir haben die Varianz eingeführt, um ein Maß für die *Streuung* zu bekommen, und auch in Beispiel 5.12 gesehen, dass tatsächlich eine breite Streuung durch einen großen Wert der Varianz angezeigt wird. Ist die Varianz groß, so kann die Zufallsvariable stark von ihrem Erwartungswert abweichen. Für diesen bisher etwas ungenau formulierten Zusammenhang kann man konkretere Abschätzungen angeben.

**Satz 5.5 (Chebyshev-Ungleichung)**

*Sei  $X$  eine Zufallsvariable deren Varianz existiert. Dann gilt für jedes  $\varepsilon > 0$*

$$\mathbb{P}(|X - \mathbb{E}[X]| > \varepsilon) \leq \frac{\mathbb{V}(X)}{\varepsilon^2}.$$

In Worten ausgedrückt ist also die Wahrscheinlichkeit, dass eine Zufallsvariable um mehr als den Wert  $\varepsilon$  vom Erwartungswert abweicht (nach oben oder unten), höchstens so groß wie die Varianz dividiert durch  $\varepsilon^2$ .

**Beweis 5.5** Wir beweisen diesen Satz hier für den Fall dass  $X$  eine diskrete Zufallsvariable ist, und auf einem diskreten Wahrscheinlichkeitsraum definiert ist. Er gilt aber allgemeiner, wobei der Beweis ganz ähnlich läuft, nur wurden die Methoden dafür bisher noch nicht eingeführt. In unserem Fall gilt mit Definition 5.2 und Satz 5.2

$$\begin{aligned} \mathbb{V}(X) &= \sum_{\omega \in \Omega} (X(\omega) - \mathbb{E}[X])^2 \mathbb{P}(\omega) \geq \sum_{\omega \in \Omega: |X(\omega) - \mathbb{E}[X]| > \varepsilon} (X(\omega) - \mathbb{E}[X])^2 \mathbb{P}(\omega) \\ &\geq \sum_{\omega \in \Omega: |X(\omega) - \mathbb{E}[X]| > \varepsilon} \varepsilon^2 \mathbb{P}(\omega) \geq \varepsilon^2 \mathbb{P}(|X(\omega) - \mathbb{E}[X]| > \varepsilon). \end{aligned}$$

Auflösen der obigen Ungleichung nach  $\mathbb{P}(|X(\omega) - \mathbb{E}[X]| > \varepsilon)$  liefert das Ergebnis.  $\square$

**Beispiel 5.13 (Coupon collector)**

Betrachte wieder das Bildchensammelproblem aus den Beispielen 3.19 und 5.8. Wieder ist die Zufallsvariable  $X = \sum_{k=0}^{n-1} Y_k$  von Interesse, wobei die  $Y_k$  unabhängige geometrisch verteilte Zufallsvariablen mit Parametern  $p_k = \frac{n-k}{n}$  sind,  $k = 0, \dots, n-1$ . Aus Beispiel 5.8

ist bekannt, dass  $\mathbb{E}[X] \approx n \log n$  gilt. Für die Varianz erhält man dank der Unabhängigkeit aus Satz 5.4 und der Formel aus Tabelle 5.1 für die geometrische Verteilung

$$\begin{aligned} \mathbb{V}(X) &= \sum_{k=0}^{n-1} \mathbb{V}(Y_k) = \sum_{k=0}^{n-1} \frac{1-p_k}{p_k^2} = \sum_{k=0}^{n-1} \frac{1-(n-k)/n}{(n-k)^2/n^2} \\ &= \sum_{k=0}^{n-1} \frac{nk}{(n-k)^2} \leq n^2 \sum_{k=0}^{n-1} \frac{1}{(n-k)^2} = n^2 \sum_{l=1}^n \frac{1}{l^2}. \end{aligned}$$

Aus der Analysis ist bekannt, dass  $\sum_{l=1}^n \frac{1}{l^2}$  für  $n \rightarrow \infty$  gegen  $\pi^2/6$  konvergiert, somit ist

$$\mathbb{V}(X) \leq \frac{\pi^2}{6} n^2.$$

Interessiert man sich nun beispielsweise für die Wahrscheinlichkeit, mehr als das Doppelte der erwarteten Anzahl Packungen kaufen zu müssen, um die Bildersammlung zu vervollständigen, also für  $\mathbb{P}(X \geq 2\mathbb{E}[X])$ , so erhält man (da  $X \geq n$  auf jeden Fall gelten muss)

$$\mathbb{P}(X \geq 2\mathbb{E}[X]) = \mathbb{P}(X - \mathbb{E}[X] \geq \mathbb{E}[X]) = \mathbb{P}(|X - \mathbb{E}[X]| \geq n \log n),$$

und durch Anwendung der Chebyshev-Ungleichung

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq n \log n) \leq \frac{\mathbb{V}(X)}{(n \log n)^2} \leq \frac{\pi^2}{6(\log n)^2},$$

was für  $n \rightarrow \infty$  gegen 0 konvergiert. Man muss also mit relativ hoher Wahrscheinlichkeit nicht mehr als  $2n \log n$  Packungen kaufen. ■

Die Markov-Ungleichung ist eine Verallgemeinerung der Chebyshev-Ungleichung. Der Beweis kann analog zum Beweis der Chebyshev-Ungleichung geführt werden. Alternativ kann die Chebyshev-Ungleichung aus der Markov-Ungleichung hergeleitet werden.

### Satz 5.6 (Markov-Ungleichung)

Sei  $X$  eine Zufallsvariable, und sei  $f : \mathbb{R} \rightarrow [0, \infty)$  eine monoton wachsende Funktion. Falls  $\mathbb{E}[f(X)]$  existiert, so gilt

$$\mathbb{P}(X > a) \leq \frac{\mathbb{E}[f(X)]}{f(a)}.$$

## 5.4 Kovarianz und Korrelation

Wir haben bis jetzt zwei wichtige Kenngrößen für Zufallsvariablen kennengelernt: Der Erwartungswert gibt den „Mittelwert“ der Zufallsvariablen an, und die Varianz ist ein Maß dafür, wie stark die Zufallsvariable um dieses Mittel herum streut. Nun betrachtet man in vielen Situationen nicht nur eine Zufallsvariable, sondern mehrere, welche eine gemeinsame Verteilung besitzen. Wir haben in diesem Zusammenhang bereits den Begriff der Unabhängigkeit kennen gelernt. Wenn zwei Zufallsvariablen nicht unabhängig sind, so möchte man eine Aussage über den Grad der Abhängigkeit treffen. Dafür führen wir nun einen weiteren Begriff ein.

**Definition 5.4 (Kovarianz)**

Seien  $X$  und  $Y$  zwei Zufallsvariablen auf  $(\Omega, \mathbb{P})$ . Die **Kovarianz** von  $X$  und  $Y$  ist definiert als

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

falls der Erwartungswert auf der rechten Seite existiert. Es werden drei Fälle unterschieden:

- Ist  $\text{cov}(X, Y) > 0$ , so heißen  $X$  und  $Y$  *positiv korreliert*.
- Ist  $\text{cov}(X, Y) < 0$ , so heißen  $X$  und  $Y$  *negativ korreliert*.
- Ist  $\text{cov}(X, Y) = 0$ , so heißen  $X$  und  $Y$  *unkorreliert*.



Sind  $X$  und  $Y$  positiv korreliert, also  $\text{cov}(X, Y) > 0$ , so bedeutet das nach der Definition  $\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] > 0$ . Produkte sind genau dann positiv, wenn beide Faktoren das gleiche Vorzeichen haben. Somit bedeutet eine positive Kovarianz, dass die beiden Zufallsvariablen  $X$  und  $Y$  im Mittel entweder beide nach oben oder beide nach unten vom Erwartungswert abweichen. Beispielsweise sind Größe und Gewicht eines Menschen positiv korreliert: Im Mittel wird ein großer Mensch auch mehr wiegen als ein kleiner (aber natürlich nicht immer). Ist umgekehrt  $\text{cov}(X, Y) < 0$ , so bedeutet das, dass im Mittel einer der beiden Werte  $X$  bzw.  $Y$  über und der andere unter dem Erwartungswert liegt. Beispielsweise sind Außentemperatur und Heizkosten negativ korreliert: Je tiefer die Temperatur, desto höher im Allgemeinen die Heizkosten, und umgekehrt. Somit gibt die Kovarianz in der Tat in eine gewisse Information über die Abhängigkeit von  $X$  und  $Y$ .

Wie bei der Varianz gibt es eine alternative Möglichkeit, die Kovarianz zu berechnen, die man durch einfache Umformung der Definition und Ausnutzung der Eigenschaften des Erwartungswerts erhält.

**Satz 5.7**

Seien  $X$  und  $Y$  Zufallsvariablen deren Kovarianz existiert. Dann gilt

$$\text{cov}(X, Y) = \mathbb{E}[X \cdot Y] - \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

**Beweis 5.7** Einsetzen in die Definition, Ausmultiplizieren und Zusammenfassen unter Verwendung von Satz 5.1 ergibt

$$\begin{aligned} \text{cov}(X, Y) &= \mathbb{E}(X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) = \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \end{aligned}$$

□

**Beispiel 5.14**

Betrachte zwei Zufallsvariablen  $X$  und  $Y$ , deren gemeinsame Verteilung von  $X$  und  $Y$  als Tabelle gegeben ist:

$\begin{array}{c} X \\ \backslash \\ Y \end{array}$	0	1	2	$\mathbb{P}(Y = y)$
0	1/16	1/4	1/8	7/16
1	3/16	1/4	1/8	9/16
$\mathbb{P}(X = x)$	1/4	1/2	1/4	1

Man möchte die Kovarianz dieser Zufallsvariablen bestimmen. Verwende Satz 5.7 um  $\text{cov}(X, Y) = \mathbb{E}[X \cdot Y] - \mathbb{E}[X]\mathbb{E}[Y]$  zu schreiben. Nach Definition des Erwartungswerts gilt  $\mathbb{E}[X] = \sum_{x \in X(\Omega)} x\mathbb{P}(X = x) = 1$  und analog  $\mathbb{E}[Y] = \frac{9}{16}$ . Zur Berechnung von  $\mathbb{E}[XY]$  muss erst der Wertebereich von  $XY$  bestimmt werden: Da  $X$  die Werte 0,1,2 und  $Y$  die Werte 0,1 annehmen kann, kann also  $X \cdot Y$  die Werte 0,1,2 annehmen. Es gilt also

$$\mathbb{E}[XY] = 0 \cdot \mathbb{P}(XY = 0) + \mathbb{P}(XY = 1) + 2 \cdot \mathbb{P}(XY = 2).$$

Aus der Tabelle sehen wir  $\mathbb{P}(XY = 1) = \mathbb{P}(X = 1, Y = 1) = \frac{1}{4}$ , und  $\mathbb{P}(XY = 2) = \mathbb{P}(X = 2, Y = 1) = \frac{1}{8}$ .  $\mathbb{P}(XY = 0)$  braucht nicht berechnet zu werden, da diese Zahl bei der Berechnung des Erwartungswerts ohnehin mit 0 multipliziert wird. Zusammengenommen erhält man  $\mathbb{E}[XY] = \mathbb{P}(XY = 1) + 2 \cdot \mathbb{P}(XY = 2) = \frac{1}{2}$ . Damit folgt aus Satz 5.7

$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \frac{8}{16} - \frac{9}{16} = -\frac{1}{16}.$$

Diese beiden Zufallsvariablen sind also negativ korreliert. ■

Wir geben nun einige Eigenschaften der Kovarianz an, welche relativ direkt aus der Definition hergeleitet werden können.

**Satz 5.8 (Eigenschaften der Kovarianz)**

Seien  $X, Y, Z$  Zufallsvariablen auf  $(\Omega, \mathbb{P})$ , deren Kovarianz existiert, und seien  $a, b \in \mathbb{R}$ . Dann gelten

- (a)  $\text{cov}(X, Y) = \text{cov}(Y, X)$ ,
- (b)  $\text{cov}(X, X) = \text{var}(X)$ ,
- (c)  $\text{cov}(aX + bY, Z) = a \cdot \text{cov}(X, Z) + b \cdot \text{cov}(Y, Z)$ ,
- (d) Falls  $X$  und  $Y$  unabhängig sind, so gilt  $\text{cov}(X, Y) = 0$ .

**Beweis 5.8** (a) und (b) folgen direkt aus der Definition, (c) folgt aus der Definition durch Ausmultiplizieren und Anwenden der Linearität des Erwartungswertes. Für den Beweis von (d) verwenden wir, dass laut Satz 5.1 (f) für unabhängige Zufallsvariablen  $X$  und  $Y$  gilt, dass dann  $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$  ist. Somit erhalten wir

$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0.$$

□

$\begin{array}{c} X \\ Y \end{array}$	-1	0	1	$\mathbb{P}(Y = y)$
0	1/4	0	1/4	1/2
1	0	1/2	0	1/4
$\mathbb{P}(X = x)$	1/4	1/2	1/4	1

Tab. 5.2: Gemeinsame Verteilung von  $X$  und  $Y$ **Beispiel 5.15 (Unabhängigkeit und Unkorreliertheit)**

Punkt (d) aus obigem Satz besagt, dass unabhängige Zufallsvariablen auch unkorreliert sind. Die Umkehrung gilt nicht, wie man an einem Beispiel einsehen kann. Seien  $X, Y$  Zufallsvariablen mit gemeinsamer Verteilung aus Tabelle 5.2. Dann gilt  $\mathbb{E}[X] = 0$  und  $\text{cov}(XY) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$ , aber  $\mathbb{P}(X = 0, Y = 0) = 0 \neq \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(X = 0)\mathbb{P}(Y = 0)$ . Somit sind  $X$  und  $Y$  zwar unkorreliert, aber nicht unabhängig. ■

Wir wissen bereits, dass für unabhängige Zufallsvariable die Varianz der Summe gleich der Summe der Varianzen ist. Allgemeiner gilt der folgende wichtige Satz.

**Satz 5.9 (Varianz der Summe zweier Zufallsvariablen)**

Seien  $X$  und  $Y$  Zufallsvariablen auf  $(\Omega, \mathbb{P})$ . Dann gilt

$$\mathbb{V}(X + Y) = \mathbb{V}(X) + 2\text{cov}(X, Y) + \mathbb{V}(Y).$$

**Beweis 5.9** Aus Satz 5.3, der Linearität des Erwartungswert, und der ersten binomischen Formel folgt

$$\begin{aligned}
 \mathbb{V}(X + Y) &= \mathbb{E}[(X + Y)^2] - (\mathbb{E}[X + Y])^2 \\
 &= \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\
 &= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - (\mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y]^2) \\
 &= \mathbb{V}(X) + 2\text{cov}(X, Y) + \mathbb{V}(Y).
 \end{aligned}$$

□

Aus diesem Satz und aus Satz 5.7 (d) folgt nun noch einmal Satz 5.3 (b).

**Beispiel 5.16 (Erdős-Rényi-Zufallsgraph)**

Betrachte einmal mehr einen Zufallsgraphen mit  $k$  Knoten und Erfolgswahrscheinlichkeit  $p$ . Seien zusätzlich zwei feste Knoten 1 und 2 ausgewählt, und sei  $Y$  die Anzahl Kanten von Knoten 1 und  $Z$  die Anzahl Kanten von Knoten 2. Da jeder Knoten mit  $k-1$  anderen Knoten verbunden werden kann, sind sowohl  $Y$  als auch  $Z$  als einzelne Zufallsvariablen binomial verteilt mit Parametern  $k-1$  und  $p$ . Die beiden Zufallsvariablen sind jedoch nicht unabhängig, denn die (mögliche) Kante zwischen Knoten 1 und 2 wird bei beiden Zufallsvariablen gezählt, und somit entsteht dadurch eine Korrelation zwischen  $Y$  und  $Z$ . Man möchte deshalb  $\text{cov}(Y, Z)$  berechnen. Schreibe dazu

$$Y = Y_2 + \dots Y_k,$$



wobei  $Y_i = 1$  ist, falls die Kante zwischen Knoten 1 und Knoten  $i$  vorhanden ist, und 0 sonst. Schreibe ebenso

$$Z = Z_1 + Z_3 + \dots + Z_k,$$

wobei  $Z_i = 1$  ist falls die Kante zwischen Knoten 2 und Knoten  $i$  vorhanden ist. Somit muss zwingend

$$Y_2 = Z_1$$

sein, und ansonsten sind die  $Y_i, Z_i$  unabhängig und Bernoulli-verteilt mit Parameter  $p$ . Wiederholte Anwendung von Satz 5.8 (c) liefert

$$\begin{aligned} \text{cov}(Y, Z) &= \text{cov}(Y_2 + \dots + Y_k, Z_1 + Z_3 + \dots + Z_k) \\ &= \text{cov}(Y_2, Z_1) + \text{cov}(Y_2, Z_3) + \dots + \text{cov}(Y_2, Z_k) + \text{cov}(Y_3, Z_1) + \dots + \text{cov}(Y_k, Z_k) \\ &= \text{cov}(Y_2, Y_2) = \mathbb{V}(Y_2) = p(1-p). \end{aligned}$$

Dabei haben wir benutzt, dass  $Z_1 = Y_2$  ist,  $\text{cov}(Y_2, Y_2) = \mathbb{V}(Y_2)$  nach Satz 5.8 (b) und  $\text{cov}(Y_2, Z_3) = \dots = \text{cov}(Y_k, Z_k) = 0$  sind wegen der Unabhängigkeit. Im letzten Schritt wurde die Varianz der Bernoulli-Verteilung eingesetzt. ■

Während das Vorzeichen der Kovarianz angibt, ob zwei Zufallsvariablen positiv korreliert, negativ korreliert oder unkorreliert sind, enthält der absolute Wert einer Kovarianz keine Information. Man *normiert* deshalb die Kovarianz in geeigneter Art und Weise.

### Definition 5.5

Seien  $X, Y$  zwei Zufallsvariablen mit positiver Varianz auf  $(\Omega, \mathbb{P})$ . Der **Korrelationskoeffizient** oder die **Korrelation** von  $X$  und  $Y$  ist definiert als

$$\text{corr}(X, Y) := \frac{\text{cov}(X, Y)}{\sqrt{\mathbb{V}(X) \cdot \mathbb{V}(Y)}}$$

◆

Notation:  $\varrho(X, Y) := \text{corr}(X, Y)$ .

Warum ist diese Normierung mit der Wurzel aus dem Produkt der Varianzen die richtige? Man kann folgenden Satz beweisen.

### Satz 5.10 (Eigenschaften der Korrelation)

Seien  $X, Y$  zwei Zufallsvariablen mit positiver Varianz auf  $(\Omega, \mathbb{P})$ . Dann gelten

- (a) Das Vorzeichen der Korrelation stimmt mit dem Vorzeichen der Kovarianz überein.
- (b) Die Korrelation nimmt nur Werte in  $[-1, 1]$  an, d.h.  $-1 \leq \text{corr}(X, Y) \leq 1$ .

### Beispiel 5.17 (Erdős-Rényi Zufallsgraph)

In Beispiel 5.16 hatten wir gesehen, dass  $\text{cov}(Y, Z) = p(1-p)$  ist. Weiter gilt  $\mathbb{V}(Y) = \mathbb{V}(Z) = (k-1)p(1-p)$ , da  $Y$  und  $Z$  einzeln binomialverteilt sind. Somit ist die Korrelation zwischen der Anzahl Kanten zweier verschiedener Knoten im Erdős-Rényi Zufallsgraph gegeben durch

$$\text{corr}(Y, Z) = \frac{\text{cov}(Y, Z)}{\sqrt{\mathbb{V}(Y)\mathbb{V}(Z)}} = \frac{p(1-p)}{(k-1)p(1-p)} = \frac{1}{k-1}.$$

Sie nimmt also insbesondere mit der Anzahl der Knoten ab, was nicht überrascht, denn je mehr Knoten der Graph hat, desto weniger Gewicht bekommt die einzelne Kante zwischen zwei festen Knoten im Vergleich zu den anderen. ■

Bis jetzt haben wir uns nur mit zwei Zufallsvariablen beschäftigt. Wenn man mehr als zwei Zufallsvariablen untersuchen möchte, schaut man sich die sogenannte *Kovarianzmatrix* an, welche folgendermaßen definiert ist:

**Definition 5.6**

Seien  $X_1, \dots, X_n$  Zufallsvariablen auf  $(\Omega, \mathbb{P})$ . Die **Kovarianzmatrix** von  $X_1, \dots, X_n$  ist

$$\begin{bmatrix} \text{cov}(X_1, X_1) & \text{cov}(X_1, X_2) & \dots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \text{cov}(X_2, X_2) & \dots & \text{cov}(X_2, X_n) \\ \vdots & & \ddots & \vdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \dots & \text{cov}(X_n, X_n) \end{bmatrix}$$



Man beachte dass nach Satz 5.7 (b) die Einträge auf der Diagonalen gleich den Varianzen sind, und dass die Matrix symmetrisch ist nach Satz 5.7 (a). Man kann im Übrigen zeigen, dass sie *positiv semidefinit* im Sinne linearen Algebra ist. Analog zur Kovarianzmatrix kann die Korrelationsmatrix aufgestellt werden.

**Zusammenfassung**

Kenngrößen von Zufallsvariablen enthalten wichtige Informationen über die entsprechende Verteilung. Der Erwartungswert entspricht dem mittleren Wert einer Zufallsvariablen, die Varianz ist ein Maß für die Streuung, welches in Abschätzungen wie der Chebyshev-Ungleichung benutzt wird. Bei mehreren Zufallsvariablen geben Kovarianz und Korrelation Informationen über das Maß der Abhängigkeit.

## Aufgaben

**5.1** In einer Urne befinden sich 5 rote, 2 blaue und eine weiße Kugel. Definiere folgende Zufallsvariablen:

$X$  = Anzahl blaue Kugeln beim 5fachen Ziehen mit Zurücklegen

$Y$  = Anzahl Kugeln bis zur ersten roten beim Ziehen mit Zurücklegen

Berechne a)  $\mathbb{E}[X]$  b)  $\mathbb{E}[Y]$  c)  $\mathbb{V}(X)$  und d)  $\mathbb{V}(Y)$ .

**5.2** Seien  $X$  und  $Y$  zwei Zufallsvariablen, deren gemeinsame Verteilung in Tabelle 5.3 gegeben ist. a) Berechnen Sie  $\mathbb{E}[X]$  und  $\mathbb{E}[Y]$ , b) Berechnen Sie  $\mathbb{V}(X)$  and  $\mathbb{V}(Y)$  c) Berechnen Sie  $\text{cov}(X, Y)$ .

$\begin{array}{c} X \\ \backslash \\ Y \end{array}$	0	1	2	$\mathbb{P}(Y = y)$
0	0.2	0.4	0	0.6
1	0.1	0	0.3	0.4
$\mathbb{P}(X = x)$	0.3	0.4	0.3	1

Tab. 5.3

**5.3** Seien  $X$  und  $Y$  Zufallsvariablen mit  $\mathbb{V}(X) = 3$ ,  $\mathbb{V}(Y) = 1.2$ , und  $\text{cov}(X, Y) = -2$ . Berechne a)  $\mathbb{V}(2X)$ , b)  $\mathbb{V}(-Y)$ , c)  $\mathbb{V}(X + 4Y)$ , d)  $\mathbb{V}(\frac{1}{2}X - \frac{1}{4}Y)$ .

**5.4** Es sei  $X$  eine Zufallsvariable mit diskreter Gleichverteilung auf der Menge  $\{-2, -1, 0, 1, 2\}$ . Weiter sei  $Y = 2X^3$ . Berechne  $\mathbb{E}[Y]$  und  $\text{cov}(X, Y)$ .

**5.5** Seien  $X$  und  $Y$  Zufallsvariablen mit gemeinsamer Verteilung

$$\mathbb{P}(X = 2, Y = 0) = \frac{1}{4}, \mathbb{P}(X = -1, Y = 1) = \frac{1}{4},$$

$$\mathbb{P}(X = -1, Y = 0) = a, \mathbb{P}(X = 2, Y = 1) = \frac{1}{2} - a.$$

Alle anderen Wahrscheinlichkeiten sind 0. a) Für welche Werte von  $a$  ist dies eine gemeinsame Wahrscheinlichkeitsverteilung? b) Berechne  $\mathbb{E}[X]$ ,  $\mathbb{E}[Y]$  und  $\mathbb{E}[X \cdot Y]$  in Abhängigkeit von  $a$ . c) Berechne  $\text{cov}(X, Y)$  in Abhängigkeit von  $a$ . Für welche Werte von  $a$  sind  $X$  und  $Y$  positiv bzw. negativ korreliert bzw. unkorreliert?

**5.6** Zeigen Sie, dass die Varianz einer geometrisch zum Parameter  $p$  verteilten Zufallsvariablen gleich  $(1 - p)/p^2$  ist

**5.7** Zeigen Sie, dass der Erwartungswert einer Poisson-verteilten Zufallsvariable mit Parameter  $\lambda$  gleich  $\lambda$  ist.

**5.8** Es sei  $X$  Bernoulli-verteilt zum Parameter  $p$  und  $Y$  binomialverteilt mit Parametern  $n, p$ . Berechne  $\mathbb{E}[e^X]$  und  $\mathbb{E}[e^Y]$ .



# 6 Zufallsvariablen mit Dichte

---

## Übersicht

6.1	Zufallsvariablen, Verteilungsfunktion, Dichten . . . . .	80
6.2	Exponentialverteilung . . . . .	85
6.3	Normalverteilung . . . . .	87
6.4	Mehrdimensionale Verteilungen . . . . .	90

---

Bisher haben wir uns (fast) nur mit diskreten Zufallsvariablen beschäftigt, d.h. mit Zufallsvariablen, deren Bild eine endliche oder abzählbar unendliche Menge ist. Wir haben gesehen, dass für solche Zufallsvariablen  $X$  die Verteilung  $p_X(x), x \in X(\Omega)$  im Prinzip alle relevanten Informationen enthält. Außerdem können wir Erwartungswert und Varianz von diskreten Zufallsvariablen berechnen.

Nicht alle in der Wirklichkeit auftretenden Zufallsvariablen sind jedoch diskret. Beispielsweise kann die Wartezeit auf das Ergebnis einer Datenbankabfrage im Prinzip jeden Wert zwischen 0 und  $\infty$  annehmen, also Werte im (überabzählbaren) Intervall  $]0, \infty[$ . Auch die Dauer eines Laserimpulses kann nicht nur bestimmte diskrete Werte annehmen. In diesem Kapitel führen wir deshalb zuerst einige Begriffe für allgemeine Zufallsvariablen ein, und lernen dann einige besonders wichtige Verteilungen kennen, die mit Hilfe einer Dichte beschrieben werden können.

### Lernziele

- Die Zusammenhänge zwischen Zufallsvariablen und Verteilungsfunktion kennen
- Dichten und ihre Eigenschaften kennen
- Mit Dichten rechnen können
- Normalverteilung und Exponentialverteilung, ihr Auftreten, ihre Dichten und ihre Kenngrößen kennen.

### Vorkenntnisse

Solide Kenntnisse der Integral- und Differentialrechnung einschließlich der Integrationstechniken durch Substitution und partielle Integration sind notwendig. Im Abschnitt über gemeinsame Dichten sind Grundkenntnisse in mehrdimensionaler Integration notwendig.

## 6.1 Zufallsvariablen, Verteilungsfunktion, Dichten

In Definition 3.6 wurde die (*kumulative*) *Verteilungsfunktion einer Zufallsvariablen*  $X$  definiert durch

$$F_X(x) = \mathbb{P}(X \leq x).$$

Aus dieser Definition kann man leicht einige Eigenschaften herleiten

### Satz 6.1 (Verteilungsfunktion)

Sei  $F$  die Verteilungsfunktion einer Zufallsvariablen  $X : \Omega \rightarrow \mathbb{R}$ . Dann gelten

- i.  $F : \mathbb{R} \rightarrow [0, 1]$  ist monoton nichtfallend
- ii.  $\lim_{x \rightarrow -\infty} F(x) = 0$
- iii.  $\lim_{x \rightarrow \infty} F(x) = 1$ .

Es ist möglich, umgekehrt vorzugehen, und zuerst den Begriff einer Verteilungsfunktion zu definieren, ohne über Zufallsvariablen gesprochen zu haben. Eine (allgemeine) Verteilungsfunktion ist eine Funktion mit den Eigenschaften aus Satz 6.1, wobei normalerweise außerdem gefordert wird, dass  $F$  *rechtsseitig stetig* ist, d.h. dass  $\lim_{h \searrow 0} F(x+h) = F(x)$  gilt, für alle  $x$  in  $\mathbb{R}$ . Dies ist bei Verteilungsfunktionen mit Sprüngen von Bedeutung. Ein Beispiel ist in Abbildung 6.1 gezeichnet. Es kann gezeigt werden, dass zu jeder solchen Funktion eine Zufallsvariable existiert, so dass  $F$  die zugehörige Verteilungsfunktion im Sinne von Definition 3.6 ist.

Mit allgemeinen Zufallsvariablen und Verteilungsfunktionen zu rechnen ist nicht immer einfach. Bisher hatten wir uns auf diskrete Zufallsvariablen beschränkt. Nun betrachten wir eine weitere wichtige Klasse von Zufallsvariablen, für welche eine *Dichte* existiert.

### Definition 6.1 (Dichte)

Sei  $X$  eine Zufallsvariable und  $F_X$  ihre Verteilungsfunktion.  $X$  bzw.  $F_X$  besitzt eine **Dichte**, falls eine Funktion  $f_X$  existiert, so dass für alle  $x \in \mathbb{R}$  gilt

$$F_X(x) = \int_{-\infty}^x f_X(t) dt.$$

Die Funktion  $f_X$  heißt dann **Dichte von**  $F_X$ , beziehungsweise von  $X$ . ◆

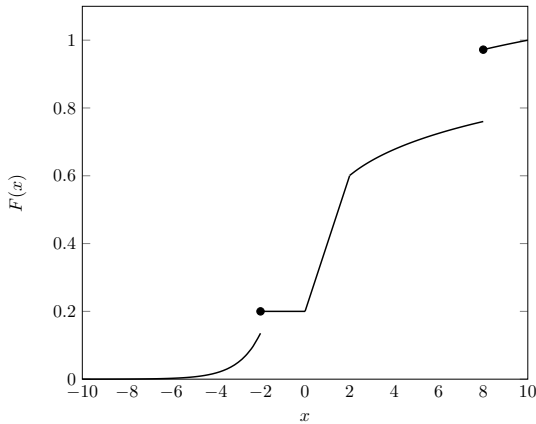


Abb. 6.1: Beispiel für eine Verteilungsfunktion: Monoton nicht fallend, linker Grenzwert ist 0, rechter ist 1. Sprünge sind erlaubt, rechtsstetig.

Falls eine Dichte  $f_X$  existiert, ist  $F_X$  also eine Stammfunktion von  $f_X$ . Ist  $F_X$  in  $x \in \mathbb{R}$  differenzierbar, so gilt nach dem Hauptsatz der Differential- und Integralrechnung

$$F'_X(x) = f_X(x).$$

Aus der Definition einer Dichte kann man folgendes schließen: Für eine Zufallsvariable mit Dichte ist die Wahrscheinlichkeit, genau einen bestimmten Wert  $x$  anzunehmen, gleich 0. Es gilt nämlich nach Definition

$$\mathbb{P}(X = x) = \mathbb{P}(X \leq x) - \lim_{h \searrow 0} \mathbb{P}(X \leq x + h) = F_X(x) - \lim_{h \searrow 0} F_X(x + h) = \int_x^x f(t) dt = 0,$$

da ein Integral über einen einzelnen Punkt gleich 0 ist. Eine Verteilungsfunktion mit Dichte ist also insbesondere stetig.

Dies mag auf den ersten Blick verwirren, betrachtet man jedoch eine konkrete Situation, wird dies anschaulich. Angenommen, die Zufallsvariable  $X$  bezeichne die Dauer eines Laser-Impulses. Die Wahrscheinlichkeit, dass der Impuls eine ganz exakt bemessene Zeit dauert, also z.B.  $2/3$  Sekunden, ist verschwindend klein. Natürlich dauert der Impuls dennoch jedes Mal eine gewisse Zeit, nur kann man diese nicht so exakt bemessen. Die Wahrscheinlichkeit, dass der Impuls eine Dauer zwischen z.B. 0.66 und 0.67 Sekunden hat, ist nun vermutlich nicht mehr 0. Denn dabei handelt es sich um eine Zeitspanne, nicht um einen exakten Wert, oder eine infinitesimal kurze Zeitspanne.

### Satz 6.2 (Eigenschaften einer Dichte)

Eine Funktion  $f : \mathbb{R} \rightarrow \mathbb{R}$  ist genau dann die Dichte einer Zufallsvariablen, wenn folgende Eigenschaften erfüllt sind:

- (a)  $f(t) \geq 0$  für alle  $t \in \mathbb{R}$ ,
- (b)  $\int_{-\infty}^{\infty} f(t) dt = 1$ .

**Beweis 6.2** Wenn  $f$  eine Dichte ist, so muss nach Definition die Funktion  $x \mapsto \int_{-\infty}^x f(t)dt$  die drei Eigenschaften aus Satz 6.1 erfüllen. Wäre  $f(t) \geq 0$  nicht erfüllt, so wäre  $x \mapsto \int_{-\infty}^x f(t)dt$  nicht monoton nichtfallend. Wäre  $\int_{-\infty}^{\infty} f(t)dt \neq 1$ , so wäre **iii.** aus Definition 6.1 nicht erfüllt. Somit muss eine Dichte diese beiden Eigenschaften haben.

Hat umgekehrt  $f$  diese beiden Eigenschaften, so erfüllt  $x \mapsto \int_{-\infty}^x f(t)dt$  die drei Eigenschaften aus Definition 6.1, und ist somit eine Verteilungsfunktion.  $\square$

### Beispiel 6.1

Sei die Funktion  $f : \mathbb{R} \rightarrow \mathbb{R}$  gegeben durch

$$f(t) = \begin{cases} \frac{1}{4}t^3 & 0 \leq t \leq 2 \\ 0 & \text{sonst.} \end{cases}$$

Ist diese Funktion eine Dichte? Es muss überprüft werden, ob die beiden Eigenschaften aus Satz 6.2 erfüllt sind. Offensichtlich gilt  $f(t) \geq 0$  für alle  $t \in \mathbb{R}$ , somit ist (a) erfüllt. Für die zweite Eigenschaft berechne

$$\int_{-\infty}^{\infty} f(t)dt = \int_0^2 \frac{1}{4}t^3 dt = \frac{1}{4} \cdot \left[ \frac{1}{4}t^4 \right]_0^2 = 1.$$

Somit ist auch (b) erfüllt, und  $f$  ist eine Dichte. Hingegen ist  $g : \mathbb{R} \rightarrow \mathbb{R}$  gegeben durch

$$g(t) = \begin{cases} t^3 & 0 \leq t \leq 1 \\ 0 & \text{sonst} \end{cases}$$

keine Dichte, denn  $\int_{-\infty}^{\infty} g(t)dt = \int_0^1 t^3 dt = \frac{1}{4} \neq 1$ .  $\blacksquare$

### Satz 6.3 (Rechnen mit Dichten)

Sei  $X$  eine Zufallsvariable mit Dichte  $f_X$ . Seien  $a, b \in \mathbb{R}$  mit  $a < b$ .

**(a)**  $\mathbb{P}(X \leq b) = \mathbb{P}(X < b) = \int_{-\infty}^b f_X(t)dt$

**(b)**  $\mathbb{P}(X \geq a) = \mathbb{P}(X > a) = \int_a^{\infty} f_X(t)dt$

**(c)**  $\mathbb{P}(X \in [a, b]) = \mathbb{P}(X \in ]a, b]) = \mathbb{P}(X \in ]a, b]) = \mathbb{P}(X \in [a, b]) = \int_a^b f_X(t)dt$

**Beweis 6.3** (a) Aus Definition 3.6 und Definition 6.1 folgt

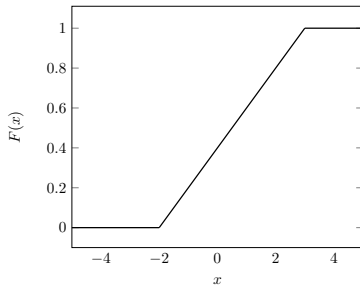
$$\mathbb{P}(X \leq b) = F_X(b) = \int_{-\infty}^b f_X(t)dt.$$

Wegen der Stetigkeit von  $F_X$  gilt  $\mathbb{P}(X \leq b) = \mathbb{P}(X < b)$ . (b) folgt aus (a), da  $\mathbb{P}(X \geq a) = 1 - \mathbb{P}(X < a)$  gilt. (c) folgt mit

$$\mathbb{P}(X \in [a, b]) = \mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X < a) = \int_{-\infty}^b f_X(t)dt - \int_{-\infty}^a f_X(t)dt.$$

Der letzte Ausdruck auf der rechten Seite kann nach den Regeln der Integralrechnung zu  $\int_a^b f_X(t)dt$  zusammengefasst werden. Analog folgen die anderen Gleichungen.  $\square$



Abb. 6.2: Gleichverteilung mit  $a = -2, b = 3$ **Beispiel 6.2 (Wahrscheinlichkeiten und Dichten)**

Sei  $f$  wiederum die Dichte aus Beispiel 6.1, also  $f(t) = \frac{1}{4}t^3$  für  $0 \leq t \leq 2$ , und  $f(t) = 0$  sonst. Dann ist

$$\mathbb{P}(X > 1) = \int_1^\infty f(t)dt = \int_1^2 \frac{1}{4}t^3 dt = \left[ \frac{1}{16}t^4 \right]_1^2 = \frac{15}{16}.$$

Man beachte, dass  $\mathbb{P}(X > 2) = 0$ , und auch  $\mathbb{P}(X < 0) = 0$  ist. Also ist beispielsweise  $\mathbb{P}(1 \leq X \leq 2) = \mathbb{P}(X \geq 1) = \mathbb{P}(X > 1) = \frac{15}{16}$ . ■

**Beispiel 6.3 (Gleichverteilung)**

Seien  $a < b$ ,  $a, b \in \mathbb{R}$ . Die Funktion  $F : \mathbb{R} \rightarrow \mathbb{R}$ ,

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & x \geq b \end{cases}$$

erfüllt die Eigenschaften einer Verteilungsfunktion, wie man in Abbildung 6.2 erkennen kann. Eine Zufallsvariable  $X$  mit dieser Verteilungsfunktion heißt **(stetig) gleichverteilt** (oder uniform verteilt) auf  $[a, b]$ , vergleiche auch Beispiel 3.8.

Hat eine Zufallsvariable mit dieser Verteilung eine Dichte? In anderen Worten, existiert eine Funktion  $f$  mit  $F(x) = \int_{-\infty}^x f(t)dt$ , bzw.  $F'_X = f_X$ ? Nach Definition ist  $F$  auf jedem Fall auf den Intervallen  $] -\infty, a[$  und  $]b, \infty[$  differenzierbar, die Ableitung ist dort 0 weil die Funktion konstant ist. Auf dem (offenen) Intervall  $]a, b[$  ist die Funktion als lineare Funktion ebenfalls differenzierbar, die Ableitung ist die konstante Funktion  $\frac{1}{b-a}$ . Diese Funktion hat rechtsseitige Grenzwerte für  $x \searrow a$  und  $x \searrow b$ , so dass die Funktion

$$f(x) := \begin{cases} 0 & x < a \\ \frac{1}{b-a} & a \leq x < b \\ 0 & x \geq b \end{cases}$$

wohldefiniert ist, und  $f(t) = F'(t)$  für  $x \in \mathbb{R} \setminus \{a\} \cup \{b\}$  gilt. Somit ist also  $f$  die Dichte einer auf  $[a, b]$  gleichverteilten Zufallsvariablen. ■

Auch für Zufallsvariablen mit Dichten sind Kenngrößen wie Erwartungswert und Varianz von Bedeutung.

**Definition 6.2**

Sei  $X$  eine Zufallsvariable mit Dichte  $f_X$ . Der **Erwartungswert** von  $X$  ist definiert als

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} t \cdot f_X(t) dt,$$

falls das Integral existiert. Die **Varianz** von  $X$  ist definiert als

$$\mathbb{V}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2],$$

falls dieser Erwartungswert existiert. ♦

**Satz 6.4**

Sei  $X$  eine Zufallsvariable mit Dichte  $f_X$ , und sei  $g : \mathbb{R} \rightarrow \mathbb{R}$  eine Funktion. Der Erwartungswert der Zufallsvariablen  $g(X)$  ist dann

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(t) \cdot f_X(t) dt,$$

falls das Integral auf der rechten Seite existiert. Somit gilt für die Varianz von  $X$

$$\mathbb{V}(X) = \int_{-\infty}^{\infty} (t - \mathbb{E}[X])^2 f_X(t) dt = \int_{-\infty}^{\infty} t^2 \cdot f_X(t) dt - \left( \int_{-\infty}^{\infty} t \cdot f_X(t) dt \right)^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

**Beweis 6.4** Da der Beweis der ersten Aussage etwas technisch (wenn auch nicht schwer) ist, wird darauf verzichtet. Bei den Aussagen über die Varianz folgt die erste Gleichung dann mit  $g(x) = (x - \mathbb{E}[X])^2$ . Der Rest folgt durch Ausmultiplizieren, vgl. auch den Beweis von Satz 5.3. □

Die Eigenschaften von Erwartungswert und Varianz aus Satz 5.1 und Satz 5.4 gelten auch im Fall von Zufallsvariablen mit Dichten, ebenso wie die Chebyshev-Ungleichung (Satz 5.5) und die Markov-Ungleichung. Es wird darauf verzichtet, die Aussagen und Beweise hier zu wiederholen.

**Beispiel 6.4 (Dichte und Erwartungswert)**

Sei  $f$  die Dichte aus Beispiel 6.1, also

$$f(t) = \begin{cases} \frac{1}{4}t^3 & 0 \leq t \leq 2 \\ 0 & \text{sonst.} \end{cases}$$

Der Erwartungswert einer Zufallsvariable  $X$  mit Dichte  $f$  ist also

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} t \cdot f(t) dt = \int_0^2 t \cdot \frac{1}{4}t^3 dt = \frac{1}{4} \int_0^2 t^4 dt = \frac{8}{5}.$$

Weiter erhalten wir

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} t^2 \cdot f(t) dt = \frac{1}{4} \int_0^2 t^5 dt = \frac{8}{3}.$$

Somit ist die Varianz  $\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{8}{75}$ . ■

**Beispiel 6.5 (Pareto-Verteilung)**

Seien  $a, b > 0$ . Die Funktion  $f : \mathbb{R} \rightarrow \mathbb{R}$  definiert durch

$$f(t) = \begin{cases} \frac{a \cdot b^a}{t^{a+1}} & t \geq b \\ 0 & \text{sonst} \end{cases}$$

erfüllt die Eigenschaften einer Dichte, da

$$\int_b^\infty \frac{a \cdot b^a}{t^{a+1}} dt = \left[ \frac{-b^a}{t^a} \right]_b^\infty = 1$$

ist. Die zugehörige Verteilungsfunktion ist dann

$$F(x) = \begin{cases} 1 - \left(\frac{b}{x}\right)^a & x \geq b \\ 0 & \text{sonst.} \end{cases}$$

Eine Zufallsvariable mit dieser Verteilung heißt **Pareto-verteilt** zu den Parametern  $a$  und  $b$ . Die Paretoverteilung ist das stetige Analog zur Zipf-Verteilung aus Kapitel 3, denn auch hier fällt die Verteilungsfunktion polynomiell in  $x$ . Der Erwartungswert einer Pareto-verteilten Zufallsvariablen  $X$  ist, falls  $a > 1$  ist,

$$\mathbb{E}[X] = \int_b^\infty t \cdot \frac{a \cdot b^a}{t^{a+1}} dt = \frac{ab}{a-1}.$$

Ist  $a < 1$ , so ist der Wert des Integrals unendlich, und der Erwartungswert existiert nicht. Für die Varianz erhält man mit einer ähnlichen Rechnung, dass sie existiert falls  $a > 2$  ist. Man erhält

$$\mathbb{V}(X) = \frac{b^2 a}{(a-1)^2(a-2)}.$$

■

## 6.2 Exponentialverteilung

Es werden nun einige Beispiele für besonders wichtige Verteilungen mit Dichte betrachtet.

**Definition 6.3 (Exponentialverteilung)**

Eine Zufallsvariable  $X$  heißt **exponentialverteilt** mit Parameter  $\lambda > 0$ , falls  $X$  die Dichte

$$f_X(x) = \begin{cases} 0, & x < 0 \\ \lambda \cdot e^{-\lambda x}, & x \geq 0 \end{cases}$$

hat.

◆

Damit diese Definition überhaupt sinnvoll ist, muss überprüft werden, dass  $f_X$  überhaupt eine Dichte ist. Da  $\lambda > 0$  und die Exponentialfunktion nur positive Werte annimmt, gilt auf jeden Fall  $f_X(t) \geq 0$  für alle  $t \in \mathbb{R}$ . Außerdem haben wir

$$\int_{-\infty}^\infty f_X(t) dt = \int_0^\infty \lambda e^{-\lambda t} dt = \left[ -e^{-\lambda t} \right]_0^\infty = 1.$$

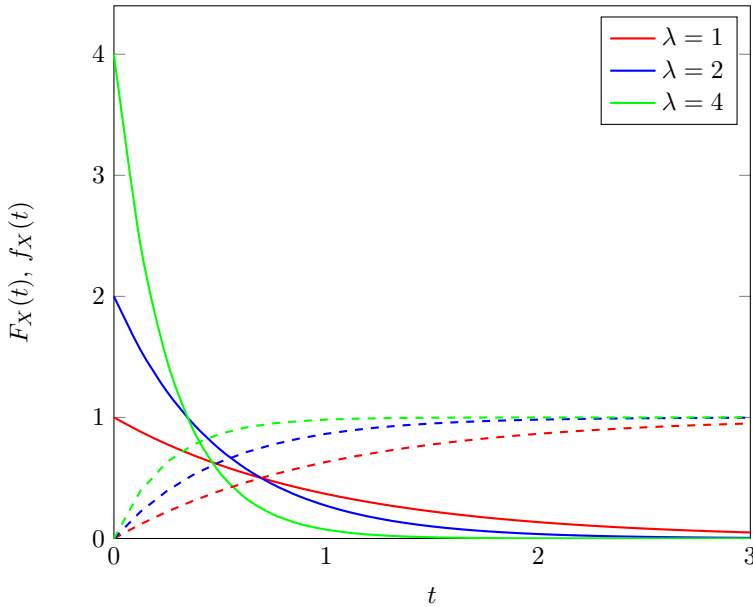


Abb. 6.3: Dichte und Verteilungsfunktion (gestrichelt) der Exponentialverteilung für verschiedene Parameterwerte

Somit ist  $f_X$  eine Dichte. Abbildung 6.3 zeigt einen Plot für verschiedene Parameterwerte.

### Satz 6.5 (Exponentialverteilung)

Sei  $X$  exponentialverteilt mit Parameter  $\lambda > 0$ . Dann ist die Verteilungsfunktion  $F_X$  gegeben durch

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0, \end{cases}$$

und es gelten

$$\mathbb{E}[X] = \frac{1}{\lambda} \quad \text{sowie} \quad \mathbb{V}(X) = \frac{1}{\lambda^2}.$$

**Beweis 6.5** Nach Definition gilt  $F_X(x) = \int_{-\infty}^x f_X(t)dt$ , somit ist klar dass  $F_X(x) = 0$  für  $x < 0$  gelten muss, und für  $x \geq 0$  erhalten wir wie oben

$$F_X(x) = \int_{-\infty}^x f_X(t)dt = \int_0^x \lambda e^{-\lambda t} dt = -e^{-\lambda x} + 1.$$

Weiter gilt nach Definition 6.2

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} t f_X(t) dt = \int_0^{\infty} t \lambda e^{-\lambda t} dt.$$

Dieses Integral kann mittels partieller Integration berechnet werden, und man findet (mit  $u'(t) = \lambda e^{-\lambda t}$ ,  $v(t) = t$ , also  $u(t) = -e^{-\lambda t}$ ,  $v'(t) = 1$ ),

$$\int_0^{\infty} t \lambda e^{-\lambda t} dt = \left[ -te^{-\lambda t} \right]_0^{\infty} + \int_0^{\infty} e^{-\lambda t} dt = \frac{1}{\lambda}.$$

Eine ähnliche Rechnung mit zweimaliger partieller Integration führt auf die Formel für die Varianz.  $\square$

Die Exponentialverteilung ist in vielerlei Hinsicht das stetige Analog zur geometrischen Verteilung. Beispielsweise gilt folgende Aussage: Sei  $(X_n)_{n \in \mathbb{N}}$  eine Folge von unabhängigen Zufallsvariablen, so dass  $X_n$  geometrisch verteilt mit Parameter  $p_n$  ist, wobei  $\lim_{n \rightarrow \infty} np_n = \lambda$  gilt. Dann gilt für alle  $x > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\frac{X_n}{n} \leq x\right) = 1 - e^{-\lambda x}, \quad (6.1)$$

d.h. die Verteilung der „reskalierten“ geometrischen Zufallsvariablen  $X_n/n$  konvergiert gegen eine Exponentialverteilung, sofern die Parameter geeignet konvergieren (vgl. Aufgabe 3.10). Dies ist ein Beispiel für einen Grenzwertsatz, ähnlich wie der Poisson-Grenzwertsatz 3.5.

Eine weitere Analogie zur geometrischen Verteilung ist die sogenannte *Gedächtnislosigkeit*, dazu mehr in Kapitel 14. Wie die geometrische Verteilung tritt auch die Exponentialverteilung oft in Situationen auf, in denen *Wartezeiten* eine Rolle spielen. Typischerweise können z.B. folgende Größen als exponentialverteilt modelliert werden:

- Wartezeit auf einen radioaktiven Zerfall
- Wartezeiten zwischen zwei Anrufen in einem Callcenter
- Restlebenszeit eines Bauteils mit konstanter Ausfallrate
- Zeit zwischen zwei Mutationen auf der DNA

## 6.3 Normalverteilung

In diesem Abschnitt wird die vermutlich bekannteste Wahrscheinlichkeitsverteilung vorgestellt. Der Grund für ihre Wichtigkeit wird in Satz 7.2 näher erläutert werden.

### Definition 6.4

Eine Zufallsvariable  $X$  heißt **normalverteilt** bzw. **Gauß-verteilt** zu den Parametern  $\mu \in \mathbb{R}, \sigma^2 > 0$ , falls  $X$  die Dichte

$$f_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

hat. Falls  $\mu = 0$  und  $\sigma^2 = 1$  gelten, also  $f_X(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}$ , so nennt man  $X$  **standardnormalverteilt**. Ist  $X$  eine normalverteilte Zufallsvariable, so schreibt man auch  $X \sim \mathcal{N}(\mu, \sigma^2)$ .  $\blacklozenge$

Für die Dichte der Normalverteilung verwendet man oft die Notation  $\varphi_{\mu, \sigma}$  oder  $\varphi_{\mu, \sigma^2}$ , für die Verteilungsfunktion  $\Phi_{\mu, \sigma^2}$ . Im Fall der Standardnormalverteilung bezeichnen wir also Verteilungsfunktion bzw. Dichte mit  $\Phi_{0,1}$  bzw.  $\varphi_{0,1}$ . In Abbildung 6.4 wird die Dichte für verschiedene Werte der Parameter abgebildet. Es handelt sich hierbei um die

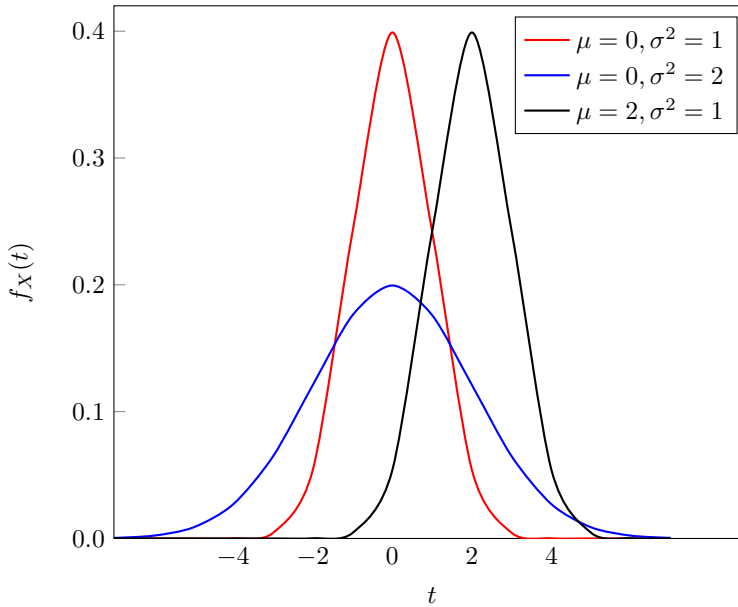


Abb. 6.4: Dichte der Normalverteilung für verschiedene Parameterwerte

berühmte *Gauß'sche Glockenkurve*. Auch hier muss wiederum überprüft werden, dass es sich bei der angegebenen Funktion tatsächlich um eine Dichte handelt. Es ist nicht schwer zu sehen, dass die Funktion nichtnegative Werte annimmt. Aus der Analysis ist bekannt, dass  $\int_{-\infty}^{\infty} e^{-y^2} dy = \sqrt{\pi}$  ist. Mittels Substitution erhält man daraus  $\int_{-\infty}^{\infty} f_X(t) dt = 1$ .

### Satz 6.6 (Normalverteilung)

Sei  $X$  eine normalverteilte Zufallsvariable mit Parametern  $\mu \in \mathbb{R}$  und  $\sigma^2 > 0$ . Dann gelten

$$\mathbb{E}[X] = \mu \quad \text{und} \quad \mathbb{V}(X) = \sigma^2.$$

**Beweis 6.6** Nach Definition 6.2 gilt

$$\begin{aligned} \mathbb{E}[X] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left( \frac{x-\mu}{\sqrt{2\sigma^2}} + \frac{\mu}{\sqrt{2\sigma^2}} \right) e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} y e^{-y^2} dy + \frac{\mu}{\sqrt{\pi}} \int_{-\infty}^{\infty} y e^{-y^2} dy, \end{aligned}$$

wobei wir in der letzten Zeile die Substitution  $y = \frac{x-\mu}{\sqrt{2\sigma^2}}$  durchgeführt haben. Das erste Integral ist gleich 0, da der Integrand eine ungerade Funktion ist. Wegen  $\int_{-\infty}^{\infty} y e^{-y^2} dy = \sqrt{\pi}$  folgt  $\mathbb{E}[X] = \mu$ . Die Formel für die Varianz wird ähnlich berechnet.  $\square$

Im Falle der Normalverteilung kann man keine geschlossene Form, d.h. keine explizite Formel, für die Verteilungsfunktion angeben. Die Werte der Verteilungsfunktion der Standardnormalverteilung sind jedoch in Tabellen aufgelistet, deren Verwendung etwas

Übung bedarf. Man kann sich die Werte auch von gängigen Statistik-Programmen (z.B. R) ausgeben lassen.

Der nächste Satz zeigt, dass eine normalverteilte Zufallsvariable  $X$  mit beliebigem (aber bekanntem) Erwartungswert und Varianz durch Subtraktion von  $\mu$  und Division durch die Standardabweichung  $\sigma = \sqrt{\sigma^2}$  auf eine standardnormalverteilte Zufallsvariable transformiert werden kann. Damit kann man aus der Standardnormalverteilung auch Werte für beliebige Normalverteilungen berechnen.

**Satz 6.7 (Standardisierung)**

Sei  $X \sim \mathcal{N}(\mu, \sigma^2)$ . Dann ist

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

**Beweis 6.7** Mittels einer Substitution  $s = \frac{t - \mu}{\sigma}$  und umformen erhält man

$$\mathbb{P}\left(\frac{X - \mu}{\sigma} \leq y\right) = \mathbb{P}(X \leq \sigma y + \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{s^2}{2}} dt,$$

was die Verteilungsfunktion der Standardnormalverteilung ist. Somit ist  $\frac{X - \mu}{\sigma}$  standardnormalverteilt.  $\square$

**Satz 6.8 (Summe normalverteilter Zufallsvariablen)**

Seien  $X_1$  und  $X_2$  unabhängige, normalverteilter Zufallsvariablen mit Parametern  $\mu_1, \sigma_1^2$  bzw.  $\mu_2, \sigma_2^2$ . Dann ist  $X_1 + X_2$  wieder normalverteilt, mit Parameter  $\mu = \mu_1 + \mu_2$  und  $\sigma^2 = \sigma_1^2 + \sigma_2^2$ . Der Satz gilt analog auch für die Summe von  $n \geq 2$  unabhängigen, normalverteilten Zufallsvariablen.

Dies folgt aus einem Resultat, welches analog zur Faltungsformel aus Satz 4.4 ist (mit Integralen statt mit Summen im Fall von Dichten). Auf den Beweis wird hier verzichtet.

Da das Rechnen mit der Normalverteilung aufgrund der fehlenden expliziten Formel für die Verteilungsfunktion etwas umständlich ist, betrachten wir hier zwei ausführliche Rechenbeispiele.

**Beispiel 6.6**

Sei  $X \sim \mathcal{N}(-1.3, 4)$ . Ziel: Berechne  $\mathbb{P}(X > 0)$ . Es gilt also  $\mu = -1.3, \sigma^2 = 4$ . Wir formen die gesuchte Wahrscheinlichkeit erst so um, dass die Verteilungsfunktion auftritt:

$$\mathbb{P}(X > 0) = 1 - \mathbb{P}(X \leq 0) = 1 - \Phi_{-1.3, 4}(0).$$

Da dieser Wert nicht direkt berechnet werden kann, ist Satz 6.7 nützlich. Wir subtrahieren im Ausdruck  $\mathbb{P}(X \leq 0)$  auf beiden Seiten den Erwartungswert und dividieren durch die Standardabweichung. Somit folgt

$$\mathbb{P}(X > 0) = 1 - \mathbb{P}(X \leq 0) = 1 - \mathbb{P}\left(\frac{X - \mu}{\sigma} \leq \frac{0 - \mu}{\sigma}\right) = 1 - \mathbb{P}\left(\frac{X + 1.3}{2} \leq 0.65\right).$$

Nach Satz 6.7 ist  $Y := \frac{X + 1.3}{2} = \frac{X - \mu}{\sigma}$  standardnormalverteilt. Somit gilt

$$1 - \mathbb{P}\left(\frac{X + 1.3}{2} \leq 0.65\right) = 1 - \mathbb{P}(Y \leq 0.65) = 1 - \Phi_{0,1}(0.65),$$

was laut der Tabelle A.1 im Anhang ungefähr gleich  $1 - 0.7422 = 0.2578$  ist.  $\blacksquare$

**Beispiel 6.7**

Sei  $X \sim \mathcal{N}(-1.3, 4)$ . Ziel: Berechne  $\mathbb{P}(X > -2)$ . Wie im vorigen Beispiel erhält man

$$\mathbb{P}(X > -2) = 1 - \mathbb{P}(X \leq -2) = 1 - \mathbb{P}\left(\frac{X + 1.3}{2} \leq -0.35\right) = 1 - \Phi_{0,1}(-0.35).$$

Allerdings taucht der Wert -0.35 nicht in Tabelle A.1 auf. Abhilfe schafft jedoch die Symmetrie der Normalverteilungsdichte: Es gilt für  $Y \sim \mathcal{N}(0, 1)$

$$\mathbb{P}(Y \leq -a) = \mathbb{P}(Y > a).$$

Somit erhält man

$$1 - \Phi_{0,1}(-0.35) = \Phi_{0,1}(0.35) \approx 0.6368.$$

Vorsicht: Die Symmetrie  $\mathbb{P}(Y \leq -a) = \mathbb{P}(Y > a)$  gilt nur für  $\mu = \mathbb{E}[Y] = 0$ . Andernfalls ist die Spiegelachse der Dichtefunktion die Gerade mit  $x = \mu$ . ■

**Beispiel 6.8 (Standardintervalle um den Erwartungswert)**

Sei  $X$  eine normalverteilte Zufallsvariable mit Parametern  $\mu$  und  $\sigma^2$ . Im Mittel nimmt  $X$  also Werte in der Nähe von  $\mu$  an, und die Varianz  $\sigma^2$  bzw. die Standardabweichung  $\sigma$  sind Maße für die Abweichung von  $X$  vom Erwartungswert  $\mu$ . Mit Hilfe von Satz 6.7 und  $Y := \frac{X - \mu}{\sigma}$  berechnet man leicht

$$\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) = \mathbb{P}(-1 \leq Y \leq 1) = \Phi_{0,1}(1) - \Phi_{0,1}(-1) = 2\Phi_{0,1}(1) - 1 \approx 0.6827.$$

In anderen Worten, eine normalverteilte Zufallsvariable nimmt mit einer Wahrscheinlichkeit von 0.6827 oder gut 2/3 Werte innerhalb eines Intervalls von  $\pm\sigma$  um den Erwartungswert an. Man sagt auch, 68% der Werte liegen *innerhalb einer Standardabweichung* vom Erwartungswert. Mit analogen Rechnungen findet man

$$\mathbb{P}(X \in [\mu - 2\sigma, \mu + 2\sigma]) \approx 0.9545 \quad \text{und} \quad \mathbb{P}(X \in [\mu - 3\sigma, \mu + 3\sigma]) \approx 0.9973.$$

Diese Werte werden in statistischen Anwendungen oft verwendet. ■

Die Normalverteilung spielt eine enorm wichtige Rolle. Sie tritt nämlich als *universelle* Verteilung immer dann auf, wenn viele unabhängige, gleichartige Zufallsexperimente ausgeführt werden. Deshalb sind in Natur und Technik viele Größen (annähernd) normalverteilt, beispielsweise das Gewicht von Lebewesen, oder die Anzahl Reisender auf einer bestimmten Flugstrecke, die Anzahl Nutzer eines bestimmten Online-Dienstes während einer festen Zeitspanne, usw. Dies wird in Kapitel 7 diskutiert werden, und spielt insbesondere in der Statistik eine wichtige Rolle.

## 6.4 Mehrdimensionale Verteilungen

Bei diskreten Zufallsvariablen hatten wir die gemeinsame Verteilung von mehreren Zufallsvariablen betrachtet, die insbesondere für zwei Zufallsvariablen in einer Tabelle dargestellt werden kann. Der Begriff der Unabhängigkeit sowie Kovarianz und Korrelation spielten dabei eine wichtige Rolle. Dies kann auch für Zufallsvariablen mit Dichte untersucht werden.



**Definition 6.5**

Seien  $X_1, \dots, X_n$  Zufallsvariablen mit Werten in  $\mathbb{R}$ , welche auf demselben Wahrscheinlichkeitsraum  $(\Omega, \mathbb{P})$  definiert sind. Die **gemeinsame Verteilungsfunktion** von  $X_1, \dots, X_n$  ist die Funktion  $F_X : \mathbb{R}^n \rightarrow [0, 1]$ , definiert durch

$$F_X = F_{(X_1, \dots, X_n)}(x_1, \dots, x_n) := \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n),$$

$x_1, \dots, x_n \in \mathbb{R}$ . ◆

Dies ist eine direkte Verallgemeinerung von Definition 3.6.

**Definition 6.6**

Zufallsvariablen  $X_1, \dots, X_n$  mit Werten in  $\mathbb{R}$ , welche auf demselben Wahrscheinlichkeitsraum  $(\Omega, \mathbb{P})$  definiert sind, besitzen eine **gemeinsame Dichte**, falls eine Funktion  $f_X : \mathbb{R}^n \rightarrow \mathbb{R}$  existiert, so dass

$$F_X(x_1, \dots, x_n) = \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_1} f_X(t_1, \dots, t_n) dt_1 \dots dt_n$$

gilt. Auf der rechten Seite steht dabei das  $n$ -dimensionale Integral. Dann heißt  $f_X$  auch **gemeinsame Dichte** von  $X_1, \dots, X_n$ . ◆

Besonders einfach ist die gemeinsame Dichte zu bestimmen, wenn die einzelnen Zufallsvariablen unabhängig sind, vgl. Definition 4.4. Der folgende Satz wird durch direktes Nachrechnen bewiesen.

**Satz 6.9 (Unabhängige Zufallsvariablen)**

Seien  $X_1, \dots, X_n$  unabhängige Zufallsvariablen, die jeweils einzeln Dichten  $f_{X_1}, \dots, f_{X_n}$  besitzen. Dann ist

$$f_X(x_1, \dots, x_n) := f_{X_1}(x_1) \cdot \dots \cdot f_{X_n}(x_n)$$

die gemeinsame Dichte von  $X_1, \dots, X_n$ .

**Beispiel 6.9**

Ist beispielsweise  $X_1$  exponentialverteilt mit Parameter  $\lambda$ , und  $X_2$  unabhängig davon gleichverteilt auf  $[1, 4]$ , so ist die gemeinsame Dichte von  $X_1$  und  $X_2$  gegeben durch

$$f(x_1, x_2) = \begin{cases} \frac{\lambda}{3} e^{-\lambda x_1}, & \text{falls } x_1 \geq 0, 1 \leq x_2 \leq 4 \\ 0 & \text{sonst.} \end{cases}$$

Somit ist

$$\mathbb{P}(X_1 \leq 5, X_2 \leq 3) = \int_0^5 \int_1^3 \frac{\lambda}{3} e^{-\lambda x_1} dx_2 dx_1 = \int_0^5 \frac{2\lambda}{3} e^{-\lambda x_1} dx_1 = \frac{2}{3} (1 - e^{-5\lambda}).$$

Da  $X_1$  und  $X_2$  unabhängig sind, gilt  $\mathbb{P}(X_1 \leq 5, X_2 \leq 3) = \mathbb{P}(X_1 \leq 5) \mathbb{P}(X_2 \leq 3)$ , was eine kurze Rechnung bestätigt. ■

**Beispiel 6.10 (Mehrdimensionale Gleichverteilung)**

Angelehnt an Beispiel 1.2 soll eine Zufallsvariable  $X$  mit Werten in  $[0, 1]^n$ , also im  $n$ -dimensionalen „Hyperwürfel“  $n$ -**dimensional gleichverteilt** heißen, wenn für jede Teilmenge  $A \subseteq [0, 1]^n$  gilt, dass  $\mathbb{P}(X \in A)$  gleich der Fläche von  $A$  ist. Dann sieht man relativ leicht ein, dass ein Vektor  $X = (X_1, \dots, X_n)$  genau dann  $n$ -dimensional gleichverteilt ist, wenn die  $X_i, i = 1, \dots, n$  unabhängig sind, und jeweils einzeln auf  $[0, 1]$  gleichverteilt sind.

Analog kann eine Gleichverteilung auf allgemeineren beschränkten Teilmengen von  $\mathbb{R}^n$  definiert werden, z.B. auf  $[a_1, b_1] \times \dots \times [a_n, b_n]$  für  $a_i, b_i \in \mathbb{R}, a_i < b_i$ . Man erhält also die gemeinsame Dichte der mehrdimensionalen Gleichverteilung auf  $[a_1, b_1] \times \dots \times [a_n, b_n]$ .

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n \frac{1}{b_i - a_i} \quad \text{für } a_i \leq x_i < b_i, i = 1, \dots, n,$$

und  $f_{X_1, \dots, X_n}(x_1, \dots, x_n) = 0$  sonst. ■

**Satz 6.10 (Randdichte)**

Seien  $X_1, \dots, X_n$  Zufallsvariable mit gemeinsamer Dichte  $f_X$ . Dann ist

$$f_i(x_i) := \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_X(x_1, \dots, x_n) dx_{x_1} \dots dx_{x_{i-1}} dx_{x_{i+1}} \dots dx_n$$

eine Dichte für die Zufallsvariable  $X_i, i = 1, \dots, n$ . Dabei ist das Integral auf der rechten Seite das  $n - 1$ -dimensionale Integral, bei dem über alle anderen Variablen integriert wird. Entsprechend gilt

$$\mathbb{E}[X_i] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} x_i \cdot f_X(x_1, \dots, x_n) dx_{x_1} \dots dx_n,$$

wobei das Integral  $n$ -dimensional ist.

**Beweis 6.7** Dies folgt durch Einsetzen in die Definition. Auf Details wird verzichtet. □

**Beispiel 6.11**

Für die gemeinsame Dichte aus Beispiel 6.9 folgt nach diesem Satz

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 = \int_1^4 \frac{\lambda}{3} e^{-\lambda x_1} dx_2 = \lambda e^{-\lambda x_1}, x_1 \geq 0,$$

und

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 = \int_0^{\infty} \frac{\lambda}{3} e^{-\lambda x_1} dx_1 = \frac{1}{3}, \quad \text{falls } 1 \leq x_2 \leq 4,$$

$f_2(x_2) = 0$  sonst. Man erhält also in der Tat die ursprünglichen Dichten zurück. ■

Die Kovarianz von zwei Zufallsvariablen mit Dichte ist analog zum diskreten Fall definiert.

**Definition 6.7 (Kovarianz)**

Die **Kovarianz** zweier Zufallsvariablen ist auch im Falle von Zufallsvariablen mit gemeinsamer Dichte wie in Definition 5.4 gegeben durch

$$\text{cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2].$$

Sind  $X_1, \dots, X_n$  Zufallsvariablen mit gemeinsamer Dichte  $f_X$ , so gilt

$$\text{cov}(X_i, X_j) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} t_i \cdot t_j \cdot f_X(t_1, \dots, t_n) dt_1 \dots dt_n - \mathbb{E}[X_i] \cdot \mathbb{E}[X_j]$$

für alle  $i, j \in \{1, \dots, n\}$ . In diesem Fall heißt der Vektor  $(\mathbb{E}[X_1], \dots, \mathbb{E}[X_n])$  **Erwartungswertvektor**. Die **Kovarianzmatrix** von  $X_1, \dots, X_n$  ist die Matrix

$$C(X) := ((\text{cov}(X_i, X_j))_{i,j=1,\dots,n}.$$

◆

**Beispiel 6.12**

Mit der Dichte aus Beispiel 6.9 erhält man  $\mathbb{E}[X_1] = \lambda, \mathbb{E}[X_2] = \frac{5}{2}$ , und  $\text{cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \frac{5\lambda}{2}$ , wobei

$$\mathbb{E}[X_1 X_2] = \int_0^{\infty} \int_1^4 \frac{\lambda}{3} x_1 x_2 e^{-\lambda x_1} dx_2 dx_1 = \int_0^{\infty} \frac{\lambda}{3} \cdot \frac{15}{2} x_1 e^{-\lambda x_1} dx_1 = \frac{5}{2} \lambda$$

ist. Somit ist  $\text{cov}(X_1, X_2) = 0$ , was zu erwarten war, da im Beispiel die Zufallsvariablen unabhängig waren, und somit auch unkorreliert. ■

Auch unter den mehrdimensionalen Verteilungen ist die Normalverteilung besonders wichtig.

**Definition 6.8 (Mehrdimensionale Normalverteilung)**

Sei  $\mu = (\mu_1, \dots, \mu_n) \in \mathbb{R}^n$ , und sei  $A = (a_{i,j})_{i,j=1,\dots,n} \in \mathbb{R}^{n \times n}$  eine symmetrische positiv definite Matrix. Ein Zufallsvektor  $X = (X_1, \dots, X_n)$  mit gemeinsamer Dichte

$$f_X(x_1, \dots, x_n) = \frac{1}{\sqrt{(2\pi)^n |\det A|}} e^{-\frac{1}{2}(t-\mu)^T A^{-1}(t-\mu)}$$

für  $t = (t_1, \dots, t_n) \in \mathbb{R}^n$  heißt **(n-dimensional) Gauß-verteilt** oder **normalverteilt**. ◆

Man kann zeigen, dass in diesem Fall  $\mu$  der Erwartungswertvektor und  $A$  die Kovarianzmatrix von  $X_1, \dots, X_n$  sind. Auch zeigt eine kurze Rechnung, dass die Randdichten (eindimensionale) Normalverteilungen sind, die einzelnen Koordinaten  $X_1, \dots, X_n$  also normalverteilte Zufallsvariablen sind (mit Erwartungswert  $\mu_i$  Varianz  $a_{i,i}, i = 1, \dots, n$ ). Für diese und viele weitere wichtige Eigenschaften der mehrdimensionalen Normalverteilung sei auf die weiterführende Literatur verwiesen. Insbesondere in der Statistik spielen mehrdimensionale Normalverteilungen eine große Rolle.

Ist in Definition 6.8 die Kovarianzmatrix  $A$  eine Diagonalmatrix, so bedeutet das nach Definition 5.4, dass die Zufallsvariablen  $X_1, \dots, X_n$  welche die Koordinaten von  $X$  bilden, unkorreliert sind. Eine kurze Rechnung zeigt, dass in diesem Fall die gemeinsame Dichte aus Definition 6.8 gleich dem Produkt der Randdichten ist, dass also  $X_1, \dots, X_n$  unabhängig sind. Damit gilt der folgende Satz.

**Satz 6.11**

*Ist der Zufallsvektor  $(X_1, \dots, X_n)$  mehrdimensional normalverteilt, so sind die  $X_i$  genau dann unabhängig, wenn sie unkorreliert sind.*

Für andere Verteilungen als der Normalverteilung gilt nur, dass Unabhängigkeit Unkorreliertheit impliziert, jedoch im Allgemeinen nicht die Umkehrung.

**Beispiel 6.13**

Für  $n = 2$  ist  $\mu = (0, 1)$  ein Erwartungswertvektor, und  $A = \begin{bmatrix} 3 & -1 \\ -1 & 1 \end{bmatrix}$  eine symmetrische, positiv definite Matrix mit  $\det A = 2$  und  $A^{-1} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}$ . Somit erhält man beispielsweise

$$\mathbb{P}(X_1 \leq 0, X_2 \leq 3) = \int_{-\infty}^0 \int_{-\infty}^3 \frac{1}{4\pi} e^{-(t_1, t_2 - 1)A^{-1}(t_1, t_2 - 1)^T} dt_2 dt_1.$$

Es gilt  $\mathbb{V}(X_1) = \text{cov}(X_1, X_1) = 3, \text{cov}(X_1, X_2) = \text{cov}(X_2, X_1) = -1$  und  $\mathbb{V}(X_2) = 1$ , sowie  $\mathbb{E}[X_1] = \mu_1 = 0, \mathbb{E}[X_2] = \mu_2 = 1$ . ■

### Zusammenfassung

Allgemein werden Zufallsvariablen mit Hilfe von Verteilungsfunktionen beschrieben. Kann die Verteilungsfunktion als Integral geschrieben werden, so ist der Integrand die zugehörige Dichte. Mit Hilfe von Dichten können Wahrscheinlichkeiten und Kenngrößen einfach berechnet werden. Die Formeln sind dabei analog zum diskreten Fall, beinhalten jedoch Integrale statt Summen. Besonders wichtig sind Exponentialverteilung und Normalverteilung. Auch mehrdimensionale Zufallsvariablen können Dichten haben, dabei wird mehrdimensional integriert.

## Aufgaben

**6.1** Welche der folgenden Funktionen  $f: \mathbb{R} \rightarrow \mathbb{R}$  sind Wahrscheinlichkeitsdichten?

- $f(x) = x$ , für  $0 \leq x \leq \sqrt{2}$ ,  $f(x) = 0$  sonst.
- $f(x) = \frac{1}{2} \sin(x)$  für  $0 \leq x \leq \pi$ ,  $f(x) = 0$  sonst.
- $f(x) = x^2 - 2x + \frac{1}{3}$  für  $0 \leq x \leq 3$ ,  $f(x) = 0$  sonst.
- $f(x) = \frac{1}{2^k}$  für  $(k-1) < x \leq k \forall k \in \mathbb{N}$ ,  $f(x) = 0$  sonst.

**6.2** Für  $c > 0$  sei  $f : \mathbb{R} \rightarrow [0, \infty)$  definiert durch

$$f(x) = \begin{cases} c \cdot x^{-3}, & x \geq 1 \\ 0 & \text{sonst.} \end{cases}$$

a) Bestimmen Sie  $c$  so dass  $f$  eine Dichte ist. b) Berechnen Sie  $\mathbb{E}[X]$  für eine Zufallsvariable  $X$  mit Dichte  $f$ . c) Berechnen Sie  $\mathbb{V}(X)$ . d) Berechnen Sie die Verteilungsfunktion  $F_X$ .

**6.3** Zeigen Sie, dass der Erwartungswert der Gleichverteilung aus Beispiel 6.3 gegeben ist durch  $\mathbb{E}[X] = \frac{b+a}{2}$ . Berechnen Sie auch die zugehörige Varianz.

**6.4** Für  $a, b \geq 0$  sei  $g : \mathbb{R} \rightarrow [0, \infty)$  definiert durch

$$g(x) = \begin{cases} a & \text{falls } x \in [-5, -4], \\ b & \text{falls } x \in [1, 3], \\ 0 & \text{sonst.} \end{cases}$$

g) Bestimme  $a$  und  $b$  so dass  $g$  die Dichte einer Zufallsvariable  $X$  mit  $\mathbb{E}[X] = 1$  ist. b) Berechne  $\mathbb{P}(X > 0)$  für eine Zufallsvariable  $X$  mit Dichte  $g$  c) Berechne  $\mathbb{P}(|X| \leq 2)$  d) Berechne  $\mathbb{P}(X = -4)$ .

**6.5** Man beweise die Aussage über die Varianz in Satz 6.5.

**6.6** Sei  $Y$  eine exponentialverteilte Zufallsvariable mit Parameter  $\lambda = 2/3$ . Berechne a)  $\mathbb{E}[Y^2]$ , b)  $\mathbb{P}(1 \leq Y < 5/2)$ , c)  $\mathbb{P}(Y^2 \geq 4)$  c)  $\mathbb{P}(2Y < 1)$ .

**6.7** Beweisen Sie die Konvergenz (6.1).

**6.8** Sei  $X$  exponentialverteilt mit Parameter  $\lambda$ . Zeigen Sie: Dann gilt  $\mathbb{P}(X > a+b | X > a) = \mathbb{P}(X > b)$  für alle  $a, b > 0$ .

**6.9** Seien  $X_1, \dots, X_n$  unabhängige, normalverteilte Zufallsvariablen, welche alle denselben Erwartungswert  $\mu$  und dieselbe Varianz  $\sigma^2$  haben. Zeigen Sie: Dann ist

$$Y := \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu)$$

standardnormalverteilt.

**6.10** Sei  $Y$  standardnormalverteilt. Zeigen Sie: Dann ist  $X = aY + b$  normalverteilt mit Parametern  $\mathbb{E}[X] = b$  und  $\mathbb{V}(X) = a$ .

**6.11** Sei  $X$  eine normalverteilte Zufallsvariable mit Parametern  $\mu = 1, \sigma^2 = 4$ . Berechne a)  $\mathbb{P}(X \geq 0)$ , b)  $\mathbb{P}(-4 \leq X \leq -3)$ , c)  $\mathbb{P}(|X| < 2)$  d)  $\mathbb{P}(3X < 4)$ .

**6.12** Sei  $X$  eine normalverteilte Zufallsvariable mit Parametern  $\mu = 0, \sigma^2 = 1$ . a) Bestimmen Sie  $x \in \mathbb{R}$  so dass  $\mathbb{P}(X \geq x) = 0.5$ , b) Bestimmen Sie  $y \in \mathbb{R}$  so dass  $\mathbb{P}(|X| \geq y) = 0.5$ .

**6.13** Es seien  $X, Y$  zwei Zufallsvariablen mit gemeinsamer Dichte

$$f(x, y) = \begin{cases} \frac{1}{4}xy^2 & \text{falls } x \in [0, 2], y \in [-1, 2], \\ 0 & \text{sonst.} \end{cases}$$

Berechnen Sie  $\mathbb{P}(X \leq 1, Y \geq 1), \mathbb{E}[X], \mathbb{E}[Y], \text{cov}(X, y)$  sowie die Randdichten.

# 7 Grenzwertsätze

---

## Übersicht

7.1	Gesetz der großen Zahlen.....	97
7.2	Zentraler Grenzwertsatz und Anwendungen.....	99

---

In diesem Kapitel werden einige der wichtigsten Grenzwertsätze formuliert, welche eine wichtige theoretische Grundlage insbesondere für die Statistik bilden. Dabei werden nun nicht mehr eine oder endlich viele Zufallsvariablen betrachtet, sondern Folgen von Zufallsvariablen, und deren Konvergenzverhalten untersucht.

### Lernziele

- Das Gesetz der großen Zahlen und seine Bedeutung kennen
- Den zentralen Grenzwertsatz und seine Bedeutung kennen
- Normalapproximation für die Binomialverteilung berechnen können.

### Vorkenntnisse

Neben dem Stoff der vorherigen Kapiteln, insbesondere zur Normalverteilung, werden Kenntnisse über die Konvergenz von Folgen benötigt.

## 7.1 Gesetz der großen Zahlen

In Kapitel 5 hatten wir einen fairen Würfel betrachtet, und mit  $X$  das Ergebnis eines Wurfs bezeichnet. Das Histogramm zeigt die Häufigkeitsverteilung der Ergebnisse beim hundertfachen Würfeln. Mit  $y_i$  bezeichnen wir Ergebnis des  $i$ -ten Wurfs.

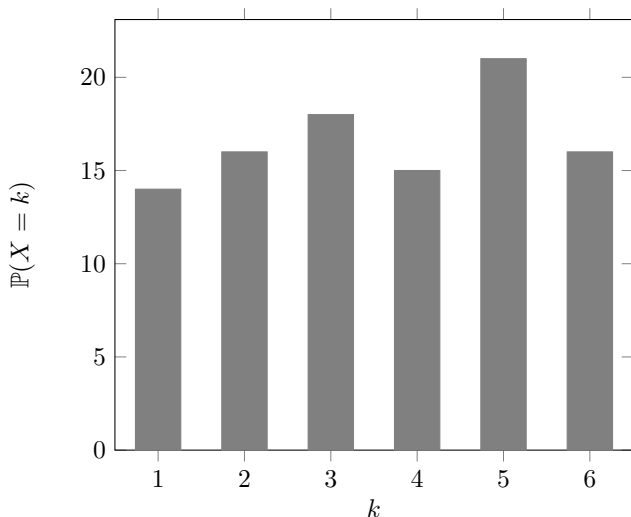


Abb. 7.1: Histogramm: Simulation von 100 Würfeln

In diesem Beispiel gilt

$$\frac{1}{100} \sum_{i=1}^{100} y_i = 3.61 \approx 3.5 = \mathbb{E}[X]$$

gilt, d.h. das beobachtete Mittel der hundert Würfe liegt nah am berechneten Erwartungswert, wie auch schon in Kapitel 5 argumentiert wurde. Wenn nun 1000 statt 100 Würfe ausgeführt werden, erwarten wir, dass das Ergebnis noch näher am Erwartungswert liegt. Dies kann man tatsächlich beweisen. Interessant ist dabei auch die Frage nach der Güte dieser Approximation.

### Satz 7.1 (Gesetz der großen Zahlen)

Sei  $(X_i)_{i \in \mathbb{N}}$  eine Folge von unabhängigen, identisch verteilten Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathbb{P})$  mit  $\mathbb{V}(X_i) < \infty$ . Sei

$$S_n := \frac{1}{n} \sum_{i=1}^n X_i$$

das empirische Mittel der ersten  $n$  dieser Zufallsvariablen. Die Folge  $(S_n)_{n \in \mathbb{N}}$  konvergiert gegen den Erwartungswert  $\mathbb{E}[X_1]$ , in dem Sinne dass für alle  $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(|S_n - \mathbb{E}[X_1]| > \varepsilon) = 0$$

gilt.

Der Begriff **identisch verteilt** in obigem Satz bedeutet, dass die Zufallsvariablen alle dieselbe Verteilung (mit denselben Parameterwerten) haben. Insbesondere sind also auch alle Erwartungswerte und Varianzen gleich, vgl. Aufgabe 7.1.

Wenn wir die  $y_i$  aus dem Würfel-Beispiel als Realisierungen der unabhängigen, identisch verteilten Zufallsvariablen  $X_i$  auffassen, dann bestätigt dieser Satz die Vermutung, dass der Durchschnitt der Beobachtungen gegen den Erwartungswert konvergiert.



**Beweis 7.1 (Beweis des Gesetzes der großen Zahlen)** Sei  $Y = S_n - \mathbb{E}[S_n]$ . Nach den Rechenregeln für die Varianz (Satz 5.4 gilt

$$\mathbb{V}(Y) = \mathbb{V}\left(\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[S_n]\right) = \frac{1}{n^2} \sum_{i=1}^n \mathbb{V}(X_i) = \frac{1}{n^2} n \cdot \mathbb{V}(X_1) = \frac{\mathbb{V}(X_1)}{n}.$$

Somit folgt aus der Chebyshev-Ungleichung (Satz 5.5) für jedes  $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(|S_n - \mathbb{E}[S_n]| > \varepsilon) \leq \lim_{n \rightarrow \infty} \frac{\mathbb{V}(X_1)}{n\varepsilon^2} = 0.$$

□

Theorem 7.1 wird normalerweise *schwaches Gesetz der großen Zahlen* genannt, wobei sich das Wort *schwach* auf die genaue Formulierung der Konvergenz bezieht. Es gelten auch ähnliche Aussagen für andere Formen der Konvergenz von Zufallsvariablen, sowie mit anderen Voraussetzungen, auf die hier jedoch nicht eingegangen werden sollen. Für praktische Zwecke wichtig ist die anschaulich einleuchtende Tatsache, dass die Folge der empirischen Mittel von unabhängigen und identisch verteilten Zufallsvariablen in einem geeigneten Sinne gegen den Erwartungswert konvergieren, dass also

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i = \mathbb{E}[X_1]$$

gilt. Für praktische Zwecke bedeutet das, dass für „große“  $n$

$$\frac{1}{n} \sum_{i=1}^n X_i \approx \mathbb{E}[X_1]$$

bzw.

$$\sum_{i=1}^n X_i \approx n \cdot \mathbb{E}[X_1]$$

gelten, sofern die Voraussetzungen von Satz 7.1 erfüllt sind. Was dabei „groß genau bedeutet, und wie  $\approx$  zu interpretieren ist, hängt von der Situation ab.

## 7.2 Zentraler Grenzwertsatz und Anwendungen

Aus dem Gesetz der großen Zahlen wissen wir nun, dass für unabhängige, identisch verteilte Zufallsvariablen  $\frac{1}{n} \sum_{i=1}^n X_i \approx \mathbb{E}[X_1]$  gilt. In diesem Abschnitt soll diese Approximation verbessert und eine Aussage über den „Approximationsfehler“ gemacht werden.

### Satz 7.2 (Zentraler Grenzwertsatz)

Sei  $(X_i)$  eine Folge von unabhängigen, identisch verteilten Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathbb{P})$ , mit  $\mathbb{E}[X_1] = \mu$ ,  $\mathbb{V}(X_1) = \sigma^2 \in (0, \infty)$ . Dann gilt für alle  $x \in \mathbb{R}$

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} \leq x\right) = \Phi_{0,1}(x).$$

Dabei ist  $\Phi_{0,1}$  die Verteilungsfunktion der Standardnormalverteilung, vgl. Kapitel 6.

**Beweis 7.2** Siehe z.B. Satz 15.37 in Klenke (2008).  $\square$

Der zentrale Grenzwertsatz besagt, dass für unabhängige, identisch verteilte Zufallsvariablen  $X_i$  mit endlicher Varianz die davon abgeleitete Zufallsvariable

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} = \frac{\sqrt{n}}{\sigma} (S_n - \mu)$$

für große  $n$  ungefähr *standardnormalverteilt* ist, wobei  $S_n$  definiert ist wie in Theorem 7.1. Dabei spielt die genaue Verteilung der  $X_i$  überhaupt keine Rolle! Diese können eine beliebige Verteilung haben, sofern sie für alle  $X_i$  gleich und die Varianz endlich ist. In jedem Fall ist der Grenzwert von  $\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma}$  durch die Standardnormalverteilung gegeben. Man spricht deshalb auch von einem *universellen Grenzwert*, da er in einer großen Zahl von Situationen auftritt. Um zu verstehen, warum durch  $\sqrt{n}\sigma$  dividiert werden muss, beachte man, dass  $\mathbb{V}(\sum_{i=1}^n (X_i - \mu)) = n^2 \sigma^2$  gilt (nach Satz 5.4). Satz 6.7 zeigt, dass zur Standardisierung durch die Standardabweichung, also durch  $\sqrt{n}\sigma$  geteilt werden muss.

Aus dem Gesetz der großen Zahlen sieht man, dass für unabhängige, identisch verteilte Zufallsvariablen  $X_i$  und für große  $n$  die Approximation  $\sum_{i=1}^n X_i \approx n \cdot \mathbb{E}[X_1]$  gilt, also eine Approximation für die Summe von  $n$  unabhängigen, identisch verteilten Zufallsvariablen durch das  $n$ -fache des Erwartungswertes einer der Zufallsvariablen. Der zentrale Grenzwertsatz liefert nun eine weitere Information, die diese Approximation genauer werden lässt: Es folgt

$$\sum_{i=1}^n X_i \approx n \cdot \mathbb{E}[X_1] + \sqrt{n} \cdot \sigma \cdot Y,$$

wobei  $Y$  eine standardnormalverteilte Zufallsvariable ist. In anderen Worten ist die Summe  $\sum_{i=1}^n X_i$  von unabhängigen, identisch verteilten Zufallsvariablen  $X_i$  (mit endlicher, positiver Varianz) ebenfalls stets normalverteilt mit Parametern  $n\mathbb{E}[X_1]$  und  $n\mathbb{V}(X_1)$  (vgl. Aufgabe 6.10).

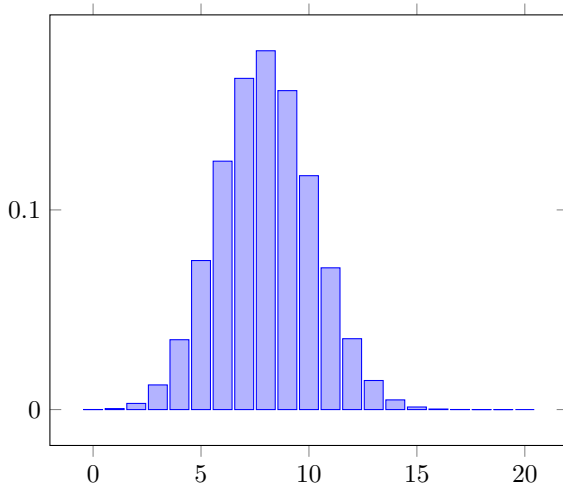
### Beispiel 7.1 (Zentraler Grenzwertsatz und Binomialverteilung)

?? Sei  $X$  eine binomialverteilte Zufallsvariable mit Parametern  $p = 0.4$  und  $n = 20$ . In Kapitel 3 hatten wir gesehen, dass man  $X$  mit Hilfe von unabhängigen, zum Parameter  $p = 0.4$  Bernoulli-verteilten Zufallsvariablen  $Y_i, i = 1, \dots, 20$  schreiben kann als

$$X = \sum_{i=1}^{20} Y_i.$$

In anderen Worten,  $X$  kann als Summe von unabhängigen, identisch verteilten Zufallsvariablen geschrieben werden – also genau die Größe, über die das Gesetz der großen Zahlen und der zentrale Grenzwertsatz Auskunft geben. Nach diesen Sätzen gilt als, dass

$$X \approx 20 \cdot \mathbb{E}[Y_1] + \sqrt{20} \cdot \sigma \cdot Y,$$

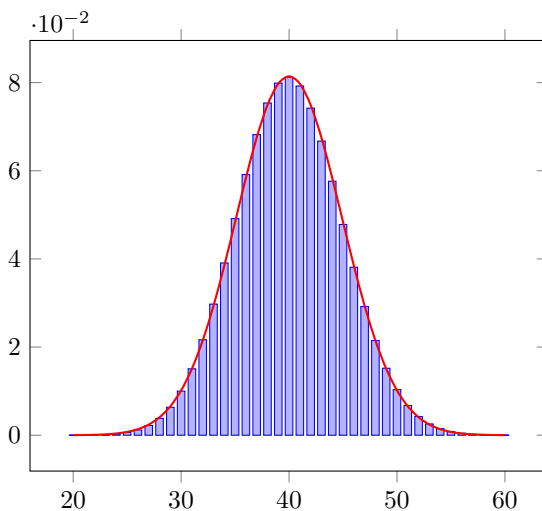
Abb. 7.2: Binomialverteilung  $\text{Bin}(20, 0.4)$ 

wobei  $\mathbb{E}[Y_1] = p = 0.4$  und  $\sigma = \sqrt{\mathbb{V}(Y_1)} = \sqrt{p(1-p)} = \sqrt{0.24}$  sind, und  $Y$  eine standardnormalverteilte Zufallsvariable ist.

Dies kann man erkennen, wenn man die Verteilungsfunktion der Binomialverteilung aufzeichnet, und die Achsen etwas streckt bzw. staucht, siehe Abbildung 7.2.

Dabei wird sichtbar, dass die Form der Binomialverteilung schon recht nah an der Gauß'schen Glockenkurve, d.h. der Dichte der Normalverteilung, liegt. Für größeres  $n$  wird dies noch deutlicher, Abbildung 7.3. ■

Als Konsequenz und Verbesserung dieser Überlegungen können wir folgenden Satz formulieren.

Abb. 7.3: Binomialverteilung  $\text{Bin}(100, 0.4)$  mit Normalverteilungsdichte  $\mathcal{N}(40, 4.9)$

**Satz 7.3 (Normalapproximation der Binomialverteilung)**

Sei  $X \sim \text{Bin}(n, p)$ . Dann gilt für  $a, b \in \{0, \dots, n\}$  mit  $a < b$

$$\mathbb{P}(a \leq X \leq b) \approx \Phi_{0,1}\left(\frac{b + 1/2 - np}{\sqrt{np(1-p)}}\right) - \Phi_{0,1}\left(\frac{a - 1/2 - np}{\sqrt{np(1-p)}}\right).$$

Dieses Ergebnis kann aus dem zentralen Grenzwertsatz hergeleitet werden, unter Berücksichtigung der Tatsache, dass die Binomialverteilung eine diskrete Verteilung ist, die Normalverteilung jedoch stetig. Für eine diskrete Verteilung gilt für  $a, b \in \mathbb{N}$  dass

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a - 1/2 \leq X \leq b + 1/2)$$

ist, was die Approximation durch die stetige Normalverteilung im Vergleich zu Beispiel ?? verbessert.

Wir wissen, dass die Approximation durch den zentralen Grenzwertsatz besser wird, je größer  $n$  ist. Im Zusammenhang mit der Binomialverteilung spielt insbesondere die Größe von  $n$  in Abhängigkeit von  $p$  eine Rolle. Als Faustregel kann verwendet werden, dass die Approximation gut ist, d.h. bis auf 2-3 Nachkommastellen mit dem tatsächlichen Wert übereinstimmt, falls  $np \geq 5$  und  $n(1-p) \geq 5$  erfüllt sind.

**Beispiel 7.2 (Normalapproximation der Binomialverteilung)**

Sei  $X$  binomialverteilt mit Parameter  $n = 90$  und  $p = 1/3$ . Es soll  $\mathbb{P}(X \in [25, 30])$  (approximativ) berechnet werden. Dabei liefert Satz 7.3

$$\begin{aligned} \mathbb{P}(25 \leq X \leq 30) &\approx \Phi_{0,1}\left(\frac{30 + 1/2 - 90 \cdot 1/3}{\sqrt{90 \cdot 1/3 \cdot 2/3}}\right) - \Phi_{0,1}\left(\frac{25 - 1/2 - 90 \cdot 1/3}{\sqrt{90 \cdot 1/3 \cdot 2/3}}\right) \\ &\approx \Phi_{0,1}(0.11) - \Phi_{0,1}(-1.23) \approx 0.4345, \end{aligned}$$

unter Verwendung der Rechenregeln für  $\Phi$  und der Normalverteilungstabelle. ■

**Zusammenfassung**

Nach dem Gesetz der großen Zahlen konvergiert das empirische Mittel von  $n$  unabhängigen, identisch verteilten Zufallsvariablen mit endlicher Varianz für  $n$  gegen  $\infty$  gegen den Erwartungswert. Der zentrale Grenzwertsatz besagt, dass in dieser Situation die Summe von zentrierten und mit  $\sqrt{n}\sigma$  skalierten Zufallsvariablen gegen eine Standardnormalverteilung konvergiert. Für große  $n$  ist damit  $\sum_{i=1}^n X_i$  annähernd gleich  $n\mathbb{E}[X_1]$  bzw. annähernd  $\mathcal{N}(n\mathbb{E}[X_1], n\mathbb{V}(X_1))$ -verteilt. Diese beiden Grenzwertsätze rechtfertigen viele Grundannahmen der Statistik, insbesondere die Allgegenwart der Normalverteilung.

## Aufgaben

**7.1** Es seien  $X_i, i \in \mathbb{N}$  unabhängige, identisch verteilte Zufallsvariablen. Zeigen Sie: Es gilt  $\mathbb{E}[X_i] = \mathbb{E}[X_1]$  und  $\mathbb{V}(X_i) = \mathbb{V}(X_1)$  für alle  $i \in \mathbb{N}$ , falls Erwartungswerte bzw. Varianz existieren.

**7.2** Es sei  $(X_n)_{n \in \mathbb{N}}$  eine Folge von unabhängigen Zufallsvariablen mit Dichte

$$f(x) = \begin{cases} \frac{1}{3}x^2, & 0 \leq x \leq 3, \\ 0 & \text{sonst.} \end{cases}$$

Berechnen Sie  $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i$ .

**7.3** Seien  $X_1, X_2, \dots, X_{10'000}$  unabhängige, exponentialverteilte Zufallsvariablen mit Parameter  $\lambda = 1.2$ . Wie können die folgenden Größen approximiert werden? a)  $X_1 + X_2 + \dots + X_{10'000}$  b)  $\mathbb{P}(X_1 + \dots + X_{10'000} \leq 8600)$ .

**7.4** Sei  $X$  eine binomialverteilte Zufallsvariable mit  $n = 4000$  und  $p = 0.6$ . Benutzen Sie die Normalapproximation zur Berechnung von a)  $\mathbb{P}(X \geq 2450)$ , b)  $\mathbb{P}(X \leq 2370)$ , d)  $\mathbb{P}(X \in [2350, 2450])$ , d)  $\mathbb{P}(\sqrt{X} > 49)$ .

**7.5** In einer Fabrik sind im Durchschnitt 4 von 100 produzierten Werkstücken nachbesserungsbedürftig. Pro Tag werden 5000 Stücke produziert. Mit welcher Wahrscheinlichkeit müssen weniger als 160 Stück nachgebessert werden?

**7.6** Es seien  $X_i, i \in \{1, \dots, n\}$  unabhängige identisch verteilte Zufallsvariablen mit  $\mathbb{E}[X_1] = \mu, \mathbb{V}(X_1) = \sigma^2$ . Leiten Sie für  $S_n := \sum_{i=1}^n X_i$  äquivalenten Formeln für die „ $\sigma$ -Regeln“ aus Beispiel 6.8 her.



## Teil II

# Einführung in die Statistik





# 8 Parameterschätzung

---

## Übersicht

8.1	Einige Grundbegriffe .....	108
8.2	Beispiele und Eigenschaften von Schätzern .....	111
8.3	Maximum Likelihood-Schätzung .....	114
8.4	Empirische Korrelation und Regressionsgerade .....	118

---

In der Einleitung zu diesem Buch wird der Begriff „Stochastik“ als Überbegriff für die beiden Gebiete *Wahrscheinlichkeitstheorie* und *Statistik* verstanden. Während wir uns bisher mit der Wahrscheinlichkeitstheorie beschäftigt haben, wird im zweiten Teil die Statistik im Vordergrund stehen. Darunter verstehen wir die Analyse und Interpretation von Messwerten und Daten aus zufälligen Vorgängen mittels wahrscheinlichkeitstheoretischen Methoden. Die Wahrscheinlichkeitstheorie bildet dafür das Fundament. In Kapitel 7 wurde bereits als wichtige theoretische Grundlage einige *Grenzwertsätze* betrachtet.

Während die Wahrscheinlichkeitstheorie von Axiomen und mathematischen Modellen ausgeht, basiert die Statistik in erster Linie auf *Daten* oder *Messwerten*. Grundsätzlich geht man in der Statistik davon aus, dass eine hinreichend große Anzahl von solchen Messwerten vorliegen. Für diese Daten soll ein geeignetes mathematisches Modell aufgestellt werden, welches diese Daten beschreibt, und welches mit wahrscheinlichkeitstheoretischen Methoden untersucht werden kann.

Dabei existieren unzählige mögliche Fragestellungen, die jeweils mit verschiedenen Methoden behandelt werden können. In diesem Buch werden nur die wichtigsten klassischen Grundprobleme behandelt:

- Punktschätzer, Bestimmung von Kenngrößen
- Konfidenzintervalle, Intervallschätzer
- Hypothesentests

Im ersten Kapitel zur Statistik widmen wir uns einem der Grundprobleme der Statistik, nämlich der Schätzung von Parametern. Das bedeutet, dass aus gemessenen Daten relevante Kenngrößen oder Verteilungsparameter der zugrundeliegenden Verteilung *geschätzt* oder approximiert werden. Dabei müssen normalerweise gewisse Annahmen getroffen werden, z.B. zur Unabhängigkeit der involvierten Zufallsvariablen.

Es gibt viele verschiedene Methoden für die Parameterschätzung. Wir betrachten hier klassische Schätzer für wichtige Kenngrößen, sowie die Methode der „Maximum Like-

lihood“-Schätzung. Ebenfalls als Schätzproblem aufgefasst werden kann die Bestimmung einer Regressionsgeraden.

### Lernziele

- Klassische Schätzer für wichtige Kenngrößen kennen
- Maximum-Likelihood-Schätzer berechnen können
- Ausgleichsgeraden mit der Methode der kleinsten Quadrate berechnen können.

## 8.1 Einige Grundbegriffe

In der Statistik stehen *Daten* im Vordergrund, aus denen mittels wahrscheinlichkeitstheoretischen Methoden Schlüsse gezogen werden sollen. Daten sind normalerweise in Form eines Vektors  $x = (x_1, \dots, x_n)$  von Messwerten  $x_i \in \mathbb{R}$  gegeben. Um die Verbindung zwischen Beobachtungen und mathematischer Theorie herzustellen, muss geklärt sein, in welcher Beziehung die erhobenen Daten zu den mathematischen Objekten stehen, die in den vorhergehenden Kapiteln dieses Buches eingeführt wurden.

**Grundannahme:** Wir gehen immer von der Gültigkeit mindestens einer der beiden folgenden Aussagen aus:

- Die gemessenen Daten sind einzelne *Realisierungen* von (unabhängigen, identisch verteilten) *Zufallsvariablen*
- Die gemessenen Daten stellen eine *Stichprobe* aus einer (noch viel größeren) *Population* dar.

Unter diesen Annahmen sollen mittels der Stichprobe Aussagen über die zugrundeliegende Zufallsvariablen bzw. über die gesamte Population gemacht werden. Man kann sich vorstellen, dass eine zu messende Größe (beispielsweise die Körpergröße von Menschen) durch Zufallsvariablen beschrieben wird, deren Verteilung im Prinzip die vollständige Information über diese Größe enthält. Solche Zufallsvariablen, oder deren Verteilung, sind jedoch nicht direkt zu beobachten und damit können wir in der Wirklichkeit nicht die vollständige Information bekommen. Was man jedoch tun kann, ist eine Anzahl von Messwerten, oder Realisierungen der Zufallsvariablen, ausgedrückt durch Zahlenwerte  $x_1, \dots, x_n$  zu erheben. Auf Basis dieser Messwerte wird nun versucht, möglichst viel Information über die zu Grunde liegenden Zufallsvariablen zu bekommen. Man kann sich vorstellen, dass die Zufallsvariablen in einem idealisierten Sinne eine Situation komplett beschreiben, wir jedoch über eine Anzahl von Messwerten nur einen Teil des Gesamtbildes erkennen können. Mittels statistischer Methoden können jedoch weiterführende (Teil-)Informationen abgeleitet werden.

Wir beginnen mit einigen Begriffen zur Beschreibung von Daten.

**Definition 8.1 (Absolute und relative Häufigkeiten)**

Sei  $(x_1, \dots, x_n) \in \mathbb{R}^n$  ein Vektor von Messwerten. Sei  $x \in \mathbb{R}$ . Die **absolute Häufigkeit** von  $x$  ist

$$H(x) := |\{i : x_i = x\}|,$$

d.h.  $H(x)$  gibt an, wie oft der Wert  $x$  im Vektor  $(x_1, \dots, x_n)$  vorkommt. Die **relative Häufigkeit** von  $x$  ist

$$h(x) := \frac{H(x)}{n}.$$

**Definition 8.2 (Klasseneinteilung, Häufigkeiten von Klassen)**

Seien  $(x_1, \dots, x_n)$  ein Vektor von Daten. Eine **Klasseneinteilung** ist eine Zusammenfassung der  $x_i$  zu disjunkten Mengen  $A_1, \dots, A_m \subseteq \mathbb{R}$  mit  $m \leq n$ , so dass jedes  $x_i$  zu genau einem  $A_j$  gehört. Gilt dabei z.B.  $x_1 = x_2$ , so müssen  $x_1$  und  $x_2$  in derselben Klasse sein. Die **absolute Häufigkeit** einer Klasse  $A$  ist

$$H(A) := |\{i : x_i \in A\}|$$

und die **relative Häufigkeit** der Klasse  $A$  bei einer Einteilung in  $m$  verschiedene Klassen ist

$$h(A) := \frac{H(A)}{n}.$$

**Beispiel 8.1 (Eine Messreihe)**

In der Produktion eines elektronischen Geräts wird in einer Testreihe die Zeit (in Sekunden, gerundet auf eine Nachkommastelle) vom Start des Geräts bis zur Betriebsbereitschaft gemessen und folgende Daten erhalten:

$$(10.9, 6.8, 9.5, 6.9, 8.2, 3.4, 6.2, 8.6, 5.3, 10.7, 8.1, 8.0, 8.9, 10.7)$$

Wir berechnen die absoluten und relativen Häufigkeiten für einige Werte. Absolute Häufigkeiten:

$$H(10.9) = 1, \quad H(10.7) = 2, \quad H(4.0) = 0.$$

Relative Häufigkeiten:

$$h(10.9) = \frac{1}{14}, \quad h(10.7) = \frac{1}{7}, \quad h(4.0) = 0.$$



Falls diese Daten  $(x_1, \dots, x_n)$  Realisierungen von *diskreten* Zufallsvariablen sind, so beschreiben die Daten ein **diskretes Merkmal**. Sind es Realisierungen von Zufallsvariablen, welche im Prinzip sämtliche Werte von  $\mathbb{R}$  oder zumindest sämtliche Werte mindestens eines Teilintervalles von  $\mathbb{R}$  annehmen können, so beschreiben die Daten ein **stetiges Merkmal**.

Stetige Merkmale oder diskrete Merkmale mit einer großen Anzahl von Ausprägungen machen üblicherweise eine Klasseneinteilung notwendig. Typische Beispiele für stetige Merkmale sind Längen, Gewicht, Zeitdauer, Spannung, usw. Diskrete Merkmale sind z.B. Teilnehmerzahlen, Anzahl Probanden mit einer bestimmten Krankheit, oder auch bereits diskretisierte, aber ursprünglich stetige Merkmale, z.B. Lebensalter in Jahren, Gewicht in Kilogramm.

Wie bei Zufallsvariablen gibt es auch für Datenmengen verschiedene Kenngrößen, welche Aufschluss über gewisse Eigenschaften. Dadurch kann die Datenmenge oft reduziert werden.

**Definition 8.3 (Empirisches Mittel)**

Sei  $(x_1, \dots, x_n) \in \mathbb{R}^n$  ein Vektor (von Messwerten bzw. Daten). Das **empirische Mittel** von  $(x_1, \dots, x_n)$  ist definiert als

$$\bar{\mu}_n = \bar{\mu}_n(x_1, \dots, x_n) := \frac{1}{n} \sum_{i=1}^n x_i$$

◆

**Definition 8.4 (Median)**

Sei  $(x_1, \dots, x_n) \in \mathbb{R}^n$  ein Vektor (von Messwerten bzw. Daten). Der **Median** von  $(x_1, \dots, x_n)$  ist definiert als der Wert in der Mitte der geordneten Liste. Falls  $n$  gerade ist, wird der Durchschnitt aus den beiden mittleren Werten gebildet. ◆

**Definition 8.5 (Empirische Varianz)**

Sei  $(x_1, \dots, x_n) \in \mathbb{R}^n$  ein Vektor (von Messwerten/Daten). Die **empirische Varianz** von  $(x_1, \dots, x_n)$  ist definiert als

$$\bar{\sigma}_n^2 = \bar{\sigma}_n^2(x_1, \dots, x_n) := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{\mu}_n)^2.$$

Die **empirische Standardabweichung** ist  $\bar{\sigma}_n := \sqrt{\bar{\sigma}_n^2}$ . ◆

Das empirische Mittel ist also nichts anderes als das arithmetische Mittel oder der Durchschnittswert der Daten. Die empirische Varianz ist, analog zur Varianz von Zufallsvariablen, ein Maß für die Streuung, bzw. für die durchschnittliche Abweichung vom Mittelwert. Zu beachten ist, dass zwar  $n$  Werte aufsummiert werden, jedoch nur durch  $n-1$  dividiert wird. Den Grund dafür werden wir später sehen.

**Beispiel 8.2 (Fortsetzung von Beispiel 8.1)**

Wir betrachten die Daten von Beispiel 8.1. Einsetzen in die Definitionen ergibt für das empirische Mittel

$$\bar{\mu}_n = 8.014,$$

und für die empirische Varianz

$$\bar{\sigma}_n^2 = 4.677.$$

Zur Bestimmung des Medians ordne man zuerst die Daten in aufsteigender Reihenfolge. Als Vektor geschrieben erhält man

$$(3.4, 5.3, 6.2, 6.8, 6.9, 8.0, 8.1, 8.2, 8.6, 8.9, 9.5, 10.7, 10.7, 10.9)$$

Insgesamt gibt es 14 Werte, also eine gerade Zahl. Die beiden Werte in der Mitte der geordneten Liste sind 8.1 und 8.2. Somit ist der Median

$$\frac{8.1 + 8.2}{2} = 8.15.$$

■

Es gibt Situationen, in denen das empirische Mittel durch einige wenige *Ausreißer*, also extrem hohe oder extrem tiefe Werte, dominiert wird, aber die Mehrheit der Werte weit vom Durchschnitt entfernt sind. In solchen Fällen ist der Median oft aussagekräftiger als das empirische Mittel.

## 8.2 Beispiele und Eigenschaften von Schätzern

Gegeben seien Messwerte  $(x_1, \dots, x_n)$ . Wie bisher nehmen wir an, dass es sich dabei um Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen  $X_1, \dots, X_n$  handelt, über deren Verteilung wir jedoch nur unvollständige Informationen haben. Unser erstes Ziel ist es, aus den gemessenen Daten Informationen über Kenngrößen oder Parameter der Verteilung der  $X_i$  zu gewinnen. Wie bisher verwenden wir die Konvention, dass Messwerte  $x_1, \dots, x_n, x_i \in \mathbb{R}$  mit Kleinbuchstaben, Zufallsvariablen mit den entsprechenden Großbuchstaben bezeichnet werden.

### Definition 8.6 (Schätzfunktion)

Seien  $X_1, \dots, X_n$  Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathbb{P})$ . Eine **Schätzfunktion** zur Stichprobengröße  $n$  ist eine Funktion  $\theta_n : \mathbb{R}^n \rightarrow \mathbb{R}$ , welche die Zufallsvariablen  $(X_1, \dots, X_n)$  auf  $\theta_n(X_1, \dots, X_n)$  abbildet. ♦

In der obigen Definition setzen wir in die Schätzfunktion  $\theta_n$  die Zufallsvariablen  $X_1, \dots, X_n$  ein. Somit ist der Wert  $\theta_n(X_1, \dots, X_n)$  wieder eine Zufallsvariable, deren Verteilung von der Verteilung des Zufallsvektors  $(X_1, \dots, X_n)$  abhängt. Setzen wir statt der Zufallsvariablen die Messwerte  $(x_1, \dots, x_n)$  ein, ist  $\theta_n(x_1, \dots, x_n)$  eine reelle Zahl.

Üblicherweise setzt man aber eine Schätzfunktion in Verbindung zu einem (in der konkreten Situation normalerweise unbekannten) Parameter der Verteilung der  $X_i$ . Dabei wird angenommen, dass die  $X_1, \dots, X_n$  alle dieselbe Verteilung haben, welche von einem unbekannten Parameter  $\theta$  abhängt. Hat man  $\theta_n$  konstruiert, dient

$$\bar{\theta}_n := \theta_n(x_1, \dots, x_n)$$

als *Schätzer* oder Approximation für den unbekannten Parameter  $\theta$ . Damit letzteres der Fall ist, sollte eine Schätzfunktion natürlich einen Zusammenhang zum unbekannten Parameter der Verteilung haben. Um die Abhängigkeit der Verteilung vom Parameter  $\theta$  zu betonen, schreiben wir  $\mathbb{P}_\theta$  für das Wahrscheinlichkeitsmaß im zugrunde gelegten Wahrscheinlichkeitsraum, und analog  $\mathbb{E}_\theta, \mathbb{V}_\theta$ .

Wir geben nun einige Beispiele für klassische Schätzer an.

### Beispiel 8.3 (Empirisches Mittel)

In Definition 8.3 hatten wir das empirische Mittel  $\bar{\mu}_x$  eines Datenvektors  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$  betrachtet, definiert als

$$\bar{\mu}_n = \bar{\mu}_n(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i.$$

Als Funktion von  $X_1, \dots, X_n$  ist  $\bar{\mu}_n$  eine Schätzfunktion. Falls die Daten Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen  $X_1, \dots, X_n$  mit  $\mathbb{E}[X_1] = \dots = \mathbb{E}[X_n] = \mu$  sind, so gilt

$$\bar{\mu}_n(x_1, \dots, x_n) \approx \mu,$$

zumindest für hinreichend große  $n$ , da laut dem Gesetz der großen Zahlen für  $n \rightarrow \infty$

$$\bar{\mu}_n(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i(\omega) \rightarrow \mathbb{E}[X_1] = \mu$$

gilt. Außerdem gilt

$$\mathbb{E}[\bar{\mu}_n(X_1, \dots, X_n)] = \frac{1}{n} \sum_{i=1}^n n \mathbb{E}[X_i] = \mu.$$

Somit ist  $\bar{\mu}_n$  wie gewünscht eine Größe, welche direkt aus den Daten  $x_1, \dots, x_n$  berechnet werden kann, und einen Schätzwert für den Erwartungswert der zugrundeliegenden Zufallsvariablen liefert. ■

### Beispiel 8.4 (Empirische Varianz)

Die empirische Varianz von  $x = (x_1, \dots, x_n)$  wurde definiert (vgl. Def. 8.5) als

$$\bar{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{\mu})^2.$$

Falls die  $x_i$  Realisierungen von unabhängigen identisch verteilten Zufallsvariablen  $X_1, \dots, X_n$  sind, d.h. so kann man mittels einiger Umformungen zeigen, dass

$$\mathbb{E}[\bar{\sigma}_n^2(X_1, \dots, X_n)] = \mathbb{E}\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{\mu}_n(X_1, \dots, X_n))^2\right] = \mathbb{V}(X_1)$$

und

$$\lim_{n \rightarrow \infty} \bar{\sigma}_n^2(X_1, \dots, X_n) = \mathbb{V}(X_1)$$

gelten. Auch hier haben wir es wieder mit einer aus den Daten berechenbaren Größe zu tun, welche einen Zusammenhang mit der Varianz der zugrundeliegenden Zufallsvariablen hat. Die empirische Varianz ist deshalb ein Schätzer für die Varianz. ■

Neben einem Zusammenhang mit dem zu schätzenden Wert ist es oft wünschenswert, dass Schätzer eine oder mehrere weitere nützliche Eigenschaften haben.

**Definition 8.7 (Schätzer mit günstigen Eigenschaften)**

Sei  $(x_1, \dots, x_n) \in \mathbb{R}^n$  ein Vektor von Daten, welche als Realisierungen von identisch verteilten Zufallsvariablen  $X_1, \dots, X_n$  mit Parameter  $\theta$  auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathbb{P}_\theta)$ . Sei  $\theta_n$  eine Schätzfunktion zur Stichprobengröße  $n$ . Dann ist  $\theta_n$  ein **erwartungstreuer Schätzer** für  $\theta$ , falls

$$\mathbb{E}_\theta[\theta_n(X_1, \dots, X_n)] = \theta$$

ist. Statt erwartungstreu verwendet man auch die Bezeichnungen **unverzerrt** oder (englisch) **unbiased**.  $\theta_n$  ist ein **konsistenter Schätzer** für  $\theta$ , falls

$$\lim_{n \rightarrow \infty} \theta_n(X_1, \dots, X_n) = \theta$$

gilt.  $\theta_n$  ist ein **effizienter Schätzer** für  $\theta$ , falls

$$\lim_{n \rightarrow \infty} \mathbb{V}_\theta(\theta_n(X_1, \dots, X_n)) = 0$$

gilt. ◆

**Beispiel 8.5 (Empirisches Mittel und empirische Varianz)**

Laut Beispiel 8.3 und Beispiel 8.4 sind das empirische Mittel ein erwartungstreu und die empirische Varianz erwartungstreu und konsistente Schätzer. Die Erwartungstreue der empirischen Varianz ist auch der Grund, weshalb die Normierung mit  $n-1$  und nicht mit  $n$  gemacht wird. Betrachtet man stattdessen

$$\hat{\sigma}_n^2 := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{\mu})^2,$$

so kann man zeigen, dass dies auch ein konsistenter Schätzer für die Varianz ist, er ist jedoch nicht erwartungstreu. ■

**Beispiel 8.6**

Auch wenn die Eigenschaften aus Satz 8.7 im Allgemeinen wünschenswert sind, so sind sie keine Garantie für einen guten Schätzer. Beispielsweise können wir als Schätzer für den Erwartungswert einer Stichprobe  $x_1, \dots, x_n$  auch einfach den ersten Wert  $x_1$  auswählen, also  $\theta_n(x_1, \dots, x_n)$  für alle  $n$ . Dies ist ein erwartungstreu Schätzer, denn

$$\mathbb{E}[\theta_n(X_1, \dots, X_n)] = \mathbb{E}[X_1].$$

Er ist jedoch z.B. nicht konsistent, es sei denn,  $X_1$  ist konstant. Es ist relativ naheliegend, dass ein einzelner Messwert außer in sehr speziellen Situationen kein besonders guter Schätzer ist.

Andererseits gibt es günstige Schätzer, welche beispielsweise nicht erwartungstreu sind. So ist z.B.  $\hat{\sigma}_n^2$  aus Beispiel 8.5 in vielen Fällen ein durchaus vernünftiger Schätzer für die Varianz. Weitere Beispiele findet man z.B. in Georgii (2009). ■

In vielen Situationen ist es sinnvoll anzunehmen, dass der unbekannte Parameter  $\theta$  aus einer gewissen Menge  $\Theta$  von möglichen Parametern stammt, beispielsweise aus  $\Theta = \mathbb{R}$  oder  $\Theta = [0, 1]$ . Die Menge aller möglichen Wahrscheinlichkeitsmaße in Definition 8.6 ist also  $(\mathbb{P}_\theta)_{\theta \in \Theta}$ . Die Familie  $(\Omega, (\mathbb{P}_\theta)_{\theta \in \Theta})$  bezeichnet man auch als *statistisches Modell*. Wir werden hier etwas weniger formal sein und den Begriff nicht weiter verwenden.

### 8.3 Maximum Likelihood-Schätzung

Die Maximum Likelihood-Schätzung ist eine spezielle Methode der Parameterschätzung, bei der unter allen theoretisch möglichen Werten eines Parameters einer Verteilung derjenige gesucht wird, für den die Wahrscheinlichkeit maximal wird, die beobachteten Messwerte  $x_1, \dots, x_n$  zu finden. Die Messwerte seien Realisierungen von unabhängigen identisch verteilten Zufallsvariablen. Dabei sei die Form der Verteilung (z.B. Poisson) bekannt, aber ein Parameter unbekannt. Sei  $\Theta \subseteq \mathbb{R}$  die Menge aller theoretisch möglichen Parameter (z.B. ist  $\Theta = ]0, \infty[$  im Fall des unbekannten Parameters  $\lambda$  der Poisson-Verteilung). Dann kann die Funktion

$$L : \Theta \rightarrow [0, 1] : L(\theta) = \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n)$$

betrachtet werden. Dabei ist zu beachten, dass  $x_1, \dots, x_n$  die Messwerte sind, also bekannte, feste Werte. Die Funktion  $L$  ordnet also den Messwerten die Wahrscheinlichkeit zu, mit der sie auftreten, falls der echte Parameter  $\theta$  ist. Die Idee besteht nun darin, dasjenige  $\theta$  zu bestimmen, für welches  $L(\theta)$  maximal wird, bei den gemessenen Werten  $x_1, \dots, x_n$ , also denjenigen Parameter, welcher die beobachtete Messung besonders wahrscheinlich macht. Im Falle von nicht-diskreten Zufallsvariable muss die Funktion  $L$  etwas modifiziert werden, da sie sonst konstant 0 wäre. Wir beschreiben dieses Verfahren nun formal, einerseits für diskrete Zufallsvariablen, andererseits für Zufallsvariablen mit Dichte.

#### Definition 8.8 (Likelihood-Funktion)

Seien  $x_1, \dots, x_n \in \mathbb{R}$  Messwerte, welche als Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen  $X_1, \dots, X_n$  aufgefasst werden können. Dabei nehmen wir an, dass die Zufallsvariablen entweder diskret sind oder eine Dichte besitzen, wobei Verteilung bzw. Dichte jeweils bekannt sind, aber von einem (unbekannten) Parameter  $\theta$  abhängen. Die **Likelihood-Funktion** ist die Funktion  $L = L((x_1, \dots, x_n); \theta)$  von  $\mathbb{R}^n \times \mathbb{R}$  nach  $[0, \infty[$ , für welche gilt:

- Falls die  $X_1, \dots, X_n$  eine diskrete Verteilung  $p_\theta$  mit Parameter  $\theta$  besitzt, also  $\mathbb{P}_\theta(X_i = x) = p_\theta(x)$  für  $i \in \{1, \dots, n\}, x \in X_i(\Omega)$  gilt, so ist  $L$  definiert als

$$L((x_1, \dots, x_n); \theta) = p_\theta(x_1) \cdot \dots \cdot p_\theta(x_n) = \prod_{i=1}^n p_\theta(x_i) = \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n).$$



- Falls  $X_i$  eine Dichte  $f_\theta$  mit Parameter  $\theta$  besitzt, so ist  $L$  definiert als

$$L((x_1, \dots, x_n); \theta) = f_\theta(x_1) \cdot \dots \cdot f_\theta(x_n) = \prod_{i=1}^n f_\theta(x_i).$$

◆

Eine Likelihood-Funktion  $L$  hängt also von  $n + 1$  Werten ab: Den  $n$  Messwerten, sowie dem (normalerweise unbekannten) Parameter  $\theta$ . In der Notation fassen wir dabei die Messwerte zu einem Vektor zusammen.

**Definition 8.9 (Maximum Likelihood-Schätzer, MLE)**

Seien  $x_1, \dots, x_n$  Messwerte, welche Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen sind, welche entweder der diskreten Verteilung  $p_\theta$  folgen, oder die Dichte  $f_\theta$  haben. Der **Maximum Likelihood-Schätzer** für  $\theta$  ist  $\theta_* = \theta_*(x_1, \dots, x_n)$ , welches die Likelihood-Funktion  $L((x_1, \dots, x_n); \theta)$  unter allen  $\theta \in \mathbb{R}$  maximiert. Formal kann man schreiben

$$\theta_*(x_1, \dots, x_n) = \operatorname{argmax}_{\theta \in \mathbb{R}} L((x_1, \dots, x_n); \theta).$$

In anderen Worten ist  $\theta_*$  derjenige Parameter, für welchen

$$L((x_1, \dots, x_n); \theta_*) \geq L((x_1, \dots, x_n); \theta) \quad \forall \theta$$

gilt. Der Maximum Likelihood-Schätzer wird dabei manchmal auch kurz als MLE bezeichnet, für das englische „maximum likelihood estimator“. ◆

Die Bestimmung eines Maximum Likelihood Schätzers geschieht nun nach folgendem Verfahren: Gegeben sind Daten  $x_1, \dots, x_n$ , von denen man weiß oder annimmt, dass sie Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen sind, deren Verteilung bis auf den unbekannten Parameter  $\theta$  bekannt ist. Danach geht man folgendermaßen vor:

1. Likelihood-Funktion aufstellen (nach Definition 8.8, abhängig davon ob die Verteilung diskret ist oder eine Dichte besitzt)
2. Den Parameter  $\theta$  so bestimmen, dass  $L((x_1, \dots, x_n); \theta)$  für das vorgegebene  $x = (x_1, \dots, x_n)$  maximal ist, d.h. die Funktion  $\theta \mapsto L((x_1, \dots, x_n); \theta)$  maximieren. Dabei ist lediglich die Stelle  $\theta_*$  des Maximums relevant, nicht der Wert  $L((x_1, \dots, x_n); \theta_*)$ .

Aus der Analysis ist bekannt, dass Maxima einer differenzierbaren Funktion durch Bestimmung der Nullstellen der ersten Ableitung dieser Funktion bestimmt werden können, wobei jeweils noch überprüft werden muss, dass die gefundenen Werte tatsächlich Maxima sind (z.B. durch den bekannten Test mit der zweiten Ableitung). Um einen MLE-Schätzer zu bestimmen, muss die Likelihood-Funktion nach  $\theta$  abgeleitet werden (und nicht etwa nach den  $x_i$ ). In vielen Fällen ist es jedoch so, dass die Likelihood-Funktion etwas aufwändig abzuleiten ist, hingegen eine einfache Transformation davon viel leichter zu behandeln ist.

**Definition 8.10**

Sei  $L((x_1, \dots, x_n); \theta)$  eine Likelihood-Funktion. Die **Log-Likelihood-Funktion** ist definiert als

$$l((x_1, \dots, x_n); \theta) = \ln L((x_1, \dots, x_n); \theta).$$

Hierbei bezeichnet  $\ln$  den natürliche Logarithmus. ♦

**Satz 8.1**

$l((x_1, \dots, x_n); \theta)$  ist genau dann maximal, wenn  $L((x_1, \dots, x_n); \theta)$  maximal ist.

**Beweis 8.1** Dies folgt aus der Tatsache, dass der Logarithmus eine streng monoton wachsende Funktion  $(0, 1] \rightarrow \mathbb{R}$  ist. □

Somit wird für praktische Zwecke das obige Vorgehen oft zu

1. Likelihood-Funktion aufstellen (nach Definition 8.8, abhängig davon ob die Verteilung diskret ist oder eine Dichte besitzt)
2. Die log-Likelihood-Funktion berechnen, und möglichst weit vereinfachen (durch Anwendung der Logarithmengesetze)
3. Die Stelle  $\theta_*$  des Maximums der log-Likelihood-Funktion  $l((x_1, \dots, x_n); \theta)$  durch Ableiten nach  $\theta$  bestimmen (und die Maximalität überprüfen).

**Beispiel 8.7 (Poisson-Verteilung)**

Wir haben Messwerte  $(x_1, \dots, x_n)$  von denen wir wissen, dass sie Poisson-verteilt sind, wobei wir jedoch den Parameter  $\lambda$  nicht kennen. Somit haben wir die Verteilung

$$p_\lambda(k) = \frac{e^{-\lambda} \lambda^k}{k!},$$

der zu schätzende Parameter  $\theta$  ist also einfach gleich Parameter  $\lambda$  der Poisson-Verteilung. Entsprechend dem oben beschriebenen Vorgehen stellen wir die Likelihood-Funktion auf:

$$L((x_1, \dots, x_n); \theta) = \prod_{i=1}^n p_\lambda(x_i) = \prod_{i=1}^n \frac{e^{-\theta} \theta^{x_i}}{x_i!} = (e^{-\theta})^n \cdot \prod_{i=1}^n \frac{\theta^{x_i}}{x_i!}.$$

Diese Funktion sollen wir jetzt bezüglich  $\theta$  maximieren. Wir berechnen zuerst die log-Likelihood-Funktion, und erhalten nach Anwendung der Logarithmengesetze

$$\begin{aligned} l((x_1, \dots, x_n); \theta) &= \ln L((x_1, \dots, x_n); \theta) = \ln \left[ (e^{-\theta})^n \cdot \prod_{i=1}^n \frac{\theta^{x_i}}{x_i!} \right] \\ &= \ln e^{-n \cdot \theta} + \sum_{i=1}^n (x_i \cdot \ln \theta - \ln(x_i!)) \\ &= -n \cdot \theta + \ln \theta \cdot \sum_{i=1}^n x_i - \sum_{i=1}^n \ln(x_i!). \end{aligned}$$

Diese Funktion ist nun leicht nach  $\theta$  abzuleiten, unter anderem weil die Summen jeweils nicht von  $\theta$  abhängt. Wir erhalten

$$\frac{\partial l((x_1, \dots, x_n); \theta)}{\partial \theta} = -n + \frac{1}{\theta} \cdot \sum_{i=1}^n x_i$$

Setzen wir die Ableitung gleich null und lösen nach  $\theta$  auf, so finden wir die einzige Nullstelle

$$\theta_* = \frac{\sum_{i=1}^n x_i}{n}.$$

Dies ist also ein Kandidat für das Maximum der Likelihood-Funktion. Wir müssen nun noch überprüfen, dass es sich hier tatsächlich um ein Maximum (und nicht etwa um ein Minimum oder einen Sattelpunkt) handelt. Berechnen der zweiten Ableitung ergibt

$$\frac{\partial^2 l((x_1, \dots, x_n); \theta)}{\partial \theta^2} = -\frac{1}{\theta^2} \cdot \sum_{i=1}^n x_i.$$

Einsetzen von  $\theta_* = \frac{\sum_{i=1}^n x_i}{n}$  führt auf

$$\frac{\partial^2 l((x_1, \dots, x_n); \theta)}{\partial \theta^2} = -\frac{n^2}{\sum_{i=1}^n x_i} < 0,$$

denn als Realisierungen einer Poisson-Verteilung ist  $x_i \geq 0$  für alle  $i = 1, \dots, n$ . Somit handelt es sich um ein Maximum.

Es fällt auf, dass der Maximum-Likelihood-Schätzer gegeben ist durch das empirische Mittel der Messwerte:

$$\lambda_* = \theta_* = \frac{\sum_{i=1}^n x_i}{n} = \bar{\mu}_n(x_1, \dots, x_n).$$

Dies überrascht möglicherweise nicht, wenn man sich daran erinnert, dass der Erwartungswert der Poisson-Verteilung durch den Verteilungsparameters gegeben ist.

Sind die Messwerte nun konkret gegeben, z.B.  $(x_1, \dots, x_{10}) = (5, 3, 4, 3, 2, 9, 3, 1, 7, 3)$ , so ergibt Einsetzen den Maximum-Likelihood-Schätzer  $\lambda_* = \bar{\mu}_{10} = 4$ . ■

### Beispiel 8.8 (Normalverteilung)

Wir nehmen an, dass unsere Messwerte  $(x_1, \dots, x_n)$  normalverteilt sind. Bekanntlich hat die Normalverteilung die Dichte

$$\varphi_{\mu, \sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Insbesondere haben wir es hier mit zwei Parametern zu tun,  $\mu$  und  $\sigma^2$ . Obwohl es auch Möglichkeiten gibt, zwei Parameter gleichzeitig zu schätzen, nehmen wir für dieses Beispiel an, dass  $\mu = 0$  gilt (z.B. weil wir wissen dass die gemessenen Daten symmetrisch um 0 verteilt sind). Somit haben wir die Dichte

$$\varphi_{0, \sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

mit einem unbekannten Parameter  $\theta = \sigma^2$ , den wir mit Hilfe der Maximum-Likelihood-Methode schätzen wollen. Für die Likelihood-Funktion erhalten wir

$$L((x_1, \dots, x_n); \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x_i^2}{2\theta}} = \frac{1}{(\sqrt{2\pi\theta})^n} e^{-\sum_{i=1}^n \frac{x_i^2}{2\theta}},$$

wobei wir im letzten Schritt die Potenzgesetze verwendet haben. Logarithmieren führt auf

$$l((x_1, \dots, x_n); \theta) = -n \cdot \ln(\sqrt{2\pi\theta}) - \frac{\sum_{i=1}^n x_i^2}{2\theta} = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \theta - \frac{\sum_{i=1}^n x_i^2}{2\theta}.$$

Dies leiten wir nun nach  $\theta$  ab und erhalten

$$\frac{\partial l((x_1, \dots, x_n); \theta)}{\partial \theta} = -\frac{n}{2} \cdot \frac{1}{\theta} - \frac{\sum_{i=1}^n x_i^2}{2} \cdot \frac{-1}{\theta^2} = \frac{n}{2\theta^2} \left( -\theta + \frac{1}{n} \sum_{i=1}^n x_i^2 \right).$$

Die (einzige) Nullstelle der ersten Ableitung ist also

$$\theta_* = \frac{\sum_{i=1}^n x_i^2}{n}.$$

Der Test mit der zweiten Ableitung ergibt

$$\frac{\partial^2 l((x_1, \dots, x_n); \theta)}{\partial \theta^2} = \frac{1}{\theta^2} \cdot \frac{n}{2} - \frac{\sum_{i=1}^n x_i^2}{2} \frac{2}{\theta^3},$$

und an der Stelle  $\theta = \theta_*$  wird dieser Ausdruck zu

$$\frac{n^3}{2(\sum_{i=1}^n x_i^2)^2} - \frac{n^3}{(\sum_{i=1}^n x_i^2)^2} < 0.$$

Somit handelt es sich tatsächlich um ein Maximum.

Wir haben also als Maximum-Likelihood-Schätzer für die Varianz der Normalverteilung

$$\sigma_*^2 = \theta_* = \frac{\sum_{i=1}^n x_i^2}{n}.$$

Dabei fällt auf, dass dies *nicht* die empirische Varianz ist. ■

## 8.4 Empirische Korrelation und Regressionsgerade

Das Statistikprogramm R enthält mehrere Beispieldatensätze, darunter einen Datensatz des Yellowstone-Geysirs „Old faithful“. Dieser beinhaltet Daten zweier gleichzeitig gemessener Größen, d.h. Paare von Daten  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , welche die Wartezeit und die Dauer eines Ausbruchs des Geysirs angeben. Mit dem Befehl `faithful` können die 272 Datenpaare in R aufgerufen werden. Das Paar  $(\bar{\mu}_x, \bar{\mu}_y)$  wird auch *Schwerpunkt* der Daten genannt, und kann in den Plot der Datenpaare als Punktwolke eingezeichnet werden (rotes Quadrat). Die Form der Punktwolke weist darauf hin, dass eine gewisse Abhängigkeit zwischen den Daten vorhanden ist. Diese Abhängigkeit soll nun quantifiziert werden.

### Definition 8.11 (Empirische Kovarianz)

Seien Datenvektoren  $x = (x_1, \dots, x_n)$  und  $y = (y_1, \dots, y_n)$  von derselben Länge  $n$  gegeben, welche Paaren von Daten  $(x_1, y_1), \dots, (x_n, y_n)$  entsprechen. Die **empirische Kovarianz** ist definiert als

$$\bar{c}_{xy} := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{\mu}_x)(y_i - \bar{\mu}_y),$$

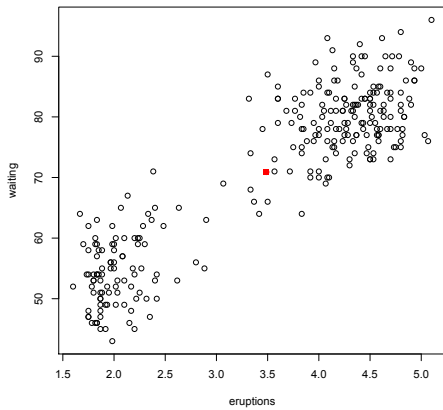


Abb. 8.1: Punktplot des R-Datensatzes `faithful`

wobei  $\bar{\mu}_x$  das empirische Mittel der  $x_i$  und  $\bar{\mu}_y$  das empirische Mittel der  $y_i$  ist.

Der **empirische Korrelationskoeffizient** ist definiert als

$$\bar{r}_{xy} = \frac{\bar{c}_{xy}}{\bar{\sigma}_x \bar{\sigma}_y},$$

wobei  $\bar{\sigma}_x := \sqrt{\bar{\sigma}_x^2}$  die *empirische Standardabweichung* von  $x$  ist (und analog  $\bar{\sigma}_y$  für  $y$ ). ◆

Man beachte, dass die Reihenfolge der Einträge der Vektoren  $x$  und  $y$  von Bedeutung ist. Insbesondere ist die Definition nur dann sinnvoll, wenn Eintrag  $x_i$  dem Eintrag  $y_i$  entspricht, also wie im Beispiel das Paar  $(x_i, y_i)$  z.B. zwei verschiedenen gemessenen Größen (Dauer und Wartezeit) derselben Messung (Ausbruch) entsprechen. Der Korrelationskoeffizient erfüllt  $-1 \leq \bar{r}_{x,y} \leq 1$  und ist ein Maß für die Stärke des linearen Zusammenhangs zwischen  $x$  und  $y$ .

### Beispiel 8.9 (Empirische Korrelation)

Im Beispieldatensatz `faithful` erhalten wir durch Einsetzen in die Definition, bzw. durch Aufrufen des R-Befehls `cov(x,y)` für die empirische Kovarianz

$$\bar{c}_{x,y} = 13.97781,$$

und mit `cor(x,y)` für die empirische Korrelation

$$\bar{r}_{x,y} = 0.9008112.$$

Dieser Wert ist relativ nah an 1, man kann also einen linearen Zusammenhang zwischen den beiden Werten vermuten. ■

Auf der Verbindung zwischen empirischer Kovarianz und linearem Zusammenhang beruht auch das Prinzip der linearen Regression. Diese kommt zur Anwendung, wenn Daten  $x = (x_1, \dots, x_n), y = (y_1, \dots, y_n)$  gegeben sind, zwischen denen ein paarweiser Zusammenhang von einer bestimmten Form (z.B. linear, quadratisch) vermutet wird, also  $y_i = f(x_i), i \in \{1, \dots, n\}$  für eine geeignete Funktion  $f$ . Dann möchte man die Daten in geeigneter Weise durch eine Kurve, z.B. eine Gerade oder eine Parabel, approximieren. Betrachten wir den Fall eines (vermuteten) linearen Zusammenhangs. Dann möchte man eine Gerade

$$y = ax + b,$$

welche durch den Schwerpunkt  $(\bar{\mu}_x, \bar{\mu}_y)$  der Daten verläuft, und welche die Abstände der Datenpunkte  $(x_1, y_1), \dots, (x_n, y_n)$  von der Geraden minimiert (unter allen Geraden durch den Schwerpunkt).

Konkret wird dabei normalerweise die Summe der quadratischen Abstände in  $y$ -Richtung minimiert. Genauer sucht man zwei Zahlen  $a, b \in \mathbb{R}$ , so dass die Gerade mit der Gleichung

$$y = ax + b$$

zwei Bedingungen erfüllt:

**(LR1)**  $(\bar{\mu}_x, \bar{\mu}_y)$  liegt auf der Geraden, also  $\bar{\mu}_y = a \cdot \bar{\mu}_x + b$ ,

**(LR2)** Der Ausdruck

$$\sum_{i=1}^n r_i^2$$

ist minimal, wobei  $r_i := y_i - ax_i - b$  der Abstand in  $y$ -Richtung von  $(x_i, y_i)$  zur Geraden ist.

### Satz 8.2 (Lineare Regression)

Die beiden Bedingungen (LR1) und (LR2) legen  $a$  und  $b$  eindeutig fest, und zwar als

$$a = \frac{\bar{c}_{xy}}{\bar{\sigma}_x^2}, \quad b = \bar{\mu}_y - a \cdot \bar{\mu}_x.$$

**Beweis 8.2** Aus (LR1) folgt sofort  $b = \bar{\mu}_y - a \cdot \bar{\mu}_x$ . Somit ist  $r_i = y_i - ax_i - b = y_i - ax_i - \bar{\mu}_y + a \cdot \bar{\mu}_x$ , und die zu minimierende Größe ist

$$\sum_{i=1}^n (y_i - \bar{\mu}_y + a(x_i - \bar{\mu}_x))^2.$$

Dies kann als Funktion von  $a$  aufgefasst werden, und mit der üblichen Methode (ableiten und nullsetzen) minimiert werden. Umformen ergibt dann  $a = \frac{\bar{c}_{xy}}{\bar{\sigma}_x^2}$ . Wir verzichten auf die Details.  $\square$

Zu beachten ist, dass die Rollen von  $x$  und  $y$  nicht symmetrisch sind.

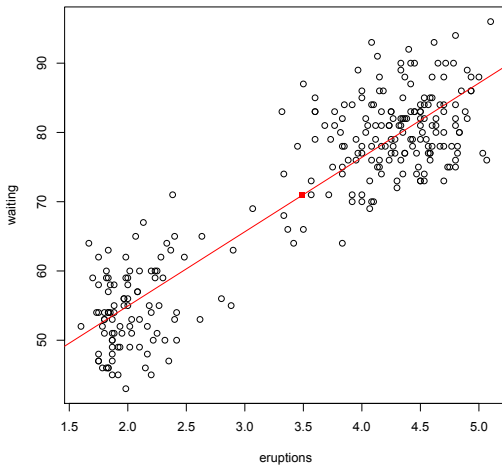


Abb. 8.2: Ausgleichsgerade für den Beispieldatensatz `faithful`

### Beispiel 8.10 (Lineare Regression, Fortsetzung von Beispiel 8.9)

Berechnet man die Regressionsgerade für den Datensatz aus Beispiel 8.9 mittels der Formel aus Satz 8.2, so erhält man In R wird mit dem Befehl `lm(y ~ x)` eine lineare Regression direkt berechnet. Angewandt auf die Daten aus Beispiel 8.9 erhält man für den  $y$ -Achsenabschnitt  $b \approx 33.4744$  und die Steigung  $a \approx 10.7296$ . Mittels des Befehls `lm(y ~ x)` kann dies auch in R ausgeführt werden, die Ausgabe erfolgt mittels des Befehls `summary`. Neben den Geradenparametern werden dort mehrere weitere Werte angegeben, welche Auskunft über die Güte der Approximation machen. Einige davon werden wir in späteren Kapiteln noch genauer kennen lernen, z. B. den sogenannten  $p$ -Wert. Ein Plot der resultierenden Regressionsgeraden durch die Punktwolke visualisiert die Approximation.

die Ausgabe

■

### Zusammenfassung

Mit Hilfe von Schätzern können Kenngrößen von Zufallsvariablen aus gemessenen Daten approximiert werden. Dabei gibt es verschiedene mögliche Schätzer, welche situationsabhängig gewählt werden können, und welche unterschiedliche günstige Eigenschaften haben. Eine wichtige Methode ist die Maximum Likelihood Schätzung. Die lineare Regression kann als Schätzmethode für den linearen Zusammenhang zwischen Messwerten zweier verschiedener Zufallsvariablen interpretiert werden.

## Aufgaben

**8.1** Gegeben seien folgende Daten:

Messung Nr.	1	2	3	4	5	6	7	8	9	10
Messwert	51	69	40	44	49	61	52	53	48	50

Berechnen Sie für diese Daten a) das empirische Mittel b) den Median c) die empirische Varianz.

**8.2** Beweisen Sie, dass die empirische Varianz erwartungstreu und konsistent ist (vgl. Beispiel 8.4 und Beispiel 8.5).

**8.3** Berechne den Maximum-Likelihood Schätzer für den Parameter  $p$  der geometrischen Verteilung.

**8.4** Berechne den Maximum-Likelihood Schätzer für den Parameter  $\lambda$  der Exponentialverteilung.

**8.5** Eine Messung ergibt folgende Datenpaare  $(x_i, y_i), i = 1, \dots, 6$ :

$i$	1	2	3	4	5	6
$x_i$	2	8	5	2	12	4
$y_i$	10	2	8	11	1	8

a) Stellen Sie die Daten graphisch als Punktmenge dar. Weist die graphische Darstellung auf positiv/negativ korrelierte Daten hin? b) Berechnen Sie den Korrelationskoeffizienten c) Berechnen Sie die Regressionsgerade.

**8.6** Es seien  $X_1, \dots, X_n$  unabhängige Zufallsvariablen mit parameterabhängiger Dichte

$$f_{\theta}(x) = \begin{cases} \frac{1}{2}(1 + \theta x), & -1 \leq x \leq 1, \\ 0 & \text{sonst,} \end{cases}$$

$\theta \in ]-1, 1[$ . Finden Sie einen erwartungstreuen, effizienten und konsistenten Schätzer für den unbekannten Parameter  $\theta$ . Hinweis: Das empirische Mittel untersuchen.



# 9 Konfidenzbereiche

---

## Übersicht

9.1	Grundprinzipien und Beispiele .....	124
9.2	Einschub: $\chi^2$ -Verteilung und $t$ -Verteilung .....	128
9.3	Konfidenzintervalle für Erwartungswert und Varianz .....	130

---

Im Kapitel 8 wurden verschiedene Methoden vorgestellt, um einen unbekannten Parameter einer Verteilung aus Messwerten zu schätzen. Dabei wurde auch die Frage der “Güte” eines Schätzer angesprochen, denn natürlich möchte man immer erreichen, dass ein geschätzter Parameter möglichst nah am echten, unbekannten Wert ist.

Jedoch ist es auch bei Schätzern mit günstigen Eigenschaften prinzipiell immer möglich, dass die gemessenen Daten für die zugrundeliegende Verteilung „untypisch“ sind, also eher extreme Ergebnisse darstellen. In solchen Fällen wird der aus den Daten berechnete Schätzer trotz guten Methoden vom echten Wert stark abweichen. Mit Hilfe von *Konfidenzbereichen* kann die Wahrscheinlichkeit von solchen Abweichungen kontrolliert werden.

### Lernziele

- Das Grundprinzip von Konfidenzbereichen kennen
- Konfidenzintervalle in Beispielen berechnen können

### Vorkenntnisse

Benötigt werden Vorkenntnisse über Zufallsvariablen und Verteilungen, insbesondere die Normalverteilung, sowie über Schätzer wie in Kapitel 8.

## 9.1 Grundprinzipien und Beispiele

Wie bisher wird von Messwerten  $x_1, \dots, x_n$  ausgegangen. Wir nehmen an, dass es sich dabei um Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen  $X_1, \dots, X_n$  handelt, deren Verteilung von einem unbekannten Parameter  $\theta$  abhängig ist.

In der obigen Situation bezeichnen wir weiterhin mit  $\Theta$  die Menge aller theoretisch möglicher Parameterwerte  $\theta$ . Beispielsweise ist also  $\Theta = \mathbb{R}$ , falls  $\theta = \mu$  der Erwartungswert einer normalverteilten Zufallsvariablen ist, oder  $\Theta = [0, 1]$  falls  $\theta = p$  die Erfolgswahrscheinlichkeit bei einem Bernoulli-Experiment ist.

Um einen Schätzer für  $\theta$  zu bestimmen, wurde in Kapitel 8 eine Schätzfunktion  $\theta_n$  benutzt. Sei  $\bar{\theta} := \theta_n(x_1, \dots, x_n)$ . Für einen Schätzer sollte nach Möglichkeit  $\bar{\theta} \approx \theta$  gelten, wobei jedoch  $\bar{\theta} = \theta$  höchstens in Ausnahmefällen zu erreichen sein wird. Die Idee hinter dem Konzept von Konfidenzbereichen ist es deshalb, eine (zufällige) Menge  $J \subset \Theta$  zu finden, welche den richtigen Wert  $\theta$  mit hinreichend großer Wahrscheinlichkeit enthält. Was „hinreichend groß“ dabei genau bedeutet, muss vor dem Zufallsexperiment festgelegt werden, konkret wird dabei eine Zahl  $\alpha \in ]0, 1[$  gewählt, das sogenannte *Fehlerniveau*. Typische Werte für  $\alpha$  sind dabei  $\alpha = 0.05$  oder  $\alpha = 0.01$ .

### Definition 9.1 (Konfidenzbereich)

Seien  $X_1, \dots, X_n$  unabhängige, identisch verteilte Zufallsvariablen, deren Verteilung von einem Parameter  $\theta \in \Theta$  abhängt. Sei ein Fehlerniveau  $\alpha \in ]0, 1[$  vorgeben. Ein **Konfidenzbereich für den Parameter  $\theta$  zum Fehlerniveau  $\alpha$**  ist ein Intervall  $J = J(X_1, \dots, X_n)$ , so dass

$$\mathbb{P}_\theta(J \ni \theta) \geq 1 - \alpha \quad \text{für jedes } \theta \in \Theta$$

gilt. Dabei bezeichnet  $\mathbb{P}_\theta$  die Wahrscheinlichkeit unter dem Parameterwert  $\theta$  der entsprechenden Verteilung. ♦

Hierbei ist es wichtig zu beachten, dass das Intervall  $J$  (als Funktion der Zufallsvariablen  $X_1, \dots, X_n$ ) *zufällig* ist, und nicht  $\theta$ . Dies soll hier durch die Schreibweise  $J \ni \theta$  (statt der üblicheren Variante  $\theta \in J$ ) deutlich gemacht werden. Insbesondere soll  $J$  nicht von  $\theta$  abhängen.

Ein Konfidenzbereich ist also eine zufällige Teilmenge des Parameterbereichs, welcher den (wahren, aber unbekannten) Parameter  $\theta$  mit einer Wahrscheinlichkeit von mindestens  $1 - \alpha$  umfasst, egal welches  $\theta \in \Theta$  tatsächlich auftritt.

Um möglichst viel Information zu enthalten, sollte  $J$  möglichst klein sein, also die Parameterwerte möglichst stark einschränken. Wie klein  $J$  sein kann, hängt jedoch von der Wahl von  $\alpha$  ab. Je kleiner das Fehlerniveau  $\alpha$  gewählt ist, desto größer muss  $J$  sein.

### Beispiel 9.1

Seien  $X_1, \dots, X_n$  unabhängige, Bernoulli-verteilte Zufallsvariablen mit Parameter  $p$ . Wir geben uns das Fehlerniveau  $\alpha = 0.01$  vor. Sei  $J$  ein Konfidenzbereich zum Fehlerniveau  $\alpha$ . Seien  $x_1, \dots, x_n$  Realisierungen von  $X_1, \dots, X_n$ . Die Definition des Konfidenzbereichs

besagt, dass mit Wahrscheinlichkeit  $0.99 = 1 - \alpha$  der Bereich  $J(X_1, \dots, X_n)$  den richtigen Wert  $p$  enthält. ■

Um einen Konfidenzbereich konkret zu bestimmen, verwendet man oft die sogenannte *Quantilmethode*.

### Definition 9.2 (Quantil)

Sei  $\beta \in [0, 1]$ , und sei  $Y$  eine Zufallsvariable mit bekannter Verteilungsfunktion  $F$ . Das *Quantil* der Verteilung von  $Y$  zum Niveau  $\beta$  ist die kleinste Zahl  $q_\beta$ , für die gilt

$$F(q_\beta) = \mathbb{P}(Y \leq q_\beta) \geq \beta.$$

Hängt die Verteilung der Zufallsvariablen  $Y$  von einem Parameter  $\theta$  ab, so schreiben wir  $q_{\theta, \beta}$  für das  $\beta$ -Quantil bei Parameterwert  $\theta$ . ♦

Quantile wichtiger Verteilungen werden in Tabellen angegeben, bzw. können von Statistik-Programmen wie R ausgegeben werden. Besitzt die Zufallsvariable  $Y$  eine Dichte, so gilt die Beziehung  $F(y) = \int_{-\infty}^y f(t)dt$ . Das Quantil  $q_\beta$  ist also der kleinste Wert  $y$ , für den  $F(y) \leq \beta$  ist. Um  $q_\beta$  zu bestimmen, müssen wir also die Integralgleichung

$$F(q_\beta) = \int_{-\infty}^{q_\beta} f(t)dt = \beta$$

lösen. In anderen Worten muss der Flächeninhalt unter dem Graphen der Dichte links von  $q_\beta$  größer oder gleich  $\beta = 0.95$  sein.

Die Grundidee der Quantilmethode zur Bestimmung eines Konfidenzbereichs basiert auf folgender Beobachtung: Für große  $\beta$  (z.B.  $\beta = 0.95$ ) ist die Wahrscheinlichkeit, dass eine Realisierung einer Zufallsvariablen  $Y$  größer als  $q_\beta$  ist, sehr klein (nämlich  $1 - \beta$ .) Umgekehrt ist für kleine  $\beta$  die Wahrscheinlichkeit, dass eine Realisierung kleiner ist als  $q_\beta$  sehr klein. Falls nun die Verteilung von  $Y$  von einem Parameter  $\theta \in \Theta$  abhängt, so suchen wir, in Abhängigkeit von einer Realisierung  $y$  einen Bereich  $J(y) \subset \Theta$ , für den die Wahrscheinlichkeit  $y$  zu beobachten groß genug ist, nämlich mindestens  $1 - \alpha$ , wenn als Fehlerniveau  $\alpha$  gewählt wurde. Somit bietet sich, je nach Problemstellung, eine der drei folgenden Möglichkeiten an. Dabei sei  $Y$  eine Zufallsvariable deren Verteilung von  $\theta \in \Theta$  abhängt, und sei  $\alpha \in ]0, 1[$ . Sei  $y$  eine Realisierung von  $Y$ .

- Der **obere Konfidenzbereich** zum Niveau  $\alpha$  ist gegeben durch

$$J(y) = \{\theta \in \Theta : y \leq q_{\theta, 1-\alpha}\},$$

d.h. man arbeitet mit dem *oberen* Quantil.

- Der **untere Konfidenzbereich** zum Niveau  $\alpha$  ist gegeben durch

$$J(y) = \{\theta \in \Theta : y \geq q_{\theta, \alpha}\}$$

- Der **beidseitige Konfidenzbereich** zum Niveau  $\alpha$  ist gegeben durch

$$J(y) = \{\theta \in \Theta : q_{\theta, \alpha/2} \leq y \leq q_{\theta, 1-\alpha/2}\}$$

Aus der Definitionen der Begriffe Quantil und Konfidenzbereich folgt unmittelbar, dass jeder der drei Fälle einen Konfidenzbereich zum Fehlniveau  $\alpha$  definiert. Zu beachten ist, dass hier eine einzelne Zufallsvariable  $Y$  betrachtet wurde, und nicht, wie in unserer ursprünglichen Formulierung der Fragestellung, mehrere Zufallsvariablen  $X_1, \dots, X_n$ . Um die Quantilmethode anwenden zu können, muss deshalb eine geeignete Zufallsvariable  $Y$  als Funktion der  $X_1, \dots, X_n$  konstruiert werden, deren Verteilung in Abhängigkeit von dem unbekannten Parameter  $\theta$  bekannt ist.

Diese Idee ist nützlich um ein Konfidenzintervall für die Erfolgswahrscheinlichkeit  $p$  in einem Bernoulli-Experiment zu finden, da wir ausnutzen können, dass die Anzahl Erfolge in einem  $n$ -fach ausgeführten Bernoulli-Experiment eine binomialverteilte Zufallsvariable ist. Später werden wir weitere Beispiele sehen.

### Beispiel 9.2 (Binomialverteilung, einseitiges Konfidenzintervall)

Bei einer Produktion eines Massenartikels wird eine gewisse Menge an Ausschuss produziert. Wir suchen ein Konfidenzintervall für die unbekannte Wahrscheinlichkeit  $p$ , mit der ein zufällig ausgewähltes Stück Ausschuss ist. Dafür testen wir  $n$  zufällig ausgewählte Stücke, wobei wir entweder mit Zurücklegen ziehen, oder eine kleine Stichprobe aus einer sehr großen Gesamtheit ziehen. Die Anzahl  $X$  an Ausschuss ist somit (zumindest annähernd) binomialverteilt mit Parametern  $n$  und  $p$ . Wenn in unserem Probelauf  $k$  Ausschuss-Stücke enthalten sind, so erhalten wir einen Schätzer für  $p$ , nämlich  $\bar{p} = \frac{k}{n}$ . Nun suchen wir jedoch einen Konfidenzbereich. Es gilt  $\Theta = [0, 1]$ , da dies die Werte sind, die  $p$  annehmen kann.

Fixiere das Fehlniveau  $\alpha$ . Nun müssen wir uns zuerst überlegen, welche der drei Varianten für den Konfidenzbereich wir verwenden wollen. Dazu müssen wir uns klar machen, welche Information wir aus dem Konfidenzbereich ziehen wollen. Da  $p$  die Wahrscheinlichkeit für Ausschuss ist, möchte man ein möglichst kleines  $p$  haben. Somit suchen wir eine obere Schranke für den Konfidenzbereich, denn man will sich dagegen absichern, versehentlich mit einem zu großen  $p$  zu rechnen. Wir benutzen also den zweiten Fall aus obiger Liste, denn unteren Konfidenzbereich, denn je größer  $p$  ist, desto größer ist  $q_{p,\alpha}$  bzw.  $q_{p,1-\alpha}$ .

Wir erhalten für  $k \in \{0, \dots, n\}$

$$\begin{aligned} J(k) &= \{p \in [0, 1] : k \geq q_{p,\alpha}\} \\ &= \{p \in [0, 1] : \mathbb{P}_p(X \leq k) \geq \mathbb{P}_p(X \leq q_{p,\alpha}) = \alpha\} \\ &= \left\{p \in [0, 1] : \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i} \geq \alpha\right\}. \end{aligned}$$

Falls wir also  $k = 0$  Stück Ausschuss in unserem Test gezählt haben, so erhalten wir

$$J(0) = \{p \in [0, 1] : (1-p)^n \geq \alpha\} = [0, 1 - \alpha^{1/n}].$$

Für andere Werte von  $k$  ist das Auflösen von  $\sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i} \geq \alpha$  nicht mehr so leicht möglich. Zur expliziten Bestimmung des Konfidenzbereichs kommen deshalb

entweder numerische Methoden zur Anwendung, oder man verwendet die Approximation der Binomialverteilung durch die Normalverteilung, Satz 7.3, welcher besagt dass

$$\mathbb{P}_p(X \leq k) \approx \Phi_{0,1}\left(\frac{k + 1/2 - np}{\sqrt{np(1-p)}}\right)$$

ist. Sei  $z_\alpha$  das  $\alpha$ -Quantil der Standardnormalverteilung. Dieses kann aus der Tabelle der Normalverteilungsquantile (siehe Anhang) abgelesen werden. Nach unseren Überlegungen muss also gelten

$$\Phi_{0,1}\left(\frac{k + 1/2 - np}{\sqrt{np(1-p)}}\right) \geq \alpha,$$

d.h.

$$\frac{k + 1/2 - np}{\sqrt{np(1-p)}} \geq z_\alpha.$$

Auflösen nach  $p$  ergibt die obere Schranke. Mit Hilfe des zweiseitigen Konfidenzbereichs kann analog ein zweiseitiges Konfidenzintervall bestimmt werden. ■

Die Grundidee der Quantilmethode ist auch in vielen anderen Fällen nützlich. Wir betrachten ein weiteres Beispiel.

### Beispiel 9.3 (Erwartungswert der Normalverteilung bei bekannter Varianz)

Seien  $x_1, \dots, x_n$  Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen  $X_1, \dots, X_n$ , welche einer Normalverteilung mit bekannter Varianz  $\sigma^2$  und unbekanntem Erwartungswert  $\mu$  folgen. Wir wählen das Fehlerniveau  $\alpha = 0.05$ . Dazu suchen wir ein Konfidenzintervall für  $\mu$ , d.h.  $J \subset \mathbb{R}$  so dass

$$\mathbb{P}_\mu(\mu \in J) \geq 1 - \alpha = 0.95 \quad \text{für alle } \mu \in \mathbb{R}.$$

Wir wählen dazu einen Ansatz:  $J$  sei um das empirische Mittel zentriert, d.h. es gibt  $h > 0$  so dass

$$J = [\bar{\mu} - h, \bar{\mu} + h]$$

gilt, mit  $\bar{\mu} = \bar{\mu}(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i$ . Da  $\bar{\mu}$  einfach berechnet werden kann, müssen wir also  $h = h(\alpha, n, \sigma^2)$  so bestimmen, dass die Bedingung für einen Konfidenzbereich zum Fehlerniveau  $\alpha$  erfüllt ist. Wir wollen also

$$\begin{aligned} 1 - \alpha &\leq \mathbb{P}_\mu\left(\mu \in \left[\frac{1}{n} \sum_{i=1}^n X_i - h, \frac{1}{n} \sum_{i=1}^n X_i + h\right]\right) \\ &= \mathbb{P}_\mu\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu) \in [-h, h]\right) \\ &= \mathbb{P}_\mu\left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \in \left[\frac{-h}{\sigma\sqrt{n}}, \frac{h}{\sigma\sqrt{n}}\right]\right). \end{aligned}$$

Nun ist nach Aufgabe 6.9 die Zufallsvariable

$$Y := \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu)$$

standardnormalverteilt, da die  $X_i$  unabhängig und  $\mathcal{N}(\mu, \sigma^2)$ -verteilt sind. Somit müssen wir  $h$  so wählen, dass für  $Y \sim \mathcal{N}(0, 1)$  gilt

$$\mathbb{P}\left(Y \in \left[\frac{-h}{\sigma\sqrt{n}}, \frac{h}{\sigma\sqrt{n}}\right]\right) \geq 1 - \alpha,$$

beziehungsweise, was wegen der Symmetrie der Standardnormalverteilung äquivalent ist,

$$\mathbb{P}\left(Y \geq \frac{h}{\sigma\sqrt{n}}\right) \leq \frac{\alpha}{2}.$$

Letzteres besagt aber genau, dass  $\frac{h}{\sigma\sqrt{n}}$  gleich dem Quantil der Standardnormalverteilung zum Wert  $1 - \frac{\alpha}{2}$  sein muss. In anderen Worten haben wir den folgenden Satz bewiesen:

**Satz 9.1 (Konfidenzintervall für den Erwartungswert der Normalverteilung)**

*Das Konfidenzintervall zum Fehlerniveau  $\alpha$  für den Erwartungswert  $\mu$  einer Normalverteilung bei bekannter Varianz  $\sigma^2$  und  $n$  unabhängigen Messungen ist gegeben durch*

$$J = J(n, \alpha, \sigma) = [\bar{\mu} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{\mu} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}]$$

wobei  $z_{1-\alpha/2}$  das Quantil der Standardnormalverteilung ist, und  $\bar{\mu}$  das empirische Mittel. ■

Wenn man sich die Herleitung im letzten Beispiel genauer anschaut, so stellt man fest, dass die Normalverteilung nur an einer Stelle benutzt wurde, um zu argumentieren, dass  $Y$  standardnormalverteilt ist. Nun hat aber  $Y$  (bis auf den konstanten Vorfaktor  $\frac{1}{\sigma}$ ) genau die Form der Zufallsvariable, die im zentralen Grenzwertsatz, Satz 7.2, vorkommt. Da der zentrale Grenzwertsatz für Folgen von unabhängigen, identisch verteilten Zufallsvariablen  $X_i, i \in \mathbb{N}$  mit beliebiger Verteilung gilt (sofern die Varianz existiert), erhält man als Konfidenzintervall für den Erwartungswert bei bekannter Varianz, zumindest im Grenzwert für  $n \rightarrow \infty$ , das gleiche Ergebnis wie für die Normalverteilung.

**Satz 9.2 (Konfidenzintervall für den Erwartungswert bei bekannter Varianz)**

*Seien  $X_1, \dots, X_n$  unabhängige, identisch verteilte Zufallsvariablen mit existierender (und bekannter) Varianz  $\sigma^2 > 0$ . Falls  $n$  hinreichend groß ist, ist das Konfidenzintervall zum Fehlerniveau  $\alpha$  für den Erwartungswert  $\mu$  gegeben durch*

$$J = J(n, \alpha, \sigma) = [\bar{\mu} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{\mu} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}]$$

wobei  $z_{1-\alpha/2}$  das Quantil der Standardnormalverteilung ist, und  $\bar{\mu}$  das empirische Mittel.

## 9.2 Einschub: $\chi^2$ -Verteilung und $t$ -Verteilung

Es sollen jetzt zwei weitere wichtige Konfidenzintervalle berechnet werden, nämlich

- Konfidenzintervall für die Varianz bei bekanntem Erwartungswert, und

- Konfidenzintervall für den Erwartungswert bei unbekannter Varianz.

Dafür werden zwei Verteilungen benötigt, die bisher noch nicht diskutiert wurden, die jedoch auch in Kapitel 10 für das Testen von Hypothesen relevant werden.

### Definition 9.3

Eine Zufallsvariable  $X$  ist  $\chi^2$ -verteilt (sprich: chi-quadrat verteilt), bzw. folgt einer  $\chi^2$ -Verteilung mit Parameter  $n \in \mathbb{N}$ , wenn sie die Dichte

$$f_n(x) = \begin{cases} \frac{x^{n/2-1} e^{-x/2}}{2^{n/2} \Gamma(n/2)} & \text{falls } x \geq 0 \\ 0 & \text{falls } x < 0 \end{cases}$$

hat. Dabei ist  $\Gamma$  die sogenannte Gamma-Funktion, welche für Werte der Form  $n/2$ ,  $n \in \mathbb{N}$  rekursiv aus  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ ,  $\Gamma(1) = 1$  und  $\Gamma(r+1) = r\Gamma(r)$  für alle  $r \in [0, \infty[$  berechnet werden kann.  $\blacklozenge$

Für die hier beschriebenen Zwecke werden die Quantile der  $\chi^2$ -Verteilung relevant sein, welche wiederum in Tabellen aufgelistet werden, siehe Tabelle A.4 im Anhang. Die Bedeutung der  $\chi^2$ -Verteilung für die Statistik liegt in folgendem Satz, der hier ohne Beweis wiedergegeben wird:

### Satz 9.3

Seien  $Y_1, \dots, Y_n$  unabhängige standardnormalverteilte Zufallsvariablen. Dann folgt die Zufallsvariable

$$Y_1^2 + \dots + Y_n^2$$

einer  $\chi^2$ -Verteilung mit Parameter  $n$ .

### Definition 9.4

Eine Zufallsvariable  $X$  heißt  $t$ -verteilt zum Parameter  $n \in \mathbb{N}$ , falls sie die Dichte

$$f_n(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

hat. Dabei ist  $\Gamma$  wieder die Gamma-Funktion.  $\blacklozenge$

Auch hier ist weniger die genaue Form der Dichte relevant, sondern die tabellierten Quantile, siehe Tabelle A.3. Die  $t$ -Verteilung tritt in folgender Situation auf:

### Satz 9.4

Seien  $X_1, \dots, X_n$  unabhängige,  $\mathcal{N}(\mu, \sigma^2)$ -verteilte Zufallsvariablen. Seien  $\bar{\mu}_n(X_1, \dots, X_n)$  und  $\bar{\sigma}_n(X_1, \dots, X_n)$  das empirische Mittel bzw. die empirische Standardabweichung dieser Zufallsvariablen. Dann folgt die Zufallsvariable

$$\frac{\sqrt{n}(\bar{\mu}_n(X_1, \dots, X_n) - \mu)}{\bar{\sigma}_n(X_1, \dots, X_n)}$$

einer  $t$ -Verteilung mit Parameter  $n - 1$ .

Die empirische Standardabweichung ist dabei die Wurzel aus der empirischen Varianz, vgl. Kapitel 8. Zu beachten ist, dass der Parameter der  $t$ -Verteilung tatsächlich  $n - 1$  ist, und nicht etwa  $n$ , obwohl wir  $n$  Zufallsvariablen  $X_1, \dots, X_n$  betrachten.

## 9.3 Konfidenzintervalle für Erwartungswert und Varianz

In den Sätzen 9.1 und 9.2 haben wir gesehen, wie das Konfidenzintervall für den Erwartungswert bei bekannter Varianz bestimmt wird. In diesem Abschnitt werden nun zwei weitere Situationen betrachtet.

### Satz 9.5 (Konfidenzintervall für die Varianz der Normalverteilung)

Seien  $x_1, \dots, x_n$  Messwerte von Realisierungen von unabhängigen normalverteilten Zufallsvariablen  $X_1, \dots, X_n$  mit bekanntem Erwartungswert  $\mu$  und unbekannter Varianz  $\sigma^2$ . Ein Konfidenzintervall für  $\sigma^2$  ist

$$J = \left[ \bar{\sigma}^2 \frac{n-1}{\chi_{n-1, 1-\alpha/2}^2}, \bar{\sigma}^2 \frac{n-1}{\chi_{n-1, \alpha/2}^2} \right]$$

wobei  $\bar{\sigma}^2$  die empirische Varianz ist, und  $\chi_{n-1, \beta}^2$  das  $\beta$ -Quantil der  $\chi^2$ -Verteilung mit Parameter  $n-1$ .

**Beweis 9.1** Der Beweis verläuft ähnlich wie der für Satz 9.1, unter Verwendung von Satz 9.3. □

Für den Fall dass die  $X_1, \dots, X_n$  eine beliebige Verteilung haben, also nicht zwingend normalverteilt sind, können wir uns im Fall von unbekannter Varianz nicht einfach auf den zentralen Grenzwertsatz zur Normalapproximation berufen. Hier sind andere Methoden notwendig, die den Rahmen dieses Buches sprengen würden.

Kennt man weder Erwartungswert noch Varianz, so kann man im Fall von normalverteilten Zufallsvariablen dennoch ein Konfidenzintervall für den Erwartungswert angeben.

### Satz 9.6 (Konfidenzintervall für den Erwartungswert der Normalverteilung)

Seien  $x_1, \dots, x_n$  Messwerte von Realisierungen von unabhängigen normalverteilten Zufallsvariablen  $X_1, \dots, X_n$  mit unbekanntem Erwartungswert  $\mu$  und unbekannter Varianz  $\sigma^2$ . Ein Konfidenzintervall für  $\mu$  ist

$$J = \left[ \bar{\mu} - \frac{\bar{\sigma}}{\sqrt{n}} t_{n-1, 1-\alpha/2}, \bar{\mu} + \frac{\bar{\sigma}}{\sqrt{n}} t_{n-1, 1-\alpha/2} \right]$$

wobei  $\bar{\mu}$  das empirische Mittel,  $\bar{\sigma}$  die empirische Standardabweichung, und  $t_{n-1, 1-\alpha/2}$  das Quantil der  $t$ -Verteilung mit Parameter  $n-1$  ist.

**Beweis 9.2** Wiederum folgt der Beweis dem Vorgehen von Satz 9.1, unter Verwendung von Satz 9.4. □

### Zusammenfassung

Konfidenzbereiche verallgemeinern die Parameterschätzung dahingehend, dass sie zufällige Teilmengen des Parameterbereichs bilden, welche den echten Parameter zu



gemessenen Daten mit hoher Wahrscheinlichkeit enthalten. Zur Bestimmung eines Konfidenzintervalls werden die Quantile wichtiger Verteilungen benötigt, situationsabhängig sind dies die Normalverteilung,  $\chi^2$ -Verteilung oder die  $t$ -Verteilung.

## Aufgaben

**9.1** Sei  $X$  eine normalverteilte Zufallsvariable mit Parametern  $\mu = 15, \sigma^2 = 4$ . Benutzen Sie die Tabelle der Standardnormalverteilung, um die Quantile  $q_\beta$  von  $X$  für a)  $\beta = 0.05$ , b)  $\beta = 0.1$ , c)  $\beta = 0.01$  zu bestimmen.

**9.2** a) Sei  $\beta > \beta'$ . Sei  $X$  eine Zufallsvariable. In welcher Beziehung stehen die Quantile  $q_\beta$  und  $q_{\beta'}$  von  $X$  zueinander? b) Sei  $\alpha > \alpha'$ . Sei  $J_\alpha$  das Konfidenzintervall zum Fehlniveau  $\alpha$  und  $J_{\alpha'}$  das Konfidenzintervall zu  $\alpha'$ , jeweils für den Erwartungswert  $\mu$  von normalverteilten Zufallsvariablen. In welcher Beziehung stehen  $J_\alpha$  und  $J_{\alpha'}$  zueinander?

**9.3** In einer Testreihe wird das Gewicht einer neu gezüchteten Schweineart untersucht. Die Daten von 10 Schweinen ergeben folgende Tabelle:

Schwein Nr. $i$	1	2	3	4	5	6	7	8	9	10
Gewicht (kg)	260	243	245	248	251	250	239	254	259	249

a) Berechnen Sie die empirische Varianz b) Unter der Annahme, die Varianz sei durch das Ergebnis von a) gegeben, und unter Annahme einer Normalverteilung, berechne das Konfidenzintervall für den Erwartungswert zum Fehlniveau  $\alpha = 0.05$ .

**9.4** Betrachte folgende Daten:

$i$	1	2	3	4	5	6	7
$x_i$	23.7	21.1	19.6	23.6	21.5	22.8	19.7

Unter Normalverteilungsannahme, bestimme das Konfidenzintervall für den Erwartungswert zum Fehlniveau  $\alpha = 0.1$ .

**9.5** In der Situation von Aufgabe 9.4, bestimmen Sie das Konfidenzintervall für  $\sigma^2$  zum Fehlniveau  $\alpha = 0.05$ .

**9.6** Seien  $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$  und  $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$  unabhängige Zufallsvariablen, wobei  $\mu_1, \mu_2 \in \mathbb{R}$ ,  $\sigma_1^2, \sigma_2^2 > 0$ . Geben Sie eine Funktion  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  an, so dass  $f(X_1, X_2)$   $\chi^2$ -verteilt mit Parameter 2 ist.

**9.7** Betrachte folgende Daten:

$i$	1	2	3	4	5	6	7	8	9	10	
$x_i$	140	136	150	144	148	152	138	141	143	151	

Diese werden als normalverteilt angenommen, mit bekanntem Erwartungswert 143. a) Bestimmen Sie die empirische Varianz b) Bestimmen Sie das Konfidenzintervall für  $\sigma^2$  zum Fehlniveau  $\alpha = 0.01$ .

**9.8** Die Daten aus Aufgabe 9.7 seien normalverteilt mit unbekanntem Erwartungswert und unbekannter Varianz. Bestimmen Sie das Konfidenzintervall für  $\mu$  zum Fehlniveau  $\alpha = 0.05$ .

# 10 Hypothesentests

---

## Übersicht

10.1 Grundprinzipien, $p$ -Wert, Fehler erster und zweiter Art .....	134
10.2 $t$ -Test .....	138
10.3 $\chi^2$ -Test .....	143

---

Hypothesentests gehören zu den grundlegenden Methoden der Statistik. Sie haben das Ziel, Annahmen über Messwerte mit wahrscheinlichkeitstheoretischen Methoden zu überprüfen. Dabei wird untersucht, ob die gemessenen Werte unter der getroffenen Annahme sehr unwahrscheinlich sind. Sind sie das, wird die Annahme verworfen, und das Gegenteil für richtig angenommen. Sind die Messwerte hingegen unter der getroffenen Annahme im normalen Bereich der Wahrscheinlichkeiten, wird die Annahme beibehalten. Zur Überprüfung der Annahme verwendet man üblicherweise Konfidenzintervalle, Vergleiche mit Quantilen entsprechender Verteilungen, oder  $p$ -Werte.

### Lernziele

- Die Grundprinzipien von Hypothesentests kennen
- $p$ -Werte bestimmen und interpretieren können
- $t$ -Test und  $\chi^2$ -Test ausführen und interpretieren können

### Vorkenntnisse

Gute Kenntnisse über Zufallsvariablen, Verteilungen und Kenngrößen, sowie über die statistischen Grundlagen aus Kapitel 8 und 9.

## 10.1 Grundprinzipien, $p$ -Wert, Fehler erster und zweiter Art

Hypothesentests beruhen auf folgender Grundidee. Vor der Messung der Daten wird eine Hypothese betreffend der zugrundeliegenden Zufallsvariablen aufgestellt. Beispiele für solche Hypothesen sind

- Die Daten folgen einer Normalverteilung
- Der Erwartungswert ist 2
- Die Realisierungen sind unabhängig

Nach der Messung der Daten wird durch Rechnung überprüft, ob die gemessenen Daten der Hypothese widersprechen, d.h. ob unter der Annahme, dass die Hypothese gilt, die Daten sehr unwahrscheinlich sind. Sind die Messwerte sehr unwahrscheinlich, wird die Hypothese verworfen. Allgemein folgt man diesem Vorgehen:

1. Formuliere die *Nullhypothese*  $H_0$ .
2. Konstruiere den *Annahmebereich* (bzw. den Ablehnungsbereich).
3. Führe das Experiment durch, erhalte Daten  $x_1, \dots, x_n$ .
4. Überprüfe, ob die Daten im Annahmebereich liegen.

Je nach Ergebnis bei 4. nimmt man die Hypothese an oder verwirft sie:

- Falls in 4. das Ergebnis *nein* lautet: Nullhypothese verwerfen
- Andernfalls: Nullhypothese annehmen

### Definition 10.1 (Nullhypothese, Annahmebereich, Fehlerniveau)

Seien  $X_1, \dots, X_n$  Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathbb{P})$ . Seien  $x_1, \dots, x_n \in \mathbb{R}$  Messwerte. Eine **Nullhypothese**  $H_0$  über  $X_1, \dots, X_n$  ist eine Aussage, deren Wahrheitsgehalt nur von der gemeinsamen Verteilung von  $X_1, \dots, X_n$  abhängt, so dass  $\{H_0 \text{ ist wahr}\}$  ein Ereignis ist. Die **Alternativhypothese**  $H_A$  ist die Negation von  $H_0$ . Ein **Annahmebereich zum Fehlerniveau**  $\alpha \in ]0, 1[$  für die Nullhypothese  $H_0$  ist eine Teilmenge  $A$  von  $\mathbb{R}^n$ , so dass

$$\mathbb{P}((X_1, \dots, X_n) \in A \mid H_0 \text{ ist wahr}) \geq 1 - \alpha$$

ist. ♦

Geeignete Annahmebereiche werden je nach Situation auf Grund von theoretischen Überlegungen konstruiert. Zum vorgegebenen Fehlerniveau  $\alpha$  soll also gelten, dass *falls* die Nullhypothese erfüllt ist, die Messwerte mit einer (hohen) Wahrscheinlichkeit von mindestens  $1 - \alpha$  im Annahmebereich liegen.

**Beispiel 10.1 (Materialbelastung)**

Es soll ein Stabilitäts-Belastungstest durchgeführt werden, um die Frage zu klären, wie viel Gewicht ein Bauteil (im Mittel) tragen kann, ohne dass es zu stark verformt wird. Es wird angenommen, dass dies im Mittel 12 kg sind. Mathematisch kann diese Frage modelliert werden, indem angenommen wird, das zu tragende Gewicht (in kg) sei durch eine Zufallsvariable  $X$  mit Werten in  $[0, \infty)$  gegeben ist. Für  $n = 20$  baugleiche Teile seien die jeweiligen Gewichte durch unabhängige, identisch verteilte Zufallsvariablen  $X_1, \dots, X_n$  mit derselben Verteilung wie  $X$  gegeben. In einer Messreihe wird die Belastung von 20 baugleichen Teilen gemessen, also Daten  $x_1, \dots, x_{20}$  als Realisierungen der Zufallsvariablen  $X_1, \dots, X_{20}$  ermittelt. Weiter soll die Varianz der Zufallsvariablen als bekannt und endlich angenommen werden, konkret sei  $\sigma^2 = \mathbb{V}(X) = 4$ . Wir führen nach dem beschriebenen Vorgehen Vorgehen einen Hypothesentest durch.

1. Nullhypothese  $H_0$ : Der Erwartungswert  $\mu$  von  $X$  ist gleich  $\mu_0 := 12$  (kg).
2. Annahmebereich: Wir betrachten hier einen der gängigen Ansätze. Zuerst wählen wir eine Fehlerschranke, z.B. 5%, also  $\alpha = 0.05$ . Wir suchen nun einen Annahmebereich  $A \subset \mathbb{R}^{20}$ , für den

$$\mathbb{P}_{\mu_0}((X_1, \dots, X_{20}) \in A) \geq 1 - \alpha$$

gilt. Das erinnert an das Vorgehen bei Konfidenzintervallen: Falls die Nullhypothese gilt, falls also  $\mu_0$  der richtige Erwartungswert ist, soll das gemessene empirische Mittel  $\bar{\mu}$  mit einer Wahrscheinlichkeit von mindestens  $1 - \alpha$  im Annahmebereich  $A$  liegen. Die Wahrscheinlichkeit, die Nullhypothese zu verwerfen, obwohl sie wahr ist, soll also höchstens  $\alpha$  sein. Wir können uns weiter an Konfidenzintervallen orientieren, und als Ansatz für den Annahmebereich

$$A := \{x = (x_1, \dots, x_{20}) : \bar{\mu}(x) \in [\mu_0 - h, \mu_0 + h]\}$$

wählen. Dabei ist  $\bar{\mu}$  wie bisher das empirische Mittel der Daten, und die Zahl  $h$  muss geeignet gewählt werden, so dass die Bedingung  $\mathbb{P}_{\mu_0}(X \in A) \geq 1 - \alpha$  erfüllt ist. Mit diesem Ansatz für  $A$  gilt

$$\mathbb{P}_{\mu_0}(X_1, \dots, X_n \in A) = \mathbb{P}(\mu_0 \in J(X_1, \dots, X_n)),$$

wobei  $J$  das Konfidenzintervall für den Erwartungswert bei bekannter Varianz  $\sigma^2$  für das Fehlniveau  $\alpha$  ist. Aus Satz 9.2 ist bekannt, dass

$$J(X_1, \dots, X_n) = [\bar{\mu}(X_1, \dots, X_n) - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{\mu}(X_1, \dots, X_n) + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}]$$

ist, dass also

$$h = z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} = z_{0.975} \frac{2}{\sqrt{20}} \approx 0.88$$

gewählt werden kann. Somit erhält man als Annahmebereich

$$A = \{x = (x_1, \dots, x_{20}) : \bar{\mu}(x) \in [11.12, 12.88]\}.$$

3. Experiment durchführen ergibt die Daten  $x = (x_1, \dots, x_{20}) = (9.50, 11.01, 12.97, 12.29, 12.34, 10.80, 13.51, 13.65, 13.01, 11.16, 14.77, 8.87, 11.15, 7.56, 12.77, 12.71, 12.10, 14.24, 9.35, 15.10)$ . Das empirische Mittel ist dann  $\bar{\mu}(x) = 11.94$ .
4. Da  $11.94 \in [11.12, 12.88]$  liegt, folgt somit  $x \in A$ , also wird die Nullhypothese angenommen, bzw. sie wird nicht verworfen – die Daten sprechen also nicht gegen die Wahrheit der Nullhypothese.

■

Wichtig: Die Annahme der Nullhypothese bedeutet nicht, dass sie tatsächlich richtig ist. In obigem Beispiel könnte der echte Erwartungswert auch etwas von der exakten Zahl 12 aus der Nullhypothese abweichen. Die Hypothese  $\mu = 14$  oder  $\mu = 10$  würde bei den vorliegenden Daten hingegen abgelehnt, da diese Werte nicht im Konfidenzintervall für den Erwartungswert liegen.

Die Konstruktion des Annahmebereiches im obigen Beispiel erfüllt das Grundprinzip, welches für Annahmebereiche bei Hypothesentests immer gelten soll: Falls *unter Annahme der Gültigkeit Nullhypothese* die beobachteten Daten *sehr unwahrscheinlich* sind, so verwirft man die Nullhypothese. Im obigen Beispiel wurde deshalb mit der Wahrscheinlichkeit  $\mathbb{P}_{\mu_0}$  gerechnet, d.h. es wurde angenommen, dass die Nullhypothese „der Erwartungswert ist gleich  $\mu_0$ “ gilt. Das Kriterium „sehr unwahrscheinlich“ wurde mit Hilfe des Konfidenzintervalls überprüft. Um dies weiter zu präzisieren, führen wir nun einige weitere Begriffe ein.

### Definition 10.2 (Fehler 1. und 2. Art)

Sei  $H_0$  eine Nullhypothese, und  $A$  ein Annahmebereich. Dann sind zwei Typen von „Fehlern“ möglich:

- **Fehler 1. Art:** Die Nullhypothese wird verworfen, obwohl sie eigentlich richtig wäre
- **Fehler 2. Art:** Die Nullhypothese wird angenommen, obwohl sie eigentlich falsch ist

◆

### Satz 10.1

Sei  $H_0$  eine Nullhypothese, und  $A$  ein Annahmebereich zum Fehlerniveau  $\alpha \in ]0, 1[$ . Dann ist die Wahrscheinlichkeit, einen Fehler 1. Art zu begehen, kleiner oder gleich  $\alpha$ .

**Beweis 10.1** Die Wahrscheinlichkeit für einen Fehler 1. Art ist nach Definition

$$\mathbb{P}(H_0 \text{ wird abgelehnt} \mid H_0 \text{ ist wahr}).$$

Andererseits gilt nach Definition des Annahmebereichs

$$\mathbb{P}((X_1, \dots, X_n) \in A \mid H_0 \text{ ist wahr}) \geq 1 - \alpha.$$

Da gemäss Definition die Nullhypothese genau dann angenommen wird, wenn  $(x_1, \dots, x_n) \in A$  ist, folgt

$$\mathbb{P}(H_0 \text{ wird abgelehnt} \mid H_0 \text{ ist wahr}) = 1 - \mathbb{P}((X_1, \dots, X_n) \in A \mid H_0 \text{ ist wahr}) \leq \alpha.$$

□

Man kann beweisen, dass die Wahrscheinlichkeit für einen Fehler 2. Art umso größer ist, je kleiner die Wahrscheinlichkeit für einen Fehler 1. Art ist. Im weiteren kann der Fehler 2. Art normalerweise nicht explizit quantifiziert werden, auch wenn er durch das Testdesign beeinflusst werden kann, beispielsweise durch Wahl einer größeren Stichprobe. Deshalb ist die Wahl der Nullhypothese es für die Konstruktion eines statistischen Tests eminent wichtig. Die Rollen von Nullhypothese und Alternativhypothese können nicht vertauscht werden, und nur einer der Fehler, nämlich der Fehler 1. Art, kann durch die Wahl des Fehlerniveaus  $\alpha$  explizit kontrolliert werden. Wenn man  $\alpha$  sehr klein wählt, also den Fehler 1. Art sehr unwahrscheinlich macht, hat man im Gegenzug zwingend eine große Unsicherheit hinsichtlich der Wahrscheinlichkeit für einen Fehler 2. Art.

Die Frage, was „sehr unwahrscheinlich“ genau bedeuten soll, kann auf verschiedene Weise beantwortet werden. Im Allgemeinen wird dazu eine sogenannte **Teststatistik** betrachtet. Das ist eine Zufallsvariable  $Y$ , welche von  $X_1, \dots, X_n$  abhängt, also

$$Y = f(X_1, \dots, X_n)$$

für eine Funktion  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . Beispielsweise war in Beispiel 10.1 die Teststatistik das empirische Mittel  $\bar{\mu}$ . Wertet man  $Y$  an der Stelle der Messwerte  $x_1, \dots, x_n$  aus, so bezeichnen wir  $y = f(x_1, \dots, x_n)$  auch als **Testwert**. „Sehr unwahrscheinlich“ wird dann üblicherweise folgendermaßen interpretiert:

- Der Testwert  $y$  liegt außerhalb des Konfidenzintervalls für die Teststatistik zum Fehlerniveau  $\alpha$ . Der Annahmebereich entspricht also dem Annahmebereich des Konfidenzintervalls (vgl. Beispiel 10.1).
- Der Testwert  $y$  ist größer (oder kleiner) als ein geeignetes Quantil  $q_\beta$  der Verteilung der Teststatistik  $Y$ . Der Annahmebereich ist also von der Form  $A = \{(x_1, \dots, x_n) : y = f(x_1, \dots, x_n) \leq (\text{bzw. } \geq) q_\beta\}$ , wobei normalerweise  $\beta = 1 - \alpha$ , bei zweiseitigen Tests  $\beta = 1 - \alpha/2$  ist.
- Der sogenannte  $p$ -Wert der Teststatistik  $Y$  (siehe Definition 10.3) ist größer als das gewählte Fehlerniveau  $\alpha$ . Der Annahmebereich ist dann von der Form  $A = \{(x_1, \dots, x_n) : p(y) \geq \alpha\}$ .

### Definition 10.3 ( $p$ -Wert)

Sei  $Y = f(X_1, \dots, X_n)$  eine Teststatistik, und  $H_0$  eine Nullhypothese diese Daten betreffend. Der **rechtsseitige  $p$ -Wert** für  $y \in Y(\Omega)$  ist definiert als

$$p(y) = \mathbb{P}(Y \geq y \mid H_0 \text{ ist wahr}),$$

der **linksseitige  $p$ -Wert** als

$$p(y) = \mathbb{P}(Y \leq y \mid H_0 \text{ ist wahr}),$$

und der **beidseitige  $p$ -Wert** als

$$p(y) = 2 \min\{\mathbb{P}(Y \geq y \mid H_0 \text{ ist wahr}), \mathbb{P}(Y \leq y \mid H_0 \text{ ist wahr})\}.$$



Der  $p$ -Wert ist die Wahrscheinlichkeit, dass  $Y$  Werte „gleich oder extremer als  $y$ “ annimmt, unter der Annahme dass die Nullhypothese gilt. Dabei bedeutet *extremer* je nach Situation größer und/oder kleiner als der gemessene Wert. Der  $p$ -Wert hängt also mit dem Quantil zusammen, welches durch  $y = f(x_1, \dots, x_n)$  begrenzt wird. Ein großer  $p$ -Wert bedeutet, dass die Messerte  $x_1, \dots, x_n$  relativ typisch für  $X_1, \dots, X_n$  unter  $H_0$  sind, denn nach Definition gibt es dann eine relativ hohe Wahrscheinlichkeit für noch „extremere“ Werte. Umgekehrt bedeutet ein kleiner  $p$ -Wert, dass die gemessenen Daten eher unwahrscheinlich sind. Entsprechend ist bei einem  $p$ -Wert  $p(y)$  zu einer Teststatistik  $Y = f(X_1, \dots, X_n)$  die Menge

$$A_\alpha := \{(x_1, \dots, x_n) : p(f(x_1, \dots, x_n)) \geq \alpha\}$$

ein Annahmehereich zum Fehlerniveau  $\alpha$ .

### Beispiel 10.2 ( $p$ -Wert)

In der Situation von Beispiel 10.1, jedoch mit unbekannter Varianz  $\sigma^2$ , betrachte die Zufallsvariable

$$T := \sqrt{n} \cdot \frac{\bar{\mu}(X_1, \dots, X_n) - \mu_0}{\bar{\sigma}_n(X_1, \dots, X_n)}.$$

Dabei sei  $\mu_0 = 12$  der Erwartungswert unter der Nullhypothese. Nach Satz 9.4 ist diese Zufallsvariable  $t$ -verteilt zum Parameter  $n - 1$ . Mit den Daten aus dem Beispiel gilt  $n = 10$ ,  $\bar{\mu}_n = 11.94$  und  $\bar{\sigma}_n^2 = 4.28$ . Somit ist  $t = \sqrt{n} \cdot \frac{\bar{\mu}(x_1, \dots, x_n) - \mu_0}{\bar{\sigma}_n(x_1, \dots, x_n)} = \sqrt{20} \frac{11.94 - 12}{\sqrt{4.28}} = -0.130$ . Dann ist der linksseitige  $p$ -Wert  $p(t) = \mathbb{P}(T \leq t) = 0.3902$ , und der rechtsseitige  $p(t) = \mathbb{P}(T \geq t) = 0.6098$ . Beides sind sehr hohe Werte, insbesondere ist  $p(t) \geq \alpha$  für alle gängigen Fehlerniveaus, z.B.  $\alpha = 0.05$ . Die Nullhypothese  $\mu_0 = 12$  wird also auch nach diesem Vorgehen angenommen. ■

## 10.2 $t$ -Test

Der  $t$ -Test ist ein Hypothesentest, der auf der  $t$ -Verteilung (Kapitel 9) beruht. Um den Test durchführen zu können, arbeitet man mit Daten  $x_1, \dots, x_n$ , von denen angenommen werden kann, dass sie Realisierungen von *unabhängigen*, identisch verteilten Zufallsvariablen  $X_1, \dots, X_n$  sind, welche außerdem (annähernd) *normalverteilt* sind, z.B. weil  $n$  groß ist und der zentrale Grenzwertsatz gilt. Insbesondere muss die Varianz existieren. Dann können mit Hilfe des  $t$ -Tests die folgenden Hypothesen getestet werden:

- Zweiseitiger  $t$ -Test: Test ob der Erwartungswert  $\mu$  der  $X_i$  gleich einer gegebenen Zahl  $\mu_0$  ist
- Einseitiger  $t$ -Test: Test ob der Erwartungswert  $\mu$  der  $X_i$  größer (kleiner) als eine gegebenen Zahl  $\mu_0$  ist

Außerdem kann mit einer Variante des  $t$ -Tests untersucht werden, ob zwei unabhängige Zufallsvariablen denselben Erwartungswert haben.



Die theoretische Grundlage für den  $t$ -Test bildet Satz 9.4, welcher besagt dass für unabhängige  $\mathcal{N}(\mu, \sigma^2)$ -verteilte  $X_1, \dots, X_n$  die Teststatistik

$$T := \sqrt{n} \cdot \frac{\bar{\mu}_n(X_1, \dots, X_n) - \mathbb{E}[X_1]}{\bar{\sigma}_n(X_1, \dots, X_n)}$$

(annähernd)  $t$ -verteilt mit Parameter  $n-1$  ist. Um den  $t$ -Test durchzuführen, wird deshalb die Testgröße

$$t = \sqrt{n} \cdot \frac{\bar{\mu}_n(x_1, \dots, x_n) - \mu_0}{\bar{\sigma}_n(x_1, \dots, x_n)}$$

in Abhängigkeit von den gemessenen Daten bestimmt. Ist dieser Wert unter der  $t$ -Verteilung sehr unwahrscheinlich, liegt also nicht im Annahmebereich, so wird die Nullhypothese verworfen.

### Vorgehen beim $t$ -Test

Annahme: Daten  $x_1, \dots, x_n$  sind Realisierungen von unabhängigen, identisch normalverteilten Zufallsvariablen  $X_1, \dots, X_n$ . Allgemeines Vorgehen:

1. Aufstellen der Nullhypothese
2. Wahl des Fehlerniveaus  $\alpha \in (0, 1)$
3. Erhebung der Daten  $x_1, \dots, x_n$ .
4. Bestimmung des *Freiheitsgrades*  $f$
5. Bestimmung des *Vergleichswertes*  $t_{\beta, f}$ , eines geeigneten Quantils der  $t$ -Verteilung, abhängig vom Freiheitsgrad
6. Berechnung des *Testwertes*  $t$ :

$$t = t(x_1, \dots, x_n) = \sqrt{n} \cdot \frac{\bar{\mu}(x_1, \dots, x_n) - \mu_0}{\bar{\sigma}_n(x_1, \dots, x_n)}$$

7. Entscheidung ob der Testwert  $t$  im Annahmebereich liegt. Dieser wird durch den Vergleichswert  $t_{\beta, f}$  gegeben. Ablehnung von  $H_0$ , falls  $t$  nicht im Annahmebereich liegt.

**Alternatives Vorgehen mittels  $p$ -Wert:** Statt des Vergleichswertes kann auch der  $p$ -Wert  $p(t)$  zu  $t = t(x_1, \dots, x_n)$  bestimmt werden, und der Annahmebereich  $A_\alpha = \{(x_1, \dots, x_n) : p(t) \geq \alpha\}$  verwendet werden. Beide Vorgehensweisen führen auf das selbe Ergebnis.

Die genaue Wahl von Nullhypothese, die Berechnung von Freiheitsgrad, Testwert, Vergleichswert sowie der Annahmebereich sind abhängig von der Variante des Tests.

### Zweiseitiger $t$ -Test für eine Stichprobe

- Nullhypothese:  $H_0: \mu = \mu_0$
- Freiheitsgrad  $f = n - 1$

- Vergleichswert  $t_{1-\alpha/2,f}$ , das  $1 - \alpha/2$ -Quantil der  $t$ -Verteilung zum Parameter  $f$
- Testwert

$$t = t(x_1, \dots, x_n) = \sqrt{n} \cdot \frac{\bar{\mu}_n(x_1, \dots, x_n) - \mu_0}{\bar{\sigma}(x_1, \dots, x_n)},$$

wobei  $\bar{\mu}_n$  das empirische Mittel und  $\bar{\sigma}_n$  die empirische Varianz der Daten ist.

- Annahmebereich:  $\{(x_1, \dots, x_n) : |t| \leq t_{1-\alpha/2,f}\}$ , also
  - Ist  $|t| \geq t_{1-\alpha/2,f}$ , so wird  $H_0$  verworfen
  - Andernfalls wird  $H_0$  angenommen

**Alternativ:** Statt des Vergleichswertes  $p(t) = 2 \min\{\mathbb{P}(T \geq t), \mathbb{P}(T \leq t)\}$  berechnen,  $T$  eine  $t$ -verteilte Zufallsvariable mit Parameter  $f$  berechnen, Annahmebereich  $\{(x_1, \dots, x_n) : p(t) \geq \alpha\}$ .

### Einseitiger $t$ -Test für eine Stichprobe

- Nullhypothese  $H_0: \mu \leq \mu_0$
- Freiheitsgrad  $f = n - 1$
- Vergleichswert  $t_{1-\alpha,f}$
- Testwert

$$t = t(x_1, \dots, x_n) = \sqrt{n} \cdot \frac{\bar{\mu}_n(x_1, \dots, x_n) - \mu_0}{\bar{\sigma}_n(x_1, \dots, x_n)}$$

- Annahmebereich  $\{(x_1, \dots, x_n) : t \leq t_{1-\alpha,f}\}$ ,

**Alternativ:** Den  $p$ -Wert  $p(t) = \mathbb{P}(T \geq t)$  (zum Parameter  $f$ ) berechnen, Annahmebereich  $\{(x_1, \dots, x_n) : p(t) \geq \alpha\}$ .

Wird beim einseitigen Test als Nullhypothese  $H_0: \mu \geq \mu_0$  gewählt, so geht man genauso vor, nur die Entscheidungsregel wird umgedreht:  $H_0$  wird verworfen falls  $t \leq t_{\alpha,f} = -t_{1-\alpha,f}$  ist, der Annahmebereich ist also  $\{(x_1, \dots, x_n) : t \geq t_{1-\alpha,f}\}$  bzw. es wird der  $p$ -Wert  $p(t) = \mathbb{P}(T \leq t)$  betrachtet.

### Beispiel 10.3 (Wirkung eines Medikaments)

Für ein neues Medikament soll überprüft werden, ob es eine fiebersenkende Wirkung besitzt. Dafür wird die Körpertemperatur von 10 Patienten vor Medikamen-

teneinnahme sowie 3 Stunden danach gemessen. Die Daten sind als Tabelle gegeben:

Nr. Patient	Temp. Einnahme	Temp. 3 h später	Differenz
$i$	$u_i$	$v_i$	$x_i = u_i - v_i$
1	39.1	38.7	0.4
2	38.3	38.1	0.2
3	37.6	37.9	-0.3
4	38.0	37.5	0.5
5	40.1	39.2	0.9
6	39.5	39.1	0.4
7	38.7	38.7	0.0
8	37.9	37.5	0.4
9	39.2	38.2	1.0
10	38.0	37.4	0.6

Es soll ein  $t$ -Test durchgeführt werden. Dazu nehmen wir an, die gemessenen Differenzen  $x_i$  seien Realisierungen von unabhängigen (annähernd) normalverteilten Zufallsvariablen. Dem Medikament wird eine Wirkung zugeschrieben, falls  $\mu := \mathbb{E}[X_i] > 0$  ist, falls also im Mittel die Körpertemperatur nach Einnahme des Medikamentes gefallen ist. Wir führen einen einseitigen  $t$ -Test durch, und wählen als Nullhypothese

$$H_0 : \mu \leq 0 := \mu_0.$$

Zur Durchführung des Tests wählen wir  $\alpha = 0.05$ , da  $n = 10$  ist, erhalten wir  $f = 9$ . Der Testwert ist also  $t_{0.95,9} = 1.833$ . Weiter ist  $\bar{\mu} = \frac{1}{10}(x_1 + \dots + x_{10}) = 0.41$  und  $\bar{\sigma} = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (x_i - 0.41)^2} = 0.39$ . Somit erhalten wir für den Testwert

$$t = \sqrt{n} \frac{\bar{\mu} - \mu_0}{\bar{\sigma}} = \sqrt{10} \frac{0.41 - 0}{0.39} = 3.35.$$

Da  $t > t_{0.95,9}$  ist, wird die Nullhypothese abgelehnt – die Daten widersprechen also der Vermutung, dass  $\mu \leq 0$  ist, und stützen die These, dass das Medikament eine fiebersenkende Wirkung hat. ■

Warum wird in diesem Beispiel Nullhypothese  $\mu \leq 0$  gewählt, und nicht ihr Gegenteil, also  $\mu > 0$ ? Bei der Durchführung des Testes soll die Wahrscheinlichkeit, das Medikament fälschlicherweise als fiebersenkend einzustufen, ihm also eine nicht vorhandene Wirkung zuzusprechen, kontrolliert werden. Dies geht nur dann, wenn der zu kontrollierende Fehler dem Fehler erster Art entspricht. Denn nur dann ist die Wahrscheinlichkeit, dass der die Nullhypothese abgelehnt wird, obwohl sie richtig ist, kleiner oder gleich dem Fehlerniveau  $\alpha$ . Entsprechend muss die Nullhypothese gewählt sein. Mit unserer Wahl wird genau dann ein Fehler erster Art begangen, wenn das Fieber im Mittel sinkt, obwohl das Medikament eigentlich *keine* fiebersenkende Wirkung hätte. Dieser Fehler soll nach Möglichkeit ausgeschlossen oder zumindest quantifiziert werden können. Diese Überlegungen

müssen bei statistischen Tests stets in die Formulierung der Nullhypothese einfließen, da nur für die Wahrscheinlichkeit für den Fehler erster Art eine exakte Schranke, nämlich die Fehlerwahrscheinlichkeit  $\alpha$ , angegeben werden kann.

Der  $t$ -Test kann auch verwendet werden, um zu testen, ob zwei verschiedene Zufallsvariablen denselben Erwartungswert haben (bzw. ob der Erwartungswert der ersten Zufallsvariablen größer oder kleiner als der der zweiten ist). Dabei ist das Vorgehen praktisch gleich, lediglich die Berechnung des Freiheitsgrades und des Testwertes weicht etwas ab.

### Zweiseitiger $t$ -Test für zwei Stichproben

Beim  $t$ -Test mit zwei zu messenden Größen hat man Daten  $x_1, \dots, x_n$  und  $y_1, \dots, y_m$ , von welchen angenommen werden kann, dass sie Realisierungen von unabhängigen Zufallsvariablen  $X_1, \dots, X_n$  und  $Y_1, \dots, Y_m$  sind. Dabei sind die  $X_i$  und die  $Y_i$  jeweils identisch verteilt, und es gilt  $\mathbb{V}(X_i) = \mathbb{V}(Y_i) < \infty$ . Außerdem seien die  $X_i$  und  $Y_i$  zumindest annähernd normalverteilt. Es sei  $\mu_x = \mathbb{E}[X_1]$  und  $\mu_y = \mathbb{E}[Y_1]$ .

- Nullhypothese  $H_0: \mu_x = \mu_y$
- Freiheitsgrad  $f = n + m - 2$
- Vergleichswert  $t_{1-\alpha/2, f}$
- Testwert

$$t = \sqrt{\frac{nm}{n+m}} \frac{\bar{\mu}_x - \bar{\mu}_y}{S},$$

dabei ist  $S := \sqrt{\frac{(n-1)\bar{\sigma}_x^2 + (m-1)\bar{\sigma}_y^2}{n+m-2}}$  mit den jeweiligen empirischen Mitteln  $\bar{\mu}_x, \bar{\mu}_y$  und empirischen Varianzen  $\bar{\sigma}_x^2, \bar{\sigma}_y^2$ .

- Annahmebereich:  $\{t : |t| \leq t_{1-\alpha/2}\}$ .

Im Falle dass man einen einseitigen Test für zwei Größen durchführen will, wird das Vorgehen leicht abgewandelt: Die Nullhypothese lautet  $\mu_x \leq \mu_y$ , sie wird verworfen wenn  $t \geq t_{1-\alpha, f}$  ist. Umgekehrt wird die Nullhypothese  $\mu_x \geq \mu_y$  verworfen, wenn  $t \leq t_{\alpha, f}$  ist. Auch im zweiseitigen Test kann alternativ der  $p$ -Wert mit dem entsprechenden Annahmebereich betrachtet werden.

### Beispiel 10.4 (Wirkung eines Medikamentes)

Die Wirkung eines Präparates zur Förderung der Konzentration soll getestet werden, indem Probanden Aufgaben lösen. Die eine Gruppe erhält das Präparat, die andere ein Placebo. Gezählt wird, wie viele Aufgaben in einer vorgegebenen Zeit gelöst wurden. Es wurden folgende Daten gemessen:

Präparat	18	22	21	17	20	17	23	20	22	21
Kontrollgruppe	16	20	14	21	20	18	13	15	17	21

Es soll ein zweiseitiger  $t$ -Test für zwei Stichproben durchgeführt werden, mit  $x_i$  werden die Daten der Gruppe mit Präparat bezeichnet, mit  $y_i$  die Daten der Kontrollgruppe. Die Nullhypothese lautet dann  $\mu_x = \mu_y$ , also wird unter der Nullhypothese kein Einfluss

des Präparates angenommen. Es ist  $n = m = 10$ , also  $f = 18$ . Bei Wahl von  $\alpha = 0.05$  ist der Testwert  $t_{0.975,18} = 2,1009$ . Für den Testwert erhält man durch Einsetzen

$$t = \sqrt{\frac{10 \cdot 10}{10 + 19}} \frac{\bar{\mu}_x - \bar{\mu}_y}{S} = 2.2573.$$

Somit ist  $t > t_{0.975,18}$ , und die Nullhypothese wird verworfen. Wäre die Nullhypothese wahr, so wäre das beobachtete Ergebnis sehr unwahrscheinlich. Die Wahrscheinlichkeit, ein solches Ergebnis zufälligerweise zu erhalten, ist höchstens  $\alpha = 0.05$ . Vermutlich hat also das Präparat tatsächlich einen (positiven) Einfluss auf das Gedächtnis. ■

## 10.3 $\chi^2$ -Test

Der  $\chi^2$ -Test beruht auf der  $\chi^2$ -Verteilung aus Kapitel 9. Er findet in zwei verschiedenen Situationen Anwendung.

- $\chi^2$ -Test auf Verteilung: Test ob die den gemessenen Daten zugrundeliegenden Zufallsvariablen einer bestimmten Verteilung folgen, z.B. einer Normalverteilung, oder einer Poisson-Verteilung, uniformen Verteilung...
- $\chi^2$ -Test auf Unabhängigkeit: Test ob die gemessenen Daten von Paaren von unabhängigen Zufallsvariablen stammen.

Der Test auf Verteilung beruht darauf, dass die Daten in *Klassen* gruppiert werden, also ähnliche Werte zusammengefasst werden. Für jede Klasse wird Häufigkeit bestimmt, d.h. gezählt, wie viele Messwerte in dieser Klasse enthalten sind. Sind die Klassen hinreichend groß, so sind diese Zufallsvariablen annähernd normalverteilt. Unter der Nullhypothese, dass die Daten einer bestimmten Verteilung folgen, können die theoretischen Häufigkeiten unter dieser Annahme bestimmt werden. Falls die Nullhypothese wahr ist, folgt nach Satz 9.3 die Summe der mittleren quadratischen Abweichungen der gemessenen Häufigkeiten vom theoretischen einer  $\chi^2$ -Verteilung. Mit Wahrscheinlichkeit  $1 - \alpha$  weichen unter der Nullhypothese die gemessenen Häufigkeiten um höchstens das  $1 - \alpha$ -Quantil der  $\chi^2$ -Verteilung vom echten Wert ab. Eine größere Abweichung ist also sehr unwahrscheinlich und deutet darauf hin, dass die Nullhypothese falsch ist, weshalb sie in diesem Fall abgelehnt wird. Wie beim  $t$ -Test muss ein geeigneter Testwert mit dem entsprechenden Quantil verglichen werden (oder alternativ der  $p$ -Wert mit dem Fehlniveau  $\alpha$ ).

### $\chi^2$ -Test auf Verteilung

Es seien Daten  $x_1, \dots, x_n$  gegeben, von denen angenommen werden kann, dass sie Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen  $X_1, \dots, X_n$  sind. Vorgehen beim Test auf Verteilung:

1. Aufstellen der Nullhypothese:  $H_0$  : Die Daten folgen einer bestimmten Verteilung (z.B. Normalverteilung, Poisson-Verteilung, uniforme Verteilung...).

2. Wahl von  $\alpha \in (0, 1)$
3. Falls nötig Schätzung der Parameter der Verteilung aus den Messwerten.
4. Gruppieren der Daten in  $k$  Klassen  $A_1, \dots, A_k$ , wobei die Klassen  $A_i$  Teilmengen von  $\mathbb{R}$  sind, so dass  $A_i \cap A_j = \emptyset$  gilt, und so dass jeder Messwert  $x_j$  in einer Klasse  $A_i$  enthalten ist. Bestimmung der Häufigkeiten

$$N_i := |\{j : x_j \in A_i\}|, i = 1, \dots, k.$$

5. Freiheitsgrad:  $f = k - m - 1$ , wobei  $m$  die Anzahl geschätzter Parameter in der angenommenen Verteilung ist.
6. Vergleichswert:  $\chi^2_{1-\alpha, f}$ , das  $1 - \alpha$ -Quantil der  $\chi^2$ -Verteilung mit Parameter  $f$ .
7. Berechnung der *theoretischen Häufigkeiten* unter der Nullhypothese:

$$F_i = n \cdot \mathbb{P}(X \in A_i | H_0), i = 1, \dots, k$$

8. Testwert:

$$\chi^2 := \sum_{i=1}^k \frac{(N_i - F_i)^2}{F_i}.$$

9. Vergleich des Testwerts mit dem Vergleichswert. Annahmehereich:  $\{\chi^2 : \chi^2 \leq \chi^2_{1-\alpha, f}\}$ .

### Beispiel 10.5 (Test auf diskrete Gleichverteilung)

Daten: Gewinnspiel einer Getränkemarkte zur Fußball-WM. Sammelbilder zu den teilnehmenden Ländern, eingeteilt in die Gruppen A bis H. Mit allen Bildern der Gruppen A und B gewinnt man den Hauptpreis, mit allen Bildern einer anderen Gruppe gewinnt man Trostpreise. Fragestellung: Sind alle Gruppen gleich häufig vertreten?

1. Nullhypothese:  $H_0$  : Die Daten sind diskret gleichverteilt. Da bei einer diskreten Gleichverteilung der Parameter durch die Anzahl Gruppen gegeben ist, hier also  $k = 8$ , muss kein Parameter geschätzt werden.
2. Wähle das Fehlniveau  $\alpha = 0.05$
3. Daten (gesammelte Bilder): Gesamtzahl  $n = 30$ . Einteilung in die 8 Gruppen,  $k = 8$ , ist bereits in natürlicher Weise vorgegeben.

Gruppe $A_i$	A	B	C	D	E	F	G	H
Häufigkeit $N_i$	2	4	2	3	3	7	5	4

4. Da  $k = 8$  ist und kein Parameter geschätzt wurde, ist der Freiheitsgrad  $f = 8 - 1 = 7$ .
5. Aus der Tabelle bestimmen wir das 0.95-Quantil der  $\chi^2$ -Verteilung mit Parameter  $n = 7$ . Dieses ist  $\chi^2_{0.95, 7} = 14.07$ .
6. Falls die Daten uniform verteilt sind, hat jede Gruppe die gleiche Häufigkeit. Somit sind die theoretische Häufigkeiten gegeben durch

$$F_i = 30/8 = 3.75, i = 1, \dots, 8$$

7. Daraus berechnet sich nun der Testwert als

$$\chi^2 = \sum_{i=1}^k \frac{(N_i - F_i)^2}{F_i} = \frac{1}{3.75} \left( (2 - 3.75)^2 + (4 - 3.75)^2 + \dots + (4 - 3.75)^2 \right) = 5.2$$

8. Somit ist der Testwert  $\chi^2 = 5.2 < 14.07 = \chi_{0.95,7}^2$ , die Nullhypothese wird also angenommen. ■

### Beispiel 10.6 (Test auf Normalverteilung)

Es werden 80 Werte erhoben, von denen vermutet wird, dass die zu Grunde liegenden Zufallsvariablen einer Normalverteilung folgt. Die Daten werden in Klassen gruppiert.

Klasse $i$	1	2	3	4	5	6	7	8	9
Bereich	<-5	[-5,0)	[0,5]	[5,10]	[10,15]	[15,20]	[20,25]	[25,30]	>30
Anzahl $N_i$	3	14	33	6	5	4	5	4	6

Weiter wurde das empirische Mittel und die empirische Varianz aus den Originaldaten berechnet, mit den Ergebnissen  $\bar{\mu} = 7.8$  und  $\bar{\sigma}^2 = 115$ .

1.  $H_0$  : Die Daten sind normalverteilt.
2.  $\alpha = 0.05$
3. Parameterschätzung ergab  $\bar{\mu} = 7.8$  und  $\bar{\sigma}^2 = 115$ , d.h.  $\bar{\sigma} = 10.7$ .
4. Klasseneinteilung in der Tabelle.
5. Freiheitsgrad  $f = 9 - 2 - 1 = 6$ , da es 9 Klassen und 2 geschätzte Parameter gibt
6.  $\chi_{0.95,6}^2 = 12.59$
7. Theoretische Häufigkeiten unter  $H_0$ : Für die Klasse Nr. 1, also die Menge  $] -\infty, -5[$  ist die theoretische Häufigkeit bei 80 Werten

$$F_1 = 80 \cdot \mathbb{P}(Y < 5),$$

wobei  $Y$  eine normalverteilte Zufallsvariable mit Parametern  $\mu = \bar{\mu}$  und  $\sigma^2 = \bar{\sigma}^2$  ist. Also genau die Anzahl Werte, welche man bei Gültigkeit der Nullhypothese im entsprechenden Intervall zu finden erwartet. Somit ist für  $Z$  standardnormalverteilt

$$F_1 = 80 \cdot \mathbb{P}\left(Z < \frac{-5 - 7.8}{10.7}\right) = 80 \cdot \Phi_{0,1}(-1.196) = 80 \cdot 0.1357 = 9.3.$$

Analog werden die anderen theoretischen Häufigkeiten berechnet. Man findet

$i$	1	2	3	4	5	6	7	8	9
Bereich	<-5	[-5,0)	[0,5]	[5,10]	[10,15]	[15,20]	[20,25]	[25,30]	>30
$F_i$	9.3	9.3	13.2	14.7	13.4	9.9	5.9	2.8	1.5

8. Für den Testwert erhält man damit

$$\chi^2 = \frac{(3 - 9.3)^2}{9.3} + \dots + \frac{(6 - 1.5)^2}{1.5} = 53.94 > 12.59 = \chi_{0.95,6}^2.$$

Die Nullhypothese wird also klar verworfen. Stellt man die gemessenen Häufigkeiten als Histogramm dar, so ist auch visuell zu erkennen, dass die beobachteten Häufigkeiten nicht einer Normalverteilung ähneln.



Der zweite Anwendungsbereich des  $\chi^2$ -Tests liegt darin, zu testen, ob Paare von Daten von unabhängigen Zufallsvariablen herrühren.

Beim  $\chi^2$ -Test auf Unabhängigkeit sind *Paare* von Messwerten  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  gegeben, für welche angenommen werden kann, dass sie Realisierungen von Zufallsvariablen  $X_1, Y_1, X_2, Y_2, \dots, X_n, Y_n$  sind, wobei die  $X_1, \dots, X_n$  jeweils untereinander unabhängig und identisch verteilt sind, und die  $Y_1, \dots, Y_n$  ebenfalls untereinander unabhängig und identisch verteilt sind. Die Verteilung der  $X_i$  und der  $Y_i$  muss dabei nicht bekannt und auch nicht gleich sein. Zu beachten ist, dass nicht gefordert wird, dass die  $X_i$  und  $Y_i$  jeweils voneinander unabhängig sind. Die Fragestellung im Test auf Unabhängigkeit lautet jeweils, zu überprüfen, ob die  $X_i$  unabhängig von den  $Y_i$  sind.

Auch bei dieser Fragestellung gruppiert man die Daten, und bestimmt die entsprechenden Häufigkeiten, welche man mit den theoretischen Häufigkeiten unter der Nullhypothese der Unabhängigkeit berechnet. Wiederum folgen nach Satz 9.3 die quadratischen Abweichungen der gemessenen und der theoretischen Häufigkeiten einer  $\chi^2$ -Verteilung.

### Vorgehen beim $\chi^2$ -Test auf Unabhängigkeit

In der obigen Situation wird der  $\chi^2$ -Test auf Unabhängigkeit folgendermaßen durchgeführt:

1. Nullhypothese:  $H_0$  : Die  $X_i$  sind unabhängig von den  $Y_i$ .
2. Wahl des Fehlerniveaus  $\alpha \in (0, 1)$ .
3. Gruppieren der  $x$ -Werte in  $k$  verschiedene Gruppen, und die  $y$ -Werte in  $m$  verschiedene Gruppen. Bestimme die Häufigkeiten

$$N_{i,j} := |\{l : x_l \text{ gehört zur Gruppe } i, y_l \text{ gehört zur Gruppe } j\}|$$

Aufstellen der *Kontingenztafel* mit diesen Häufigkeiten (siehe Tabelle 10.1). Berechnung der Randhäufigkeiten  $N_{i,*} := \sum_{j=1}^m N_{i,j}$  und  $N_{*,j} := \sum_{i=1}^k N_{i,j}$ .

4. Freiheitsgrad:  $f = (k - 1)(m - 1)$
5. Vergleichswert:  $\chi^2_{1-\alpha, f}$ , das  $1 - \alpha$ -Quantil der  $\chi^2$ -Verteilung mit Parameter  $f$
6. Berechnung der theoretischen Häufigkeiten unter der Nullhypothese:

$$F_{i,j} = \frac{N_{i,*} \cdot N_{*,j}}{n}$$

7. Testwert  $\chi^2 = \sum_{i,j} \frac{(F_{i,j} - N_{i,j})^2}{F_{i,j}}$
8. Vergleich von Testwert und Vergleichswert. Annahmereich:  $\{\chi^2 : \chi^2 \leq \chi^2_{1-\alpha, f}\}$ .

Dabei bezeichnet  $N_{i,j}$  die Anzahl der Paare mit  $x$ -Wert  $i$  und  $y$ -Wert  $j$ ,  $N_{i,*}, N_{*,j}$  die Randhäufigkeiten, d.h.  $N_{i,*}$  die Anzahl Paare mit  $x$ -Wert  $i$ , und  $N_{*,j}$



$\begin{matrix} \diagup \\ X \backslash Y \end{matrix}$	1	2	...	$m$	Total
1	$N_{1,1}$	$N_{1,2}$	...	$N_{1,m}$	$N_{1,*}$
2	$N_{2,1}$				
$\vdots$					
$k$	$N_{k,1}$			$N_{k,m}$	$N_{k,*}$
Total	$N_{*,1}$			$N_{*,m}$	$n$

Tab. 10.1: Kontingenztabelle für den  $\chi^2$ -Test auf Unabhängigkeit.

die Anzahl Paare mit  $y$ -Wert  $j$ . Diese Randhäufigkeiten werden zur Berechnung der theoretischen Häufigkeiten verwendet. Sind die Zufallsvariablen unabhängig, so sind nach Definition die theoretischen Häufigkeiten gleich dem Produkt der Randhäufigkeiten (vgl. Kapitel 4).

**Beispiel 10.7 (Unabhängigkeit von Genveränderungen)**

100 Personen werden auf zwei verschiedene Genveränderungen (A und B) getestet. Dabei findet man folgende Häufigkeiten:

$N_{i,j}$	A vorhanden	A nicht vorhanden	Total
B vorhanden	16	14	30
B nicht vorhanden	24	46	70
Total	40	60	100

Mittels eines  $\chi^2$ -Tests soll nun ermittelt werden, ob das Auftreten der beiden Genveränderungen unabhängig voneinander ist.

1. Nullhypothese: Die Merkmale sind unabhängig.
2. Es sollen zwei  $\alpha$  untersucht werden:  $\alpha = 0.05$  und  $\alpha = 0.1$ .
3. Die Kontingenztabelle ist bereits gegeben.
4. Wir haben  $k = m = 2$ , also ist  $f = 1$ .
5. Die Quantile sind  $\chi^2_{0.95,1} = 3.84$  und  $\chi^2_{0.9,1} = 2.71$
6. Theoretische Häufigkeiten  $F_{i,j}$  unter der Nullhypothese, bei gleicher Gesamtzahl der einzelnen Merkmale. Die Randhäufigkeiten sind also dieselben wie in der Kontingenztabelle der gemessenen Häufigkeiten, und die Werte im Inneren der Tabelle berechnen sich durch Multiplikation der jeweiligen Randwerte, normiert mit der Gesamtzahl, also  $F_{AB} = \frac{30 \cdot 40}{100} = 12$ . Insgesamt erhält man

$F_{i,j}$	A vorhanden	A nicht vorhanden	Total
B vorhanden	12	18	30
B nicht vorhanden	28	42	70
Total	40	60	100

7. Der Testwert wird somit

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(N_{i,j} - F_{i,j})^2}{F_{i,j}} = \frac{(16 - 12)^2}{12} + \frac{(14 - 18)^2}{18} + \frac{(24 - 28)^2}{28} + \frac{(46 - 42)^2}{42} = 3.17$$

8. Es ist  $\chi^2 < \chi_{0.95,1}^2$ , aber  $\chi^2 > \chi_{0.9,1}^2$ . Somit wird die Nullhypothese für  $\alpha = 0.05$  angenommen, jedoch für  $\alpha = 0.1$  verworfen. ■

### Zusammenfassung

Mittels Hypothesentests oder auch mittels  $p$ -Werten kann überprüft werden, ob gemessene Daten einer Annahme über die zu Grunde liegenden Zufallsvariablen so stark widersprechen, dass diese Annahme als falsch verworfen werden muss. Je nach Situation werden unterschiedliche Tests angewandt. Besonders häufig werden  $t$ -Test und  $\chi^2$ -Test verwendet. Dabei werden stets nach Aufstellen einer Nullhypothese und Wahl eines Fehlerniveaus ein dem Problem angepasster Testwert mit einem geeigneten Quantil der passenden Verteilung verglichen, und je nach Ergebnis die Nullhypothese verworfen.

## Aufgaben

**10.1** Zeigen Sie: Wird eine Nullhypothese mit  $\alpha = 0.01$  verworfen, so wird sie (bei den selben Daten) auch für  $\alpha = 0.05$  verworfen.

**10.2** In einer Abfüllstation soll die Abfüllmenge von Halbliterflaschen mittels eines  $t$ -Tests überprüft werden. Dazu werden 75 Flaschen abgefüllt und nachvermessen. Die Füllmenge wird als normalverteilt angenommen. Man findet als empirisches Mittel der Füllmenge  $\bar{\mu} = 0.5099$ , und als empirische Varianz  $\bar{\sigma}^2 = 0.01$ . Bestimmen Sie den  $p$ -Wert zu diesem Problem. Für welche Wahl von  $\alpha$  wird die Nullhypothese verworfen, für welche angenommen?

**10.3** Gegeben seien Daten  $x = (69, 70, 71, 68, 67, 70, 70, 70, 67, 69)$ . Mittels eines  $t$ -Tests zum Fehlerniveau  $\alpha = 0.1$  bzw.  $\alpha = 0.05$  untersuche man die Frage, ob der Erwartungswert der zugehörigen Zufallsvariablen  $\mu_0 = 70$  ist.

**10.4** In der Situation von Aufgabe 10.3 unteruche man mittels eines einseitigen  $t$ -Tests, ob der Erwartungswert kleiner als  $\mu_0 = 70$  ist.

**10.5** Mittels eines  $\chi^2$ -Tests auf Gleichverteilung soll eine Münze auf die Frage untersucht werden, ob sie fair ist. Dazu wird die Münze 10 Mal geworfen, und die Anzahl Köpfe mit  $k$  bezeichnet werden. Stellen Sie eine geeignete Nullhypothese auf und bestimmen Sie

Freiheitsgrad und Testwert. Für welche Werte von  $k$  wird die Nullhypothese bei einem Fehlniveau von  $\alpha = 0.05$  verworfen?

**10.6** Untersuchen Sie mittels eines  $\chi^2$ -Testes zum Fehlniveau  $\alpha = 0.01$ , ob für die Daten aus Aufgabe 9.3 die dort gemachte Normalverteilungsannahme gerechtfertigt ist.

**10.7** Mittels eines statistischen Tests soll die Frage untersucht werden, ob Rauchen abhängig vom Geschlecht ist. In einer Umfrage unter 100 Personen wurde 19 rauchende Männer, 48 nichtrauchende Männer, 12 rauchende Frauen und 31 nichtrauchende Frauen gezählt. Stellen Sie eine geeignete Nullhypothese auf, und testen Sie dieses mit Hilfe eines geeigneten Tests zum Fehlniveau  $\alpha = 0.05$ .

**10.8** Im Versandhandel wird die Bearbeitungszeit von Aufträgen an jeweils 1000 Aufträgen in zwei Monaten überprüft. In Monat 1 beträgt die mittlere Zeit 54 Sekunden, in Monat 2 50 Sekunden. Die Varianz liegt jeweils bei 30 Sekunden. Überprüfen Sie mittels eines geeigneten Tests zum Fehlniveau  $\alpha = 0.05$  ob sich die Bearbeitungszeiten in Monat 1 und Monat 2 signifikant unterscheiden.



## Teil III

# Markov-Ketten und stochastische Algorithmen



# 11 Markov-Ketten

---

## Übersicht

11.1 Markov-Ketten und stochastische Matrizen .....	154
11.2 Invariante Verteilungen .....	160
11.3 Langzeitverhalten von Markov-Ketten .....	164
11.4 Anwendungen .....	166

---

Markov-Ketten modellieren den zeitlichen Verlauf gewisser Prozesse, bei denen zufällige Übergänge ausgeführt werden. Zu ihrer Beschreibung werden stochastische Matrizen verwendet. In diesem Kapitel lernen wir die wichtigsten Aspekte der Theorie von Markov-Ketten in diskreter Zeit kennen, insbesondere was deren Langzeitverhalten angeht. Anwendungen gibt es beispielsweise in der Modellierung von Wartesystemen oder im PageRank-Algorithmus von Google.

### Lernziele

- Den Begriff der Markov-Kette und den Zusammenhang zu stochastischen Matrizen kennen
- Mit stochastischen Matrizen rechnen können
- Invariante Verteilungen berechnen können, Kriterien für Existenz und Eindeutigkeit kennen, ihre Bedeutung für Markov-Ketten kennen
- Anwendungen von Markov-Ketten kennen.

### Vorkenntnisse

In diesem Kapitel spielen insbesondere bedingte Wahrscheinlichkeiten eine wichtige Rolle. Außerdem ist der souveräne Umgang mit Matrizen und linearen Gleichungssystemen wichtig, sowie mit Folgen und Grenzwerten.

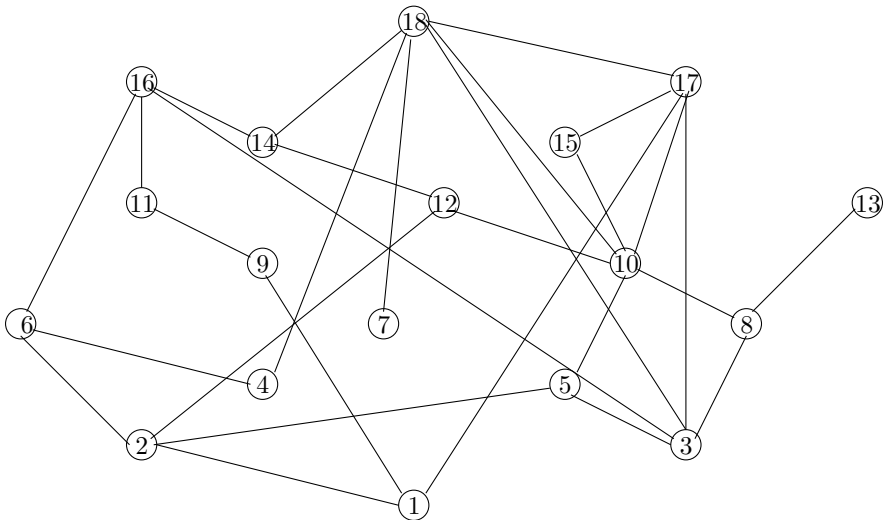


Abb. 11.1: Netzwerk aus Knoten und Kanten

## 11.1 Markov-Ketten und stochastische Matrizen

Als erstes Beispiel betrachten wir ein Netzwerk, welches aus Knoten und Kanten besteht. Eine Kanten verbindet jeweils zwei Knoten miteinander, jedoch hat nicht jeder Knoten gleich viele Verbindungen.

Wir möchten die Struktur des Netzwerks untersuchen, haben aber nicht die Möglichkeit, das ganze Netzwerk zu betrachten (z.B. weil wir uns selbst innerhalb des Netzwerks befinden). Wir können uns jedoch innerhalb des Netzwerks bewegen, indem wir von Knoten zu Knoten springen, falls dazwischen eine Kante liegt. Folgendes soll gelten:

- Von einem Knoten innerhalb des Netzwerks sieht man nur seine direkten Nachbarn
- Von einem Knoten aus springt man zu einem zufällig ausgewählten Nachbarknoten, und zwar wählt man jeden Nachbarknoten mit gleicher Wahrscheinlichkeit.

Dieses einfache Beispiel bildet es die Grundlage des Google-Algorithmus „PageRank“. Dabei stellen die Knoten einzelne Websites dar, und die Kanten die Verbindungen dazwischen, also die Links. Mehr dazu in Beispiel 11.11. Man kann sich beispielsweise folgende Fragen stellen:

- Wie lange dauert es, bis man jeden Knoten einmal besucht hat?
- Wie lange dauert es, bis man wieder beim Ausgangspunkt ist?
- Wie oft besucht man (auf lange Sicht) einen bestimmten Knoten/einen durchschnittlichen Knoten? Besucht man alle gleich oft?
- Welche Rolle spielt die Wahl des Startpunktes in den obigen Fragen?
- Was kann man aus solchen Informationen über die Struktur des Netzwerks sagen?

Um dieses Beispiel mathematisch zu behandeln, soll es zuerst etwas formalisiert werden.



- Ein Netzwerk wird beschrieben als ein *Graph* bestehend aus Knoten und Kanten, die Knoten sind nummeriert als  $1, 2, \dots, k$ .
- Mit  $X_0$  bezeichnen wir den Startknoten angibt, d.h.  $X_0$  ist eine Zufallsvariable mit Werten in  $\{1, \dots, k\}$ .
- Wähle zufällig (gleichverteilt) einen Nachbarknoten aus (d.h. einen Knoten der mit  $X_0$  durch eine Kante verbunden ist), und springe dorthin. Bezeichne diesen Knoten mit  $X_1$ .
- Iterativ:  $X_n$  sei die Nummer des Knoten, die man nach diesem Verfahren im  $n$ -ten Schritt besucht.

Dieses Vorgehen beschreibt einen *stochastischen Algorithmus*, dessen Ausgabe eine Folge von Zufallsvariablen  $X_0, X_1, \dots$  ist, welche jeweils Werte in  $\{1, \dots, k\}$  annehmen. Diese Folge hat eine ganz bestimmte Eigenschaft, die wir herausheben wollen, da sie in diesem Kapitel entscheidend ist:

Der Wert bzw. die Verteilung von  $X_n$  hängt vom Wert von  $X_{n-1}$  ab (und der zufälligen Auswahl beim Sprung), aber nicht von  $X_{n-2}, X_{n-3}, \dots$ . In anderen Worten: Welchen Knoten man im nächsten Schritt besucht, hängt davon ab, wo man sich gerade befindet, jedoch von den Knoten die man davor schon besucht hat. Außerdem wird der nächste Knoten immer nach der gleichen zufälligen Regel ausgewählt, egal wie viele Sprünge man schon gemacht hat. Diese Eigenschaften sind in der folgenden Definition formalisiert.

### Definition 11.1 (Markov-Kette)

Sei  $S$  eine Menge. Eine **Markov-Kette** auf  $S$  ist eine Folge von Zufallsvariablen  $X_0, X_1, X_2, \dots$  mit Werten in  $S$ , so dass für alle  $n \in \mathbb{N}_0$  und für alle  $a_0, a_1, \dots, a_n \in S$  gilt:

$$\begin{aligned} \mathbb{P}(X_n = a_n \mid X_0 = a_0, \dots, X_{n-1} = a_{n-1}) &= \mathbb{P}(X_n = a_n \mid X_{n-1} = a_{n-1}) \\ &= \mathbb{P}(X_1 = a_n \mid X_0 = a_{n-1}). \end{aligned}$$

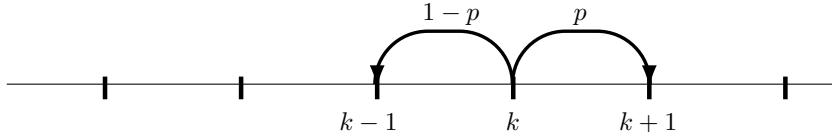
Die Menge  $S$  heißt **Zustandsraum** der Markov-Kette, wir nehmen stets an, dass sie endlich oder abzählbar unendlich sei. Ein Element  $a \in S$  heißt **Zustand**. ♦

Der eingangs beschriebene Prozess auf dem Netzwerk aus Abbildung 11.1 ist also eine Markov-Kette.

### Beispiel 11.1 (Irrfahrten auf $\mathbb{Z}$ bzw. $\mathbb{Z}^d$ )

Ein Beispiel für eine Markov-Kette mit unendlichem Zustandsraum ist die sogenannte *Irrfahrt* auf den ganzen Zahlen  $\mathbb{Z}$ . Diese kann man sich als unendlich großen Graphen vorstellen, wobei die Knoten gegeben sind durch die einzelnen Zahlen, und die Kanten sich jeweils zwischen benachbarten Zahlen befinden. Startet man also in  $k \in \mathbb{Z}$ , z.B.  $k = -217$ , so springt man mit jeweils gleicher Wahrscheinlichkeit zu einem Nachbarknoten, in diesem Fall sind das  $-216$  und  $-218$ . Formal ausgedrückt:

$$\mathbb{P}(X_n = k + 1 \mid X_{n-1} = k) = \mathbb{P}(X_n = k - 1 \mid X_{n-1} = k) = \frac{1}{2}, \quad \text{für alle } k \in \mathbb{Z}, n \in \mathbb{N}.$$

Abb. 11.2: Irrfahrt auf  $\mathbb{Z}$ 

Da man mit gleicher Wahrscheinlichkeit nach links bzw. rechts springt, nennt man das auch die *symmetrische Irrfahrt* auf  $\mathbb{Z}$ . Die *asymmetrische Irrfahrt* springt mit einer Wahrscheinlichkeit  $p \in [0, 1]$  nach links, und mit der Gegenwahrscheinlichkeit nach rechts, also

$$\mathbb{P}(X_n = k + 1 \mid X_{n-1} = k) = p, \quad \mathbb{P}(X_n = k - 1 \mid X_{n-1} = k) = 1 - p, \quad k \in \mathbb{Z}, n \in \mathbb{N}.$$

Auf dem zweidimensionalen regelmäßigen Gitter  $\mathbb{Z}^2$  hat man von jedem Knoten aus 4 Möglichkeiten (Abb. 11.3), somit springt die symmetrische Irrfahrt in diesem Fall mit Wahrscheinlichkeit  $\frac{1}{4}$  zu jedem der 4 Nachbarknoten. Analog in höheren Dimensionen. ■

### Definition 11.2

Sei  $X_0, X_1, \dots$  eine Markov-Kette auf einem Zustandsraum  $S$ . Seien  $a, b \in S$ . Dann heißt

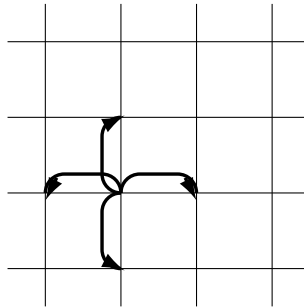
$$p_{a,b} := \mathbb{P}(X_n = b \mid X_{n-1} = a)$$

**Übergangswahrscheinlichkeit** von  $a$  nach  $b$ . Nach Definition 11.1 ist sie unabhängig von  $n$ . ♦

### Beispiel 11.2 (Symmetrische Irrfahrt auf $\mathbb{Z}^d$ )

Betrachte die symmetrische Irrfahrt auf  $\mathbb{Z}^d, d \geq 1$  siehe Beispiel 11.1. Dann gelten für die Übergangswahrscheinlichkeiten

$$p_{xy} = \begin{cases} \frac{1}{2d} & \text{falls } x \text{ und } y \text{ Nachbarn sind} \\ 0 & \text{sonst.} \end{cases}$$

Abb. 11.3: Symmetrische Irrfahrt auf  $\mathbb{Z}^2$

**Beispiel 11.3 (Irrfahrt auf Graphen)**

Wir betrachten den eingangs beschriebenen Prozess auf dem Graphen aus Abbildung 11.1. Vom jeweiligen Ausgangsknoten springt die Kette entlang einer der Kanten zu einem angrenzenden Knoten, wobei jede Kante mit gleicher Wahrscheinlichkeit benutzt wird. Die Übergangswahrscheinlichkeiten können dann durch Zählen der Nachbarknoten abgelesen werden. Für Knoten 1, welcher drei Nachbarknoten hat, ergibt sich so eine Wahrscheinlichkeit von  $1/3$  für jeden davon, also

$$p_{1,2} = p_{1,9} = p_{1,17} = \frac{1}{3}.$$

Alle anderen Übergänge ausgehend von Knoten Nr. 1 haben Wahrscheinlichkeit 0. Knoten 18 hat 6 Nachbarn, also

$$p_{18,3} = p_{18,4} = p_{18,7} = p_{18,10} = p_{18,14} = p_{18,17} = \frac{1}{6}.$$

Knoten 13 hat nur einen Nachbarn, man springt also zwingend zu diesem, also  $p_{13,8} = 1$ . ■

**Definition 11.3**

Die **Übergangsmatrix** einer Markovkette auf einem endlichen Zustandsraum  $S = \{a_1, \dots, a_K\}$  ist gegeben durch die quadratische Matrix

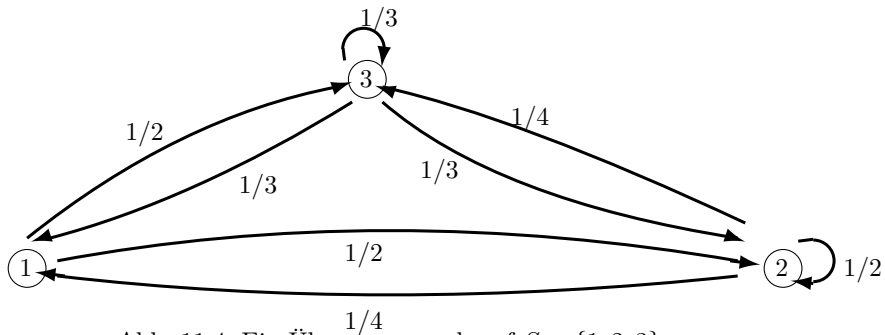
$$P := (p_{a_n, a_m})_{n,m=1,\dots,K} = \begin{bmatrix} p_{a_1, a_1} & p_{a_1, a_2} & \cdots & p_{a_1, a_K} \\ p_{a_2, a_1} & p_{a_2, a_2} & \cdots & p_{a_2, a_K} \\ \vdots & & \ddots & \vdots \\ p_{a_K, a_1} & p_{a_K, a_2} & \cdots & p_{a_K, a_K} \end{bmatrix},$$

wobei die  $p_{a,b}$ ,  $a, b \in S$  die jeweiligen Übergangswahrscheinlichkeiten der Kette sind. In der ersten Zeile stehen also die Übergangswahrscheinlichkeiten für Sprünge, welche von Zustand  $a_1$  ausgehen, in der zweiten Zeile für Sprünge ab  $a_2$ , etc. ♦

**Beispiel 11.4 (Übergangsgraph)**

Markov-Ketten und ihre Übergangswahrscheinlichkeiten können gut an Hand ihres *Übergangsgraphen* dargestellt werden, indem man mögliche Übergänge mit Pfeilen zwischen den Zuständen markiert, und die entsprechenden Wahrscheinlichkeiten über die Pfeile schreibt. Hier ein einfaches Beispiel mit einem dreielementigen Zustandsraum  $S = \{1, 2, 3\}$ . Die Übergangswahrscheinlichkeiten seien mit Hilfe der Übergangsmatrix angegeben:

$$P := \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

Abb. 11.4: Ein Übergangsgraph auf  $S = \{1, 2, 3\}$ 

Es gilt also z.B. dass der Übergang von Zustand 1 in den Zustand 2 die Wahrscheinlichkeit  $p_{1,2} = 1/2$  hat (erste Zeile, zweite Spalte). Man beachte, dass beispielsweise  $p_{3,3} = 1/3 \neq 0$  ist, dass die Kette also mit positiver Wahrscheinlichkeit von 3 nach 3 springen kann, beziehungsweise in 3 stehen bleibt. Als Übergangsgraph sieht dies aus wie in Abbildung 11.4 dargestellt. Übergänge mit Wahrscheinlichkeit 0 werden in der Skizze normalerweise weggelassen.

■

### Satz 11.1 (Stochastische Matrix)

Sei  $P$  die Übergangsmatrix einer Markov-Kette. Dann gelten

(a)  $p_{a,b} \leq 1$  für alle  $a, b \in S$ , und

(b)  $\sum_{b \in S} p_{a,b} = 1 \quad \forall a \in S$ .

**Beweis 11.1** Nach Definition gilt  $p_{a,b} = \mathbb{P}(X_1 = b \mid X_0 = a)$ , also muss  $p_{a,b}$  als Wahrscheinlichkeit zwischen 0 und 1 liegen. Außerdem haben wir

$$\sum_{b \in S} p_{ab} = \sum_{b \in S} \mathbb{P}(X_1 = b \mid X_0 = a) = \mathbb{P}(X_1 \in S \mid X_0 = a),$$

und die rechte Seite ist einfach die Wahrscheinlichkeit, bei Start in  $a$  nach einem Schritt in einem beliebigen Zustand zu sein. Da man aber auf jeden Fall in irgend einem Zustand sein muss, ist die rechte Seite gleich 1.  $\square$

Eine Matrix  $P$  mit den Eigenschaften (a) und (b) aus Satz 11.1 heißt *stochastische Matrix*. So wie man zu jeder Markov-Kette eine Übergangsmatrix finden kann, welche diese beiden Eigenschaften hat, existiert auch zu jeder stochastischen Matrix eine entsprechende Markov-Kette.

Die Einträge der Übergangsmatrix geben nach Definition die Sprungwahrscheinlichkeiten in einem Schritt der Markov-Kette an. Interessiert man sich für längere Zeitabstände, kommen die Potenzen der Übergangsmatrix ins Spiel.

**Satz 11.2 (Mehrstufige Übergangswahrscheinlichkeiten)**

Sei  $P$  die Übergangsmatrix einer Markov-Kette auf einem endlichen Zustandsraum, und sei

$$P^n = P \cdot \dots \cdot P$$

die  $n$ -te Matrix-Potenz von  $P$ , von der Form  $P^n = (p_{a,b}^{(n)})_{a,b \in S}$ . Dann gilt für die  $n$ -Schritt-Übergangswahrscheinlichkeiten:

$$\mathbb{P}(X_n = b \mid X_0 = a) = p_{a,b}^{(n)},$$

d.h. die Wahrscheinlichkeit, dass die Kette in genau  $n$  Schritten von  $a$  nach  $b$  gelangt, ist gegeben durch den Eintrag an der Stelle  $(a, b)$  der  $n$ -ten Potenz der Übergangsmatrix.

**Beweis 11.2** Dies folgt aus der Formel von der Gesamtwahrscheinlichkeit. Wir verzichten auf die formalen Details.  $\square$

**Definition 11.4**

Für  $a \in S$  definieren wir  $\nu_a := \mathbb{P}(X_0 = a)$ . Der (Spalten-)Vektor  $\nu = (\nu_a)_{a \in S}$  heißt **Startverteilung** der Markov-Kette.  $\blacklozenge$

Soll die Startverteilung betont werden, so schreiben wir auch  $\mathbb{P}_\nu(X_n = b)$ ,  $b \in S, n \in \mathbb{N}$ , um anzuzeigen, dass die Verteilung von  $X_0$  durch  $\nu$  gegeben ist. Wenn klar ist, wo die Kette startet, kann der Index  $\nu$  hier auch weggelassen werden.

Kennt man zu einer Markov-Kette die Startverteilung sowie die Übergangsmatrix, so kann man die Verteilung der Kette zu einem festen Zeitpunkt  $n$  berechnen.

**Satz 11.3 (Eindimensionale Verteilungen)**

Sei  $X_0, X_1, \dots$  eine Markov-Kette mit Übergangsmatrix  $P$  und Startverteilung  $\nu$  auf einem Zustandsraum  $S$ . Dann gilt für alle  $b \in S, n \in \mathbb{N}$

$$\mathbb{P}_\nu(X_n = b) = \sum_{a \in S} \nu_a p_{a,b}^{(n)},$$

wobei  $P^n = (p_{a,b}^{(n)})_{a,b \in S}$  die  $n$ -fache Matrix-Potenz von  $P$  aus Satz 11.2 bezeichnet, und  $\nu^T$  der Zeilenvektor ist, den man durch Transposition von  $\nu$  erhält.

Als Vektor-Matrix Produkt aufgefasst gilt  $\sum_{a \in S} \nu_a p_{a,b}^{(n)} = (\nu^T P^n)_b$ , d.h.  $\mathbb{P}_\nu(X_n = b)$  ist der Eintrag an der Stelle  $b$  des Ergebnisses der Multiplikation des Zeilenvektors  $\nu^T$  mit der Matrix  $P^n$ . Für praktische Zwecke ist für große  $n$  und großen Zustandsraum  $S$  aber der Rechenaufwand zur Berechnung von  $\nu^T P^n$  oft zu hoch. Deshalb werden wir uns im nächsten Abschnitt mit strukturellen Eigenschaften von Markov-Ketten beschäftigen, welche es erlauben werden, das *Langzeitverhalten* einer Markov-Kette zu untersuchen.

**Beispiel 11.5**

Wir betrachten eine Markov-Kette mit folgender Übergangsmatrix:

$$P = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/2 & 1/4 & 1/4 \\ 1/2 & 1/2 & 0 \end{bmatrix}.$$

Man überprüft leicht, dass das tatsächlich eine stochastische Matrix ist, die zu einer Markov-Kette mit Zustandsraum  $\{1, 2, 3\}$  gehört. Wähle die Startverteilung  $\nu$  so, dass  $\nu^T = [1/2, 0, 1/2]$  gilt, d.h.  $\mathbb{P}(X_0 = 1) = \mathbb{P}(X_0 = 3) = 1/2, \mathbb{P}(X_0 = 2) = 0$ . Was ist dann die Verteilung von  $X_3$ ? Nach Satz 11.3 gilt

$$\mathbb{P}(X_3 = \cdot) = \nu^T P^3.$$

Wir berechnen die dritte Potenz von  $P$  (z.B. mit Matlab) und erhalten

$$P^3 = \begin{bmatrix} 1/2 & 3/16 & 5/16 \\ 1/2 & 13/64 & 19/64 \\ 1/2 & 7/32 & 9/32 \end{bmatrix}.$$

Damit erhalten wir

$$\mathbb{P}(X_3 = \cdot) = \nu^T P^3 = \begin{bmatrix} 1/2 & 0 & 1/2 \end{bmatrix} \cdot \begin{bmatrix} 1/2 & 3/16 & 5/16 \\ 1/2 & 13/64 & 19/64 \\ 1/2 & 7/32 & 9/32 \end{bmatrix} = \begin{bmatrix} 1/2 & 13/64 & 19/64 \end{bmatrix}.$$

Wir haben also  $\mathbb{P}_\nu(X_3 = 1) = 1/2, \mathbb{P}_\nu(X_3 = 2) = 13/64, \mathbb{P}_\nu(X_3 = 3) = 19/64$ . ■

## 11.2 Invariante Verteilungen

Eine Frage vom Anfang dieses Kapitels lautete: „Welche Zustände werden auf lange Sicht von einer Markov-Kette besonders oft besucht?“ Um diese und ähnliche Fragen zu beantworten, ist der folgende Begriff zentral.

### Definition 11.5 (Invariante Verteilung)

Sei  $X_0, X_1, \dots$  eine Markov-Kette mit Zustandsraum  $S$ . Ein (Spalten-)Vektor  $(\pi_a)_{a \in S}$  heißt **invariante Verteilung** für die Markov-Kette  $X = (X_n)_{n \geq 0}$ , falls gelten

$$\pi_a = \sum_{b \in S} \pi_b p_{b,a}, \text{ für alle } a \in S,$$

$$\pi_a \geq 0 \text{ für alle } a \in S, \text{ und } \sum_{a \in S} \pi_a = 1.$$

Invariante Verteilungen werden auch als **stationäre Verteilung** oder **Gleichgewichtsverteilung** bezeichnet. ♦

Wegen der Eigenschaft (b) ist eine invariante Verteilung eine diskrete Wahrscheinlichkeitsverteilung auf dem Zustandsraum  $S$ . Die Eigenschaft (a) kann in Vektor-Matrix Kurzschreibweise als  $\pi^T = \pi^T P$  geschrieben werden.

### Satz 11.4

Falls  $\pi$  eine stationäre Verteilung für  $(X_n)_{n \geq 0}$  ist, und  $\mathbb{P}(X_0 = a) = \pi_a$  für alle  $a \in S$  gilt, so gilt für alle  $n \in \mathbb{N}$ ,

$$\mathbb{P}_\pi(X_n = a) = \pi_a.$$

**Beweis 11.3** Dies folgt aus Satz 11.3 und der Eigenschaft (a) einer invarianten Verteilung.  $\square$

Dieser Satz sagt aus, dass falls die *Startverteilung* der Kette gerade eine invariante Verteilung ist, die Verteilung der Kette zu jedem festen Zeitpunkt gegeben ist durch ebendiese invariante Verteilung. Mit anderen Worten, die Verteilung bleibt *invariant*, also unverändert, im Laufe der Zeit. Dies erklärt die Bezeichnung invariante Verteilung.

**Satz 11.5 (Existenz und Berechnung der invarianten Verteilung)**

*Eine Markov-Kette auf einem endlichen Zustandsraum  $S$  besitzt immer mindestens eine invariante Verteilung. Invariante Verteilungen sind Lösungen des Gleichungssystems*

$$(P - I)^T \pi = 0, \quad \sum_{a \in S} \pi_a = 1,$$

wobei  $I$  die Einheitsmatrix bezeichnet, d.h.  $I$  ist die quadratische Matrix der Größe  $|S|$ , welche auf der Diagonalen Einsen, und ansonsten nur Nullen enthält.  $(P - I)^T$  bezeichnet die transponierte Matrix zu  $P - I$ , d.h. die Matrix die man erhält indem man Zeilen zu Spalten macht.

**Beweis 11.4** Gegeben die Übergangsmatrix  $P$  muss eine invariante Verteilung  $\pi = (\pi_a)_{a \in S}$  die Eigenschaften (a) und (b) aus Definition 11.5 erfüllen. Die Bedingung (a) können wir in Vektor-Matrix-Schreibweise umformen zu

$$\pi^T = \pi^T P,$$

wobei  $\pi^T$  den *Zeilenvektor* bezeichnet, den man durch Transposition des *Spaltenvektors*  $\pi$  erhält. Äquivalenzumformungen führen auf

$$(P - I)^T \pi = 0.$$

Dies ist ein lineares Gleichungssystem wie aus der linearen Algebra bekannt. Lösen des Gleichungssystems unter der zusätzlichen Bedingung (b) liefert also eine invariante Verteilung. Mit den Methoden der linearen Algebra kann man zeigen, dass der Lösungsraum von (a) für stochastische Matrizen immer mindestens eindimensional ist.  $\square$

Dieser Satz klärt die Frage der Existenz von invarianten Verteilungen zumindest im Fall von endlichen Zustandsräumen. Auf unendlichen Zustandsräumen ist die Frage schwieriger, und man kann beispielsweise zeigen, dass die symmetrische Irrfahrt auf  $\mathbb{Z}^d$  keine invariante Verteilung besitzt. Jedoch gibt es durchaus auch auf unendlichen Zustandsräumen Beispiele von invarianten Verteilungen. Wir diskutieren dies hier jedoch nicht.

Zur Frage der Eindeutigkeit von invarianten Verteilungen benötigen wir noch einen weiteren Begriff.

**Definition 11.6 (Irreduzibilität)**

Sei  $(X_n)_{n \in \mathbb{N}}$  eine Markov-Kette mit Zustandsraum  $S$ . Eine Folge von Zuständen  $(a_0, a_1, \dots, a_k)$  heißt **Pfad**, falls  $p_{a_i, a_{i+1}} > 0$  ist für alle  $i = 0, \dots, k-1$ . Das bedeutet, dass die Kette von  $a_0$  nach  $a_1$  und weiter nach  $a_2, a_3, \dots$  bis  $a_k$  tatsächlich springen kann. Eine Markov-Kette heißt **irreduzibel**, falls für jede Wahl von zwei Zuständen  $a, b \in S$  ein Pfad existiert, welcher  $a$  und  $b$  verbindet, d.h. ein Pfad mit  $a_0 = a$  und  $a_k = b$ . Die Länge  $k$  ist dabei egal.  $\blacklozenge$

**Beispiel 11.6**

Betrachte die Markov-Kette aus Beispiel 11.4. In Abbildung 11.5 ist leicht zu erkennen, dass es Pfade von jedem Zustand in jeden anderen Zustand gibt. Somit ist diese Markov-Kette irreduzibel. Im Gegensatz dazu ist die Markov-Kette mit stochastischer Matrix

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1/4 & 1/4 \\ 1/2 & 1/2 & 0 \end{bmatrix}$$

nicht irreduzibel, denn vom Zustand 1 aus kann keiner der anderen Zustände erreicht werden, es gibt keinen Pfad von 1 nach 2 oder von 1 nach 3.  $\blacksquare$

**Satz 11.6**

*Falls die Markov-Kette Fall von Satz 11.5 zusätzlich irreduzibel ist, so ist die invariante Verteilung eindeutig. Außerdem gilt dann  $\pi_a > 0$  für alle  $a \in S$ .*

**Beweis 11.5** Siehe z.B. Korollar 1.7 aus Levin et al. (2017).  $\square$

**Beispiel 11.7**

Betrachte wieder die Matrix

$$P = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/2 & 1/4 & 1/4 \\ 1/2 & 1/2 & 0 \end{bmatrix}.$$

Man kann sich, z.B. durch Aufzeichnen des Übergangsgraphen, davon überzeugen, dass die zugehörige Markov-Kette irreduzibel ist. Berechnen wir die invariante Verteilung. Dazu müssen wir die Gleichung  $(P - I)^T \pi = 0$  lösen, zuerst also  $(P - I)^T$  berechnen. Das ist

$$\left( \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/2 & 1/4 & 1/4 \\ 1/2 & 1/2 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right)^T = \begin{bmatrix} -1/2 & 0 & 1/2 \\ 1/2 & -3/4 & 1/4 \\ 1/2 & 1/2 & -1 \end{bmatrix}^T = \begin{bmatrix} -1/2 & 1/2 & 1/2 \\ 0 & -3/4 & 1/2 \\ 1/2 & 1/4 & -1 \end{bmatrix}.$$

Das Gleichungssystem  $(P - I)^T \pi = 0$  wird also zu

$$\begin{bmatrix} -1/2 & 1/2 & 1/2 \\ 0 & -3/4 & 1/2 \\ 1/2 & 1/4 & -1 \end{bmatrix} \cdot \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$



Zusätzlich muss die Gleichung  $\pi_1 + \pi_2 + \pi_3 = 1$  erfüllt sein, wir haben also ein Gleichungssystem mit drei Unbekannten und vier Gleichungen. Lösen mittels den bekannten Methoden der linearen Algebra (z.B. Gauß-Elimination) liefert für die invariante Verteilung

$$\pi = \begin{bmatrix} 1/2 \\ 1/5 \\ 3/10 \end{bmatrix}.$$

■

### Beispiel 11.8 (Markov-Kette mit mehreren invarianten Verteilungen)

Sei  $(X_n)_{n \in \mathbb{N}}$  definiert durch die stochastische Matrix

$$P = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/4 & 3/4 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix}.$$

Gesucht sind die invariante Verteilungen  $\pi = (\pi_i)_{i=1,\dots,4}$ . Der übliche Ansatz  $(P-I)^T \pi = 0$ ,  $\pi_1 + \dots + \pi_4 = 1$  führt (nach Elimination) auf das Gleichungssystem

$$\pi_2 = 2\pi_1$$

$$\pi_4 = \pi_3$$

$$\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1.$$

Dies sind drei linear unabhängige Gleichungen für vier Unbekannte, das Gleichungssystem hat unendlich viele Lösungen bei einem freien Parameter. Wir setzen  $\pi_1 = a$ , wobei  $a \in [0, 1]$  sein muss, da es sich um Wahrscheinlichkeiten handelt. Damit folgt  $\pi_2 = 2a$ , und  $\pi_3 = \pi_4 = \frac{(1-3a)}{2}$ , mit  $a \in [0, 1/3]$ . Somit ist die Menge  $L$  aller invarianten Verteilungen von  $P$  gegeben durch

$$L = \left\{ \pi^{(a)} = \left[ a, 2a, \frac{(1-3a)}{2}, \frac{(1-3a)}{2} \right]^T, a \in [0, 1/3] \right\}$$

Spezialfälle von Lösungen sind  $\pi^{(0)} = \left[ 0, 0, \frac{1}{2}, \frac{1}{2} \right]^T$  und  $\pi^{(1/3)} = \left[ \frac{1}{3}, \frac{2}{3}, 0, 0 \right]^T$ . Damit kann die Lösungsmenge auch geschrieben werden als

$$L = \{ \alpha \pi^{(1/3)} + (1 - \alpha) \pi^{(0)}, \alpha \in [0, 1] \}.$$

Eine Skizze der Situation zeigt, dass die Markov-Kette nicht irreduzibel ist, sondern der Zustandsraum in zwei disjunkte Komponenten zerfällt. Dabei ist  $\pi^{(0)}$  die invariante Verteilung der Kette, die auf der Komponente  $\{3, 4\}$  startet, und immer dort bleibt, und  $\pi^{(1/3)}$  ist die invariante Verteilung der Kette, die auf der Komponente  $\{1, 2\}$  startet. Die allgemeine Situation kann so interpretiert werden, dass mit Wahrscheinlichkeit  $\alpha$  auf der Komponente  $\{1, 2\}$  gestartet wird, und mit Wahrscheinlichkeit  $1 - \alpha$  auf  $\{3, 4\}$ . ■

## 11.3 Langzeitverhalten von Markov-Ketten

Die große Bedeutung invarianter Verteilungen liegt in folgendem zentralen Satz 11.7, für dessen Formulierung wir erst noch einen neuen Begriff einführen müssen.

### Definition 11.7

Eine irreduzible Markov-Kette heißt **aperiodisch**, falls für alle  $a \in S$  gilt

$$\text{ggT}\{k \in \mathbb{N} : \exists \text{Pfad der Länge } k \text{ von } a \text{ nach } a\} = 1.$$



### Beispiel 11.9

Die Markov-Kette aus Beispiel 11.4 ist aperiodisch, denn es gibt z.B. einen Pfad der Länge 2 von Zustand 1 nach Zustand 1 (springe nach 3 und wieder zurück), aber auch einen Pfad der Länge 3 (von 1 nach 2 nach 3 nach 1). Der ggT von 2 und 3 ist gleich 1 ist. Hingegen ist die symmetrische Irrfahrt auf  $\mathbb{Z}$  aus Beispiel 11.2 nicht aperiodisch, denn um zu einem Knoten zurückzukehren, muss die Kette sicher eine gerade Anzahl Sprünge machen. Somit ist der obige ggT gleich 2. ■

Wir kommen nun zum Hauptergebnis dieses Kapitels.

### Satz 11.7 (Konvergenz gegen die invariante Verteilung)

Sei  $(X_n)_{n \geq 0}$  eine irreduzible und aperiodische Markov-Kette mit Zustandsraum  $S$ . Sei  $(\pi_a)_{a \in S}$  eine invariante Verteilung. Dann gilt für alle  $a \in S$ , unabhängig von der Startverteilung,

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n = a) = \pi_a.$$

**Beweis 11.6** Siehe z.B. Georgii (2009), Cor. 6.17. □

In Worten: Ganz egal, wo die Kette startet, für  $n \rightarrow \infty$  wird ihre Verteilung durch die invariante Verteilung  $\pi$  beschrieben. In praktischen Anwendungen wird man nicht unbedingt an Grenzwerten  $n \rightarrow \infty$  interessiert sein, sondern an festen, aber großen Werten  $n \in \mathbb{N}$ . Dann besagt der Satz, dass stets

$$\mathbb{P}(X_n = a) \approx \pi_a$$

gilt, unabhängig von der Startverteilung der Kette. Auf lange Frist gibt also die invariante Verteilung  $\pi_a$  zumindest näherungsweise die Wahrscheinlichkeit an, mit der sich die Kette im Zustand  $a$  aufhält, unabhängig von ihrem Startwert.

Dieses liefert wichtige Informationen über Struktur der Markov-Kette: Zustände mit großem Wert für  $\pi$  werden auf lange Sicht häufiger besucht als Zustände mit kleinem Wert.

Falls keine invariante Verteilung existiert, kann unter den Voraussetzungen des Satzes gezeigt werden, dass  $\lim_{n \rightarrow \infty} \mathbb{P}(X_n = a) = 0$  für alle  $a \in S$ .

**Beispiel 11.10 (Netzwerk)**

Beim Prozess auf dem Netzwerk aus Abbildung 11.1 bedeutet Satz 11.7 konkret: Knoten mit vielen Kanten werden oft besucht, denn für solche Knoten  $a$  ist  $\pi_a$  größer als für solche die selten besucht werden. Man kann die invariante Verteilung berechnen, indem wir das Gleichungssystem

$$(P - I)^T \pi = 0, \quad \pi_1 + \dots + \pi_{18} = 1$$

für die aus dem Graphen abgelesene Übergangsmatrix löst (z.B. mit Matlab). Die Lösung ist ein Vektor mit 18 Einträgen, und man findet (unter anderem)  $\pi_{10} = \pi_{18} = 0.1034$ , also relativ große Werte, hingegen ist  $\pi_{13} = 0.0172$  relativ klein. Dies liegt daran, dass Knoten 10 und 18 mit vielen anderen Knoten verbunden sind, und deshalb öfter besucht werden als Knoten 13, welcher nur einen Nachbarn hat. ■

Man kann in diesem Zusammenhang noch mehr aussagen. Häufige Besuche eines Zustandes bedeuten insbesondere, dass kurze Wartezeiten zwischen zwei Besuchen gibt.

**Satz 11.8**

Sei  $(X_n)_{n \in \mathbb{N}_0}$  eine irreduzible und aperiodische homogene Markov-Kette mit Zustandsraum  $S$  und invarianter Verteilung  $\pi$ . Sei

$$T_a := \inf\{n \geq 1 : X_n = a\}$$

der erste Zeitpunkt nach dem Start, zu dem sich die Kette im Zustand  $a$  befindet. Dann gilt für die invariante Verteilung  $\pi$ ,

$$\pi_a = \frac{1}{\mathbb{E}[T_a \mid X_0 = a]}.$$

Dabei bezeichnet  $\mathbb{E}[\cdot \mid X_0 = a]$  den Erwartungswert unter der bedingten Verteilung  $\mathbb{P}(\cdot \mid X_0 = a)$ .

**Beweis 11.7** Siehe z.B. Kapitel 7 von Mitzenmacher Upfal (2017). □

Es gilt also, je größer  $\pi_a$  ist desto kürzer ist die erwartete Zeit, bis die Kette bei Start in  $a$  zu  $a$  zurückkehrt. In diesem Zusammenhang spielen oft folgende Begriffe eine Rolle:

**Definition 11.8 (Rekurrenz, Transienz)**

Sei  $(X_n)_{n \in \mathbb{N}_0}$  eine irreduzible Markov-Kette mit Zustandsraum  $S$ . Die Kette heißt *rekurrent*, falls es ein  $a \in S$  gibt, so dass

$$\mathbb{P}(\exists n \in \mathbb{N} : X_n = a \mid X_0 = a) = 1$$

ist. Andernfalls heißt die Kette *transient*. Eine rekurrente Kette heißt *positiv rekurrent*, falls  $\mathbb{E}[T_a \mid X_0 = a] < \infty$  ist für alle  $a \in S$ , wobei wieder  $T_a := \inf\{n \geq 1 : X_n = a\}$  die erste Rückkehrzeit bezeichnet. Ist eine Markov-Kette rekurrent, aber nicht positiv rekurrent, so heißt sie *nullrekurrent*. ♦

Die Eigenschaft

$$\mathbb{P}(\exists n \in \mathbb{N} : X_n = a \mid X_0 = a) = 1$$

bedeutet, dass die Kette, wenn sie im Zustand  $a$  startet, mit Wahrscheinlichkeit 1, also sicher, irgendwann wieder in diesen Zustand zurückkehrt. Satz 11.8 besagt insbesondere, dass eine irreduzible, aperiodische Markov-Kette mit invarianter Verteilung stets positiv rekurrent und damit auch rekurrent ist.

Man kann beweisen, dass für irreduzible Markov-Ketten die Eigenschaft dass es ein  $a \in S$  gibt mit  $\mathbb{P}(\exists n \in \mathbb{N} : X_n = a \mid X_0 = a) = 1$  bereits impliziert, dass *jeder* Zustand  $a \in S$  diese Eigenschaft hat, und dass es auch bedeutet, dass man, egal wo man startet, jeden Zustand mit Wahrscheinlichkeit 1 mindestens einmal besucht (sogar, dass man dann jeden Zustand unendlich oft besucht).

Ist die Kette *nicht irreduzibel*, so kann es Zustände geben, zu denen man mit Wahrscheinlichkeit 1 wieder zurückkehrt, und solche, zu denen man mit positiver Wahrscheinlichkeit nie wieder zurückkehrt. Das bedeutet, die Eigenschaften rekurrent, transient und positiv rekurrent sind in diesem Fall nicht Eigenschaften der ganzen Kette, sondern nur einzelner Zustände.

## 11.4 Anwendungen

### Beispiel 11.11 (PageRank)

Das Ziel des PageRank Algorithmus von Google besteht in der Bewertung oder Gewichtung von Dokumenten mit Verknüpfungen, wie beispielsweise Webseiten mit Links. Das verwendete *Zufallssurfermodell* ist deshalb angelehnt an die Irrfahrt auf einem Graphen, wie im Eingangsbeispiel dieses Kapitels, siehe Beispiel 11.10. Das Netzwerk besteht dabei aus  $k$  Webseiten,  $S = \{1, \dots, k\}$  sowie aus Verbindungen mittels Links, wobei  $c_i$  = die Anzahl Links bezeichnet, welche von Seite  $i \in S$  *ausgehen*. Im Gegensatz zu Beispiel 11.10 gibt es natürlich für einen Link von Seite  $i$  auf Seite  $j$  nicht unbedingt einen entsprechenden Link von  $j$  auf  $i$ , im Gegensatz zu vorhin arbeiten wir also mit einem *gerichteten Graphen*, d.h. die Kanten können jeweils nur in eine Richtung benutzt werden (wie bei einem Übergangsgraphen). Setze

$$q_{ij} := \begin{cases} \frac{1}{c_i} & \text{falls } i \text{ auf } j \text{ verlinkt} \\ 0 & \text{sonst.} \end{cases}$$

Dann ist  $Q := (q_{ij})_{i,j \in S}$  ist die Übergangsmatrix für die Irrfahrt auf dem (gerichteten) Graphen, der durch die Webseiten und die Links beschrieben wird, d.h.  $q_{ij}$  ist die Wahrscheinlichkeit mit der man von  $i$  nach  $j$  springt. Deshalb auch der Begriff Zufallssurfermodell: Die Irrfahrt auf diesem Graphen entspricht genau dem Verhalten eines "Surfers" im Netz, der auf jeder Seite zufällig einen Link auswählt und diesen anklickt. Die Grundidee von PageRank ist nun einfach: Finde die invariante Verteilung von  $Q$ . Nach Satz 11.7 wissen wir: Je größer deren  $i$ -te Komponente, desto häufiger wird die

Seite  $i$  besucht, in anderen Worten, desto wichtiger ist die Seite  $i$ . Eine kurze Rechnung zeigt, dass die invariante Verteilung gegeben ist durch die Lösung des Gleichungssystems

$$\pi_j = \sum_{i: i \text{ verlinkt auf } j} \frac{\pi_i}{c_i}, \quad j \in S.$$

Die praktische Berechnung in großen Netzwerken führt natürlich auf diverse numerische Probleme, auf die wir hier nicht eingehen können. Wir wollen jedoch kurz eine konzeptuelle Schwierigkeit ansprechen. Gibt es nämlich Seiten, auf die gar nicht verlinkt wird, so können diese mit dem Zufallssurfer-Algorithmus nicht gefunden werden, und erhalten so Gewicht 0. Eine Möglichkeit zur Abhilfe besteht darin, dass zusätzlich eine Zahl  $d \in (0, 1)$  gewählt und

$$p_{ij} := dq_{ij} + \frac{1-d}{k}$$

definiert wird, d.h. der Zufallssurfer folgt mit Wahrscheinlichkeit  $d$  den Links, mit Wahrscheinlichkeit  $1-d$  springt er zu einer zufällig ausgewählten Seite. Der PageRank der Seite  $i \in S$  ist nun definiert BrinPage (1998) als

$$PR_i := \tilde{\pi}_i,$$

wobei  $\tilde{\pi}$  die invariante Verteilung von  $P = (p_{ij})_{i,j=1,\dots,k}$  ist.  $PR$  ist also die (normierte) Lösung des Gleichungssystems

$$PR_i = \frac{1-d}{k} + d \sum_{j=1}^k \frac{1_{\{j \text{ verlinkt auf } i\}}}{c_j} PR_j.$$

■

### Beispiel 11.12 (Stabilität von Warteschlangen)

Warteschlangenmodelle spielen beispielsweise in der IT, in der Verkehrsmodellierung, in Callcentern oder in der Logistik eine Rolle. Sie beschreiben Systeme, in denen Kunden zu zufälligen Zeiten eintreffen, und deren Anliegen dann abgearbeitet werden muss. Wir modellieren hier ein einfaches Warteschlangenmodell mittels einer Markov-Kette, auf folgende Weise: Betrachte Zeitabschnitte von fester Länge. Gegeben seien Parameter  $k \in \mathbb{N}$ ,  $\mu \in [0, 1]$ , und  $\lambda \in [0, 1]$ , mit  $\lambda + \mu \leq 1$ . Modellannahme: In jedem Zeitabschnitt geschieht *eines* der folgenden Ereignisse:

- Falls weniger als  $k$  Kunden da sind, so kommt mit Wahrscheinlichkeit  $\lambda$  ein neuer Kunde an
- Falls mindestens ein Kunde da ist, so verlässt mit Wahrscheinlichkeit  $\mu$  ein Kunde die Schlange
- Andernfalls (mit Wahrscheinlichkeit  $1 - \mu - \lambda$ ) geschieht nichts

Sei  $X_n := \text{Anzahl Kunden in der Schlange im } n\text{-ten Zeitabschnitt}$ . Dann ist  $(X_n)_{n \in \mathbb{N}_0}$  eine Markov-Kette mit Zustandsraum  $S = \{0, \dots, k\}$ . Diese Markov-Kette kann folgende Übergänge machen:

$$p_{i,i+1} = \lambda \text{ falls } i < k$$

$$p_{i,i-1} = \mu \text{ falls } i > 0.$$

d.h. die zugehörige Übergangsmatrix ist gegeben durch

$$P = \begin{bmatrix} 1-\lambda & \lambda & 0 & \cdots & \cdots & 0 \\ \mu & 1-\mu-\lambda & \lambda & 0 & \cdots & 0 \\ 0 & \mu & 1-\mu-\lambda & \lambda & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & & & 0 & \mu & 1-\mu-\lambda & \lambda \\ 0 & \cdots & & & 0 & \mu & 1-\mu & \end{bmatrix}.$$

Eine zentrale Frage lautet nun: Wie viele Kunden sind auf lange Sicht in der Schlange? Ist das Warteschlangenmodell auf lange Sicht, also für  $k \rightarrow \infty$  stabil? Stabilität wird hierbei dahingehend interpretiert, dass die Anzahl Kunden in der Schlange nicht über alle Maßen wächst. Nach Satz 11.7 ist das Langzeitverhalten einer Markov-Kette durch ihre invariante Verteilung bestimmt wird. Mittels des üblichen Ansatzes und Lösung des entsprechenden Gleichungssystems findet man

### Satz 11.9

[Invariante Verteilung der Warteschlange] Die invariante Verteilung der beschriebenen Warteschlange ist gegeben durch

$$\pi_i = \frac{1}{\sum_{j=0}^k (\lambda/\mu)^j} \cdot \left(\frac{\lambda}{\mu}\right)^i, \quad i = 0, \dots, k.$$

Was bedeutet das für die Stabilität? Falls  $\lambda/\mu < 1$  ist, gilt  $\sum_{j=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^j = \frac{\mu}{\mu-\lambda} < \infty$ , die Warteschlange hat für  $k \rightarrow \infty$  eine invariante Verteilung, ist stabil. Anschaulich bedeutet das einfach, dass im Mittel pro Zeiteinheit mehr Kunden das System verlassen als hereinkommen. Andernfalls ist  $\sum_{j=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^j = \infty$ , die Warteschlange ist instabil, auf lange Sicht kommen also immer mehr Kunden rein und können nicht schnell genug abgearbeitet werden, es entsteht also Stau. ■

### Zusammenfassung

Markov-Ketten beschreiben zufällig ablaufende Prozesse, deren charakteristische Eigenschaft darin besteht, dass der Zustand im nächsten Schritt nur vom aktuellen Zustand abhängt, und nicht von der Vergangenheit. Die Übergangsmatrix enthält die wesentlichen Informationen über die zugehörige Markov-Ketten. Das Langzeitverhalten von Markov-Ketten wird durch deren invarianten Verteilungen bestimmt, welche durch Lösen eines Gleichungssystems berechnet werden können. Diese Tatsache spielt in verschiedenen praktischen Anwendungen eine Rolle, beispielsweise beim PageRank Algorithmus.

## Aufgaben

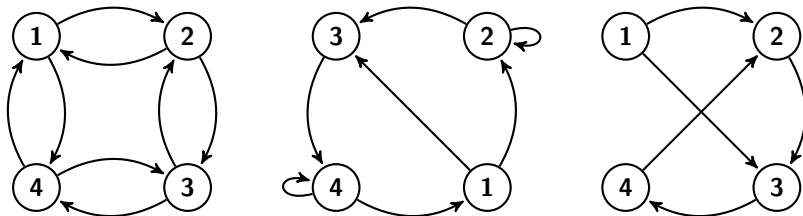
**11.1** Überprüfen Sie, dass

$$P = \begin{bmatrix} 0.2 & 0 & 0.7 & 0.1 \\ 0.4 & 0 & 0 & 0.6 \\ 0.3 & 0.2 & 0.2 & 0.3 \\ 0 & 0.8 & 0.1 & 0.1 \end{bmatrix}$$

eine stochastische Matrix ist, und zeichnen Sie den Übergangsgraphen.

**11.2** In Abbildung 11.5 ist der Übergangsgraph einer Markov-Kette gegeben. Bestimmen Sie die zugehörige stochastische Matrix  $P$ , und berechnen Sie die bedingte Verteilung von  $X_3$  gegeben  $X_0 = 1$ . Ist die zugehörige Kette irreduzibel bzw. aperiodisch?

**11.3** Welche der Markovketten zu den folgenden Übergangsgraphen sind irreduzibel bzw. aperiodisch? Pfeile stehen dabei für positive Übergangswahrscheinlichkeiten.



**11.4** Berechnen Sie die invarianten Verteilungen der Markov-Ketten aus Aufgabe 11.1 und 11.2.

**11.5** Bestimmen Sie für die Markov-Kette aus Aufgabe 11.2  $\lim_{n \rightarrow \infty} \mathbb{E}[X_n]$ .

**11.6** Es sei  $(X_n)_{n \in \mathbb{N}_0}$  eine Markov-Kette mit Zustandsraum  $\{1, 2, 3\}$  und Übergangswahrscheinlichkeiten  $p_{1,2} = p_{1,3} = 1/2, p_{2,1} = 1/2, p_{2,2} = 1/4, p_{2,3} = 1/4, p_{33} = 1$ . a) Ist die Markov-Kette irreduzibel? b) Bestimmen Sie alle möglicherweise vorhandenen invarianten Verteilungen dieser Kette. c) Bestimme  $\lim_{n \rightarrow \infty} \mathbb{P}(X_n = i | X_0 = 1)$  für  $i \in \{1, 2, 3\}$ .

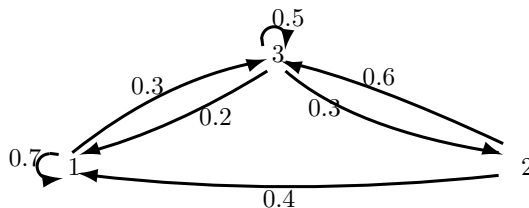


Abb. 11.5: Übergangsgraph zu Aufgabe 7.3

**11.7** Es sei  $(X_n)_{n \in \mathbb{N}_0}$  eine Markov-Kette mit Zustandsraum  $\{1, 2, 3\}$  und Übergangswahrscheinlichkeiten  $p_{1,2} = p_{1,3} = 1/2, p_{2,2} = p_{33} = 1$ . a) Ist die Markov-Kette irreduzibel? b) Bestimmen Sie alle möglicherweise vorhandenen invarianten Verteilungen dieser Kette. c) Bestimmen Sie  $\lim_{n \rightarrow \infty} \mathbb{P}_\nu(X_n = i)$  für  $i \in \{1, 2, 3\}$ , wenn die Startverteilung gegeben ist durch  $\nu = [1/2, 1/2, 0]^T$ .

**11.8** Man beweise Satz 11.9.

**11.9** In Beispiel 11.12 betrachte man im Fall  $\lambda < \mu$  die Situation für  $k \rightarrow \infty$ . Wie viele Kunden befinden sich auf lange Sicht im Mittel in der Warteschlange?



# 12 Randomisierte Algorithmen: Beispiele und Anwendungen

---

## Übersicht

12.1 Erwartete Laufzeit von Quicksort .....	172
12.2 Erfolgswahrscheinlichkeit des MinCut .....	174
12.3 Ein probabilistischer Existenzbeweis: Graphenfärbung .....	177

---

Ein randomisierter (auch: zufälliger oder stochastischer) Algorithmus verwendet zufällige Ereignisse bei seiner Durchführung. Der Zufall kann dabei in der zufälligen Auswahl, Ordnung oder Partitionierung der Eingabe stecken, in zufälligen Entscheidungen bei der Ausführung des Algorithmus, oder auch in einer zufälligen Verteilung der Lasten auf die verfügbaren Rechenkapazitäten. In vielen Situationen haben zufällige Algorithmen Vorteile gegenüber deterministischen Algorithmen, beispielsweise kann es möglich sein, einen deterministischen Algorithmus mit schlechtem worst-case Verhalten so zu randomisieren, dass der resultierende zufällige Algorithmus sich mit hoher Wahrscheinlichkeit günstig verhält. Stochastische Algorithmen sind oft einfach zu verstehen und zu implementieren, haben häufig eine kurze Laufzeit und wenig Speicherbedarf.

In diesem Abschnitt wird keine allgemeine Theorie entwickelt, sondern einige randomisierte Algorithmen und häufig verwendete Methoden an Hand relativ einfacher Beispiele illustriert. Solche Beispiele können in der Praxis als Teilprobleme eines größeren Algorithmus auftreten. Unabdingbar zur Programmierung eines stochastischen Algorithmus ist die Fähigkeit, Zufallsvariablen mit einer vorgegebenen Verteilung zu simulieren. Zu dieser Problemstellung verweisen wir auf Kapitel 15.

### Lernziele

- Grundprinzipien und Grundbegriffe randomisierter Algorithmen kennen
- Beispiele von randomisierten Algorithmen kennen
- Typische Fragestellungen und Ansätze zur Analyse von randomisierten Algorithmen kennen.

### Vorkenntnisse

Benötigt werden die Grundlagen aus dem ersten Teil, insbesondere Zufallsvariablen, Erwartungswerte und Verteilungen.

## 12.1 Erwartete Laufzeit von Quicksort

Der randomisierte Quicksort-Algorithmus ist ein elementarer Sortieralgorithmus. Er sortiert in einem rekursiven Verfahren die Eingabe, welche aus einer Liste von Zahlen besteht. Der Zufall steckt dabei in einer stochastischen Partitionierung der Eingabe, bzw. der zufälligen Wahl eines Vergleichselements bei jedem Schritt.

### Randomisierter Quicksort

Eingabe: Eine Menge  $S = \{x_1, \dots, x_n\}$  von paarweise verschiedenen Zahlen.

1. Wähle *zufällig* (d.h. gleichverteilt) ein Element  $Y$  von  $S$
2. Bestimme die Mengen

$$S_{<} := \{x \in S : x < Y\} \quad \text{und} \quad S_{>} := \{x \in S : x > Y\}$$

bestehend aus den Elementen von  $S$  welche kleiner bzw. größer sind als  $Y$ , indem jedes Element von  $S$  mit  $Y$  verglichen wird.

3. Ordne rekursiv die Mengen  $S_{<}$  und  $S_{>}$  durch Wiederholen von 1. und 2.
4. Gebe die sortierte Menge  $S_{<}$  gefolgt von  $Y$  gefolgt von der sortierten Menge  $S_{>}$  aus

Ausgabe: Die Elemente von  $S$  aufsteigend sortiert.

Die Mengen  $S_{<}$  und  $S_{>}$  haben in jedem Fall eine kleinere Kardinalität als die ursprüngliche Menge  $S$ . Eine ein- oder nullelementige Menge ist offensichtlich geordnet. Somit terminiert dieser Algorithmus nach endlich vielen Schritten.

### Beispiel 12.1

Ist die Liste  $S = \{17, 6, 56, 23, 4, 19, 58, 28, 14, 43, 62, 3\}$  zu sortieren, und wird als Vergleichselement  $Y = 23$  ausgewählt, so erhält man  $S_{<} = \{17, 6, 4, 19, 14, 3\}$  und  $S_{>} = \{56, 58, 28, 43, 62\}$ . ■

**Definition 12.1 (Monte Carlo und Las Vegas)**

Ein Algorithmus der (wenn er terminiert) immer eine richtige Lösung des Problems ausgibt, heißt **Las-Vegas-Algorithmus**. Ein Algorithmus der mit positiver Wahrscheinlichkeit ein falsches Ergebnis ausgibt, heißt **Monte-Carlo-Algorithmus**. ♦

Der zufällige Quicksort ist ein Beispiel für einen Las-Vegas-Algorithmus. Im nächsten Abschnitt werden wir ein Beispiel für einen Monte-Carlo-Algorithmus sehen.

**Definition 12.2 (Laufzeit bei Quicksort)**

Als Maß für die Laufzeit des randomisierten Quicksort definieren wir die Zufallsvariable  $X_S$  als die Anzahl durchgeführter Paarvergleiche bei Eingabe einer Liste  $S$ . ♦

Es gilt  $X_S = X_n$ , d.h. die Laufzeit hängt nur von der Länge  $n$  der Liste ab. Bei einer unglücklichen Wahl der Vergleichselemente ist die Laufzeit des Algorithmus

$$\frac{(n-1)n}{2} \approx \frac{n^2}{2}.$$

Dies geschieht im worst-case Szenario, bei dem man in jedem Schritt das aktuell kleinste (bzw. größte Element) der Liste auswählt. In diesem Fall ist jeweils  $S_<$  die leere Menge, und  $S_>$  nur um ein Element, nämlich um  $Y$ , kleiner als die ursprüngliche Liste. Somit nimmt die Anzahl benötigter Vergleiche pro Schritt nur um einen ab. Durch die Randomisierung ist jedoch in Erwartung die Laufzeit sehr viel kürzer. Dies liegt daran, dass es nur zwei worst cases, aber viele recht gute Möglichkeiten gibt.

**Satz 12.1 (Erwartete Laufzeit von Quicksort)**

*Im randomisierten Quicksort gilt*

$$\mathbb{E}[X_n] \leq 2nH_n,$$

wobei  $H_n := \sum_{k=1}^n \frac{1}{k}$  von der Größenordnung  $\ln n$  ist. Die Laufzeit ist also bei einer Liste der Länge  $n$  von der Ordnung  $n \ln n$ .

**Beweis 12.1** Wir schreiben  $x_{(1)}, \dots, x_{(n)}$  für die Elemente von  $S$  in der *richtigen* Reihenfolge, d.h.  $x_{(1)}$  ist das kleinste Element von  $S$ . Betrachte

$$A_{i,j} := \{\text{es kommt zu einem direkten Vergleich von } x_{(i)} \text{ und } x_{(j)}\}.$$

Dies ist ein Ereignis im Sinne von Kapitel 1, welches im Verlauf des Algorithmus auftreten kann. Nicht jedes dieser Ereignisse tritt tatsächlich auf, da unter Umständen  $x_{(i)}$  und  $x_{(j)}$  in verschiedenen Teillisten landen, bevor sie direkt verglichen werden.

Wir nehmen an, es gelte  $i < j$ . Damit tritt  $A_{i,j}$  genau dann ein, wenn entweder  $x_{(i)}$  oder  $x_{(j)}$  als erstes aus den insgesamt  $j - i + 1$  Elementen  $x_{(i)}, x_{(i+1)}, \dots, x_{(j)}$  als Vergleichselement  $Y$  ausgewählt wird. Da jedes Element mit gleicher Wahrscheinlichkeit Vergleichselement wird, ist die Wahrscheinlichkeit dafür nach den Rechenregeln in Laplace-Räumen

$$\mathbb{P}(A_{i,j}) = \frac{2}{j - i + 1}.$$

Mit Hilfe der Indikatorfunktion dieser Ereignisse können wir die Anzahl Vergleiche schreiben als

$$X_n = \sum_{(i,j): 1 \leq i < j \leq n} 1_{A_{i,j}} = \sum_{i=1}^{n-1} \sum_{j=i+1}^n 1_{A_{i,j}}.$$

Damit folgt mit den Rechenregeln für den Erwartungswert

$$\begin{aligned} \mathbb{E}[X_n] &= \mathbb{E} \left[ \sum_{i=1}^{n-1} \sum_{j=i+1}^n 1_{A_{i,j}} \right] \\ &= \sum_{i=1}^{n-1} \sum_{j=i+1}^n \mathbb{E}[1_{A_{i,j}}] = \sum_{i=1}^{n-1} \sum_{j=i+1}^n [1 \cdot \mathbb{P}(A_{i,j}) + 0 \cdot \mathbb{P}(A_{i,j}^c)] \\ &= \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{2}{j-i+1} \leq 2 \sum_{i=1}^n \sum_{j=1}^n \frac{1}{k} = 2nH_n. \end{aligned}$$

□

## 12.2 Erfolgswahrscheinlichkeit des MinCut

In diesem Abschnitt lernen wir einen Algorithmus kennen, welcher durch Entfernen von Kanten einen zusammenhängende *Graphen* in einen unzusammenhängenden Teilgraphen überführt. Das Ziel dabei ist, möglichst wenige Kanten zu entfernen. Eine solche Problematik spielt beispielsweise bei der Stabilitätsanalyse von Netzwerken eine Rolle, oder auch in der sogenannten Clusteranalyse.

### Definition 12.3 (Graph)

Ein **Graph**  $G$  besteht aus Knoten und Kanten, wobei Mehrfachkanten zulässig sind.  $G$  heißt **zusammenhängend**, falls je zwei Knoten durch eine Folge von Kanten verbunden sind. Ein **einfacher Graph** ist ein Graph ohne Mehrfachkanten. Einen Graphen, bei dem Mehrfachkanten explizit zugelassen sind, nennt man auch **Multigraph**, siehe Abbildung 12.1

◆

Notation: Wir schreiben  $G = (V, E)$ , wobei  $V$  die Menge der Knoten und  $E$  die Menge der Kanten des Graphen bezeichnet. Kanten sind von der Form  $e = \{v_1, v_2\}$  mit  $v_1, v_2 \in V$ . Bei einem Multigraph schreiben wir  $G = (V, E, m)$ , wobei  $m : E \rightarrow \mathbb{N}_0$  für jede Kante angibt, wie oft sie auftritt.

### Definition 12.4 (Cut und MinCut)

Sei  $G$  ein zusammenhängender Multigraph. Ein **Cut** ist eine Teilmenge der Kanten von  $G$ , deren Entfernung dazu führt dass  $G$  unzusammenhängend wird. Ein **MinCut** ist ein Cut von minimaler Kardinalität.

◆

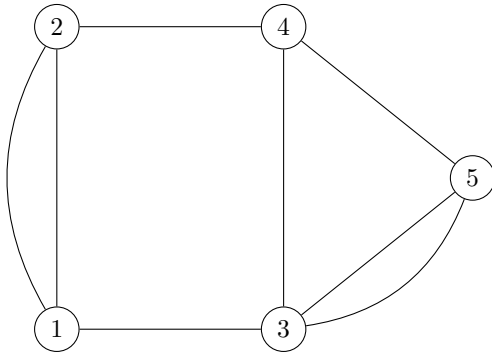


Abb. 12.1: Ein Multigraph mit 5 Knoten

**Beispiel 12.2**

Entfernt man beispielsweise im Multigraphen aus Abbildung 12.1 die drei Kante zwischen Knoten 1 und Knoten 3 sowie zwischen Knoten 2 und Knoten 4, so erhält man einen MinCut: Der Graph zerfällt in die Komponenten bestehend aus Knoten 1 und 2, sowie aus den Knoten 3, 4, und 5. Entfernt man die Kante zwischen Knoten 4 und Knoten 5 sowie die beiden Kanten zwischen Knoten 3 und Knoten 5, so erhält man einen Cut, da Knoten 5 vom Rest des Graphen getrennt wird, aber keinen MinCut, da drei Kanten entfernt werden müssen, obwohl es auch mindestens einen Cut der Größe zwei gibt. ■

Wir wollen nun einen einfachen stochastischen Algorithmus für das *Min-Cut Problem* angeben, d.h. der Algorithmus sollte einen MinCut eines vorgegebenen Graphen finden.

**Ein randomisierter MinCut Algorithmus**

Eingabe: Knotenmenge  $V$  und Kantenmenge  $E$ .

1. Wähle *zufällig* eine Kante aus (gleichverteilt aus allen vorhandenen)
2. Verschmelze die Endpunkte der gewählten Kante zu einem neuen Knoten, und entferne alle Kanten zwischen diesen beiden Knoten
3. Wiederhole 1. und 2. bis nur noch zwei Knoten da sind.

Ausgabe: Die zwei verbliebenen Knoten und eine Anzahl Kanten zwischen diesen Knoten.

Wird dieser Algorithmus auf einen Graphen mit  $|V| = n$  Knoten angewandt, so terminiert der Algorithmus offensichtlich nach  $n - 2$  Schritten, da der Graph in jedem Schritt um genau einen Knoten verringert wird. Man kann sich leicht überlegen, dass die ausgegebenen Kanten einen Cut des ursprünglichen Graphen bilden, jedoch ist es auch einfach, Beispiele dafür anzugeben, dass die Ausgabe nicht immer ein MinCut ist. Somit handelt es sich hier um einen Las Vegas-Algorithmus, und von Interesse ist insbesondere die Frage, mit welcher Wahrscheinlichkeit die Ausgabe tatsächlich ein MinCut ist.

**Satz 12.2 (Erfolgswahrscheinlichkeit beim MinCut)**

Sei  $M$  das Ereignis dass die Ausgabe im vorgestellten Algorithmus ein MinCut ist. Wenn  $|V| = n$  ist, so gilt

$$\mathbb{P}(M) \geq \frac{2}{n(n-1)}.$$

**Beweis 12.2** Sei  $C$  ein MinCut des vorgegebenen Graphen, also eine Teilmenge der Kantenmenge  $E$ . Falls der Algorithmus genau  $C$  ausgibt, ist also ein MinCut gefunden. Da es mehr als einen MinCut geben kann, ist die gesuchte Wahrscheinlichkeit also mindestens so groß wie die Wahrscheinlichkeit, dass der Algorithmus  $C$  ausgibt.

Wir stellen eine Vorüberlegung zur Kardinalität von  $E$  und  $C$  an. Falls  $|C| = k$ , der MinCut  $C$  also  $k$  Kanten hat, so hat jeder Knoten von  $G$  mindestens  $k$  Kanten, denn andernfalls könnte man einen kleineren Cut als  $C$  finden, indem man alle Kanten dieses Knoten entfernt. Somit ist also

$$|E| \geq \frac{kn}{2},$$

denn nach Voraussetzung gibt es im Graphen mindestens  $n$  Knoten, von denen jeder mindestens zwei Kanten hat, wobei jede Kante zu zwei Knoten gehört.

Nach dieser Vorüberlegung definiere

$$E_i := \{\text{im } i\text{-ten Schritt wird kein Element von } C \text{ ausgewählt}\}.$$

Dies ist ein Ereignis im zugrunde liegenden Wahrscheinlichkeitsraum, auf dem die gleichverteilten Zufallsvariablen definiert sind. Es gilt, da der Algorithmus genau  $n-2$  Schritte macht,

$$\mathbb{P}(\text{Algorithmus gibt } C \text{ aus}) = \mathbb{P}(E_1 \cap E_2 \cap \dots \cap E_{n-2}).$$

Aus der Definition der bedingten Wahrscheinlichkeit (wiederholt angewendet) erhalten wir

$$\begin{aligned} \mathbb{P}(E_1 \cap E_2 \cap \dots \cap E_{n-2}) &= \mathbb{P}(E_{n-2} \mid E_1 \cap E_2 \cap \dots \cap E_{n-3}) \cdot \mathbb{P}(E_1 \cap E_2 \cap \dots \cap E_{n-3}) \\ &= \mathbb{P}(E_{n-2} \mid \cap_{i=1}^{n-3} E_i) \cdot \mathbb{P}(\cap_{i=1}^{n-3} E_i) \\ &= \dots = \mathbb{P}(E_{n-2} \mid \cap_{i=1}^{n-3} E_i) \cdot \mathbb{P}(E_{n-3} \mid \cap_{i=1}^{n-4} E_i) \cdot \dots \cdot \mathbb{P}(E_2 \mid E_1) \cdot \mathbb{P}(E_1) \end{aligned}$$

Aus der obigen Vorüberlegung sehen wir, dass

$$\mathbb{P}(E_1) \geq 1 - \frac{k}{k \cdot n/2} = 1 - \frac{2}{n} = \frac{n-2}{n}$$

ist, sowie

$$\mathbb{P}(E_2 \mid E_1) \geq 1 - \frac{k}{k \cdot (n-2)/2} = 1 - \frac{2}{n-1} = \frac{n-3}{n-1},$$

und induktiv

$$\mathbb{P}(E_j \mid \cap_{i=1}^{j-1} E_i) \geq 1 - \frac{2}{n-j+1} = \frac{n-j-1}{n-j+1}$$

gelten. Damit folgt zusammengefasst

$$\begin{aligned}
 \mathbb{P}(\text{Algorithmus gibt } C \text{ aus}) &= \mathbb{P}(E_1 \cap E_2 \cap \dots \cap E_{n-2}) \\
 &= \mathbb{P}(E_{n-2} \mid \cap_{i=1}^{n-3} E_i) \cdot \mathbb{P}(E_{n-3} \mid \cap_{i=1}^{n-4} E_i) \cdot \dots \cdot \mathbb{P}(E_2 \mid E_1) \cdot \mathbb{P}(E_1) \\
 &= \prod_{j=1}^{n-2} \frac{n-j-1}{n-j+1} = \frac{2}{n(n-1)}.
 \end{aligned}$$

□

Führt man den MinCut-Algorithmus mehrmals hintereinander mit *unabhängiger* Wahl der Kanten aus, so hat jede Wiederholung also mindestens Wahrscheinlichkeit  $\frac{2}{n(n-1)}$ , einen MinCut zu liefern. Wegen der Unabhängigkeit ist damit die Wahrscheinlichkeit, in  $m$  Versuchen mindestens einen MinCut zu finden, mindestens

$$\begin{aligned}
 \mathbb{P}(\text{in } m \text{ Versuchen mindestens ein MinCut}) &= 1 - \mathbb{P}(\text{in } m \text{ Versuchen kein MinCut}) \\
 &= 1 - \left(1 - \frac{2}{n(n-1)}\right)^m \\
 &\geq 1 - e^{-2m/n}.
 \end{aligned}$$

Durch Erhöhung der Anzahl Wiederholungen  $m$  kann diese Wahrscheinlichkeit beliebig nah an 1 gebracht werden.

## 12.3 Ein probabilistischer Existenzbeweis: Graphenfärbung

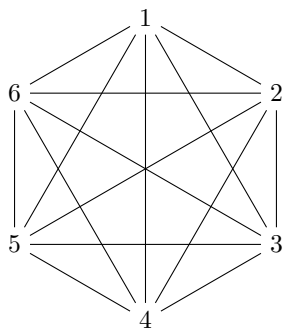
Als “probabilistische Methode” bezeichnet man eine Strategie zur Existenz von mathematischen Objekten mit bestimmten Eigenschaften, welche auf der Wahrscheinlichkeitstheorie basiert. Die Grundidee besteht in der Konstruktion eines Wahrscheinlichkeitsraumes, aus dem mit positiver Wahrscheinlichkeit ein Objekt mit der gewünschten Eigenschaft ausgewählt wird. Das bedeutet aber, dass es ein solches Objekt tatsächlich geben muss.

Die Beweismethode ist dabei oft in einen stochastischen Algorithmus zur expliziten Konstruktion eines solchen Objektes übersetzbar. Wir illustrieren die Methode an einem einfachen Beispiel.

### Definition 12.5

Der **vollständige Graph der Größe  $n$**  ist der (eindeutige) Graph  $G = (E, V)$  mit  $n$  Knoten und  $\binom{n}{2}$  Kanten, bei dem keine Mehrfachkanten zugelassen sind. ♦

Beim vollständigen Graphen mit  $n$  Knoten ist also jedes der  $\binom{n}{2}$  Paare von Knoten mit einer Kante verbunden, siehe Abbildung 12.2.

Abb. 12.2: Der vollständige Graph mit  $n = 6$  Knoten**Definition 12.6**

Ein **Subgraph** eines Graphen  $G = (E, V)$  besteht aus einer Teilmenge  $E'$  von  $E$  und einer Teilmenge  $V'$  von  $V$ , so dass die Elemente von  $E'$  Kanten zwischen Elementen von  $V'$  sind.  $\blacklozenge$

Wir betrachten folgendes *Färbeproblem*: Gegeben sei der vollständige Graph  $G$  der Größe  $n$ . Ist es möglich, die Kanten von  $G$  mit zwei verschiedenen Farben derart einzufärben, dass kein vollständiger Subgraph der Größe  $k$  entsteht, dessen Kante alle dieselbe Farbe haben? Ein Kriterium ist in folgendem Satz gegeben.

**Satz 12.3**

Falls  $\binom{n}{k} 2^{-(\binom{k}{2}+1)} < 1$  ist, so gibt es eine Färbung mit zwei Farben des vollständigen Graphen der Größe  $n$  derart, dass kein vollständiger Subgraph der Größe  $k$  einfarbig ist.

**Beweis 12.3** Der vollständige Graph mit  $n$  Knoten hat genau  $\binom{n}{2}$  Kanten. Wir stellen uns vor, dass wir einfach jede Kante unabhängig von den anderen Kanten mit einer der beiden Farben einfärben, wobei jede Farbe mit Wahrscheinlichkeit  $1/2$  ausgewählt wird.

Im vollständigen Graphen mit  $n$  Knoten gibt es genau  $\binom{n}{k}$  vollständige Subgraphen mit  $k$  Knoten, denn die Anzahl Möglichkeiten,  $k$  von den  $n$  Knoten auszuwählen ist gleich dem Binomialkoeffizienten  $\binom{n}{k}$ , und der vollständige Subgraph ist durch seine Knotenmenge festgelegt. Wir nummerieren diese Subgraphen mit  $i = 1, \dots, \binom{n}{k}$ , und schreiben  $A_i$  für das Ereignis, dass der  $i$ -te Subgraph einfarbig eingefärbt wurde.

Wenn wir nun einen festen Subgraphen  $i$  der Größe  $k$  betrachten, so hat dieser  $\binom{k}{2}$  Kanten. Damit ist die Wahrscheinlichkeit, dass alle Kanten dieses Subgraphen dieselbe Farbe haben, gleich

$$\mathbb{P}(A_i) = 2 \cdot \frac{1}{2^{\binom{k}{2}}} = 2^{-(\binom{k}{2}+1)}.$$

Es gilt nämlich 2 Möglichkeiten für die Wahl der Farbe. Somit ist die Wahrscheinlichkeit, dass es mindestens einen einfarbigen Subgraphen gibt, gegeben durch

$$\mathbb{P}(A_1 \cup \dots \cup A_{\binom{n}{k}}) \leq \sum_{i=1}^{\binom{n}{k}} \mathbb{P}(A_i) = \binom{n}{k} 2^{-(\binom{k}{2}+1)}.$$



Nach der Voraussetzung des Satzes ist die gesuchte Wahrscheinlichkeit, dass es keinen einfarbigen Subgraphen gibt, gleich

$$P(A_1^c \cup \dots \cup A_{\binom{n}{k}}^c) = 1 - \mathbb{P}(A_1 \cup \dots \cup A_{\binom{n}{k}}) \geq 1 - \binom{n}{k} 2^{-\binom{k}{2}+1} > 0.$$

□

Der “Algorithmus” der diesem Beweis zu Grunde liegt, färbt lediglich jede Kante unabhängig mit Wahrscheinlichkeit  $1/2$  mit Farbe 1 ein, und andernfalls mit Farbe 2. Anschließend muss lediglich jeder vollständige Subgraph auf Einfarbigkeit überprüft werden.

### Zusammenfassung

Randomisierte Algorithmen können in einer Vielzahl von Situationen nützlich sein. Man unterscheidet insbesondere zwischen Monte-Carlo und Las-Vegas Algorithmen. Typische Fragestellungen sind beispielsweise die Laufzeitanalyse oder die Erfolgswahrscheinlichkeit.

## Aufgaben

**12.1** Zeigen Sie, dass im vorgestellten MinCut Algorithmus die erwartete Anzahl unabhängiger Versuche, bis zum ersten Mal ein MinCut gefunden wird, höchstens  $\frac{n(n-1)}{2}$  ist.

**12.2** Man zeige, dass für  $k \geq 3$  und  $n, k$  derart, dass  $n^k \leq 2^{k/2}$  gilt, dass  $\binom{n}{k} 2^{-\binom{k}{2}+1} \leq \frac{2^{k/2}+1}{k!}$  gilt. Daraus leite man eine weitere hinreichende Bedingung für die Existenz einer Zweifärbung eines vollständigen Graphen mit  $n$  Knoten ohne vollständigen Subgraphen mit  $k$  Knoten her.

**12.3** Man zeige, dass eine Zweifärbung des vollständigen Graphen mit  $n = 1000$  Knoten ohne vollständigen Subgraphen mit  $k = 20$  Knoten existiert.



# 13 Verzweigungsprozesse und erzeugende Funktionen

---

## Übersicht

13.1 Galton-Watson Verzweigungsprozesse .....	182
13.2 Wahrscheinlichkeitserzeugende Funktionen .....	184
13.3 Aussterbewahrscheinlichkeit im Galton-Watson Prozess .....	185
13.4 Weitere Anwendungen der wahrscheinlichkeitserzeugenden Funktion .....	188

---

Verzweigungsprozesse können als Algorithmus für die Erzeugung zufälliger Bäume aufgefasst werden. Wir betrachten in diesem Kapitel eine spezielle Konstruktion, welche auf den sogenannten Galton-Watson Prozess führt. Wir führen wahrscheinlichkeitserzeugenden Funktionen ein, um das Aussterbeverhalten solcher Prozesse zu analysieren.

### Lernziele

- Eine Konstruktion von Verzweigungsprozessen kennen
- Wahrscheinlichkeitserzeugende Funktionen berechnen können und ihre Eigenschaften kennen
- Die Aussterbewahrscheinlichkeit eines Galton-Watson-Prozesses berechnen können

### Vorkenntnisse

Dabei werden Kenntnisse aus den Bereichen Zufallsvariablen, Verteilungen, Erwartungswerte, bedingte Wahrscheinlichkeiten und Markov-Ketten benötigt.

## 13.1 Galton-Watson Verzweigungsprozesse

Ziel dieses Abschnitts ist es, eine bestimmte Klasse von zufälligen Bäumen zu konstruieren. Dabei denken wir in Ebenen oder Generationen. Ein *Baum* im Sinne der Graphentheorie ist ein zusammenhängender Graph bestehend aus Knoten und Kanten, (vgl. Definition 12.3), bei dem keine Mehrfachkanten und keine geschlossenen Pfade vorkommen. Es gibt keine Möglichkeit, durch Springen entlang von Kanten zum Ausgangspunkt zurückzukommen, ohne dass man Kanten mehrfach benutzt.

In unserer Konstruktion gehört jeder Knoten (oder jede Verzweigung) des Baumes zu einer bestimmten Ebene oder Generation. Die Kanten bilden Verbindungen zwischen den Knoten verschiedener Ebenen. Wir starten in Ebene 0 mit einer (einzigen) Wurzel. Diese hat eine zufällige Anzahl *Nachkommen*  $X$ , welche die Knoten in Generation 1 bilden, und jeweils mit der Wurzel durch eine Kante verbunden sind. So wird iterativ verfahren: Jeder Knoten in Ebene  $i \in \mathbb{N}$  hat wieder eine Anzahl Nachkommen, die in Ebene  $i + 1$  liegen, und mit ihrem eindeutigen Vorfahren in Ebene  $i$  durch eine Kante verbunden sind. Die Anzahl Nachkommen wird dabei für jeden Knoten unabhängig und gemäß derselben Verteilung ausgewürfelt.

### Definition 13.1 (Nachkommenverteilung)

Die **Nachkommenverteilung** eines Verzweigungsprozesses ist die Verteilung  $(q_k)_{k \in \mathbb{N}_0}$  einer Zufallsvariablen mit Werten in  $\mathbb{N}_0$ . Dabei ist  $q_k$  die Wahrscheinlichkeit, dass ein beliebiger Knoten genau  $k$  Nachkommen in der nächsten Ebene hat. Insbesondere gilt also für die Anzahl Nachkommen  $X$  der Wurzel  $q_k = \mathbb{P}(X = k)$ . ♦

Wir geben einen Algorithmus an, welcher obiges Vorgehen formalisiert. Dabei startet man nicht zwingend mit einer einzigen Wurzel, der Prozess kann mit einer beliebigen Anzahl Knoten in Generation 0 gestartet werden.

### Algorithmus: Galton-Watson Prozess

Vorgegeben sei eine Nachkommenverteilung  $q_k, k \in \mathbb{N}_0$ .

1.  $i = 0$  : Wähle  $Z_0 \in \mathbb{N}$ . Oft:  $Z_0 = 1$ , oder auch zufällig
2.  $i \rightarrow i + 1$  :
  - Gegeben  $Z_i = m$ , erzeuge  $m$  unabhängige Zufallsvariablen  $X_1^{(i)}, \dots, X_m^{(i)}$  mit Verteilung  $(q_k)_{k \in \mathbb{N}_0}$ , also mit
 
$$\mathbb{P}(X_j^{(i)} = k) = q_k$$
  - Setze  $Z_{i+1} = X_1^{(i)} + \dots + X_m^{(i)}$
3. Wiederhole 2. bis zum ersten Mal  $Z_i = 0$  ist, oder bis zu einer vorher festgelegten Anzahl Ebenen  $i$ .

**Definition 13.2**

Die so erzeugte Folge von Zufallsvariablen  $(Z_i)_{i \in \mathbb{N}_0}$  heißt **Galton-Watson Prozess**. Die Zufallsvariablen  $X_j^{(i)}$  werden dabei interpretiert als die Anzahl der Nachkommen von Knoten Nr.  $j$  aus Generation  $i$ , wobei die Nachkommen zu Generation  $i + 1$  gehören.



Der Galton-Watson Prozess kann auf einfache Weise graphisch dargestellt werden: Auf Ebene 0 zeichne  $Z_0$  Knoten. Auf Ebene 1 zeichne  $X_1^{(0)}$  Knoten mit Verbindung zu Knoten 1 von Ebene 0,  $X_2^{(0)}$  Knoten mit Verbindung zu Knoten 2 von Ebene 0, etc. Analog für jede Ebene.

(Skizze)

Man kann sich leicht klarmachen, dass ein Galton-Watson Prozess eine Markov-Kette bildet. Die Übergangswahrscheinlichkeiten dieser Markov-Kette können zwar im Prinzip mit Hilfe der Faltungsformel aus Satz 4.4 berechnet werden, was jedoch schnell kompliziert und unpraktikabel wird. Im nächsten Abschnitt werden wir deshalb eine andere Methode kennen lernen, um das Langzeitverhalten von Galton-Watson Prozessen zu studieren. Wir betrachten jedoch zuerst ein Beispiel.

**Beispiel 13.1 (Binärer Galton-Watson Prozess)**

Ein Programm ruft zu einem bestimmten Zweck einen Prozess auf, welcher wiederum bei jedem Aufruf eine zufällige Anzahl gleichartiger Prozesse triggert. Dabei stellt sich in natürlicher Weise die Frage: Wie oft wird dieser Prozess insgesamt aufgerufen? Stoppt das Programm von selbst, weil irgendwann keine Prozesse mehr ausgelöst werden, oder muss abgebrochen werden, weil es immer mehr Prozesse werden?

Für eine erste Untersuchung dieser Fragen betrachten wir einen Spezialfall. Sei  $p \in ]0, 1[$ , und sei die Nachkommenverteilung gegeben durch

$$q_0 = p, \quad q_2 = 1 - p, \quad \text{und} \quad q_k = 0 \quad \text{für alle anderen } k \in \mathbb{N}.$$

Mit anderen Worten, jeder Knoten hat entweder 0 oder 2 Nachkommen.

Es ist nun relativ plausibel zu denken, dass es zwei Szenarien gibt: Falls  $p$  nahe an 0 liegt, kommen 2 Nachkommen deutlich häufiger vor als 0 Nachkommen, so dass die Anzahl Knoten pro Ebene schnell ansteigt (im Grenzfall  $p = 0$ , der nicht mehr zufällig ist, gibt es in Ebene  $i$  offensichtlich  $2^i$  Knoten). Ist umgekehrt  $p$  nahe an 1, so haben nur sehr wenige Knoten überhaupt Nachkommen, und es ist davon auszugehen, dass irgendwann eine Ebene ohne Knoten vorkommt, der Prozess also abbricht. Es gilt

$$\mathbb{E}[Z_1 | Z_0 = 1] = 2(1 - p),$$

und allgemeiner

$$\mathbb{E}[Z_i | Z_{i-1} = a] = 2a(1 - p), \quad a \in \mathbb{N}.$$

Das bedeutet, für  $p < 1/2$  nimmt die Anzahl Knoten pro Ebene in Erwartung stets zu. In der Tat werden wir später zeigen, dass der Prozess mit Wahrscheinlichkeit 1 stoppt, falls  $p \geq 1/2$ , und mit positiver Wahrscheinlichkeit unendlich lange weiter laufen würde, falls  $p < 1/2$  ist.



**Definition 13.3**

Sei  $X$  die Anzahl Nachkommen eines festen Knoten, d.h.  $X$  ist eine Zufallsvariable mit Verteilung  $(q_k)_{k \in \mathbb{N}_0}$ . Ein Galton-Watson-Prozess heißt **subkritisch**, falls  $\mathbb{E}[X] < 1$  ist, **kritisch** falls  $\mathbb{E}[X] = 1$  ist, und **superkritisch** falls  $\mathbb{E}[X] > 1$  ist.  $\blacklozenge$

Im obigen Beispiel ist der Prozess subkritisch für  $p > 1/2$ , superkritisch für  $p < 1/2$ , und kritisch für  $p = 1/2$ . Daher liegt die folgende Vermutung nahe: Ein subkritischer Galton-Watson Prozess endet irgendwann von selbst (wir sagen *er stirbt aus*), ein superkritischer Prozess stirbt nicht zwingend aus. Der übernächste Abschnitt liefert eine Beweisskizze dieser Vermutung. Zuerst führen wir aber noch ein allgemeines Hilfsmittel aus der Wahrscheinlichkeitstheorie ein.

## 13.2 Wahrscheinlichkeitserzeugende Funktionen

Die Analyse der Aussterbewahrscheinlichkeit von Galton-Watson Prozessen beruht auf sogenannten wahrscheinlichkeitserzeugenden Funktionen. Diese Funktionen sind auch in anderen Bereichen der Wahrscheinlichkeitstheorie ein nützliches Hilfsmittel.

**Definition 13.4**

Sei  $X$  eine Zufallsvariable. Die **wahrscheinlichkeitserzeugende Funktion** (manchmal auch nur erzeugende Funktion) von  $X$  ist definiert durch

$$g_X(t) = \mathbb{E}[t^X], \quad t \in [0, 1].$$

 $\blacklozenge$ 

Im Fall von Galton-Watson Prozessen arbeiten wir mit Zufallsvariablen, welche Werte in  $\mathbb{N}_0$  annehmen. In diesem Fall berechnet sich die erzeugende Funktion nach Definition des Erwartungswertes und Satz 5.1 (e) als

$$g_X(t) = \sum_{k=0}^{\infty} t^k \mathbb{P}(X = k).$$

**Beispiel 13.2**

Im binären Galton-Watson Prozess aus Beispiel 13.1 war die Verteilung  $\mathbb{P}(X = 0) = p, \mathbb{P}(X = 2) = 1 - p$ . Somit ist die erzeugende Funktion gleich  $g_X(t) = p + (1 - p)t^2$ .  $\blacksquare$

**Beispiel 13.3 (Poisson-Verteilung)**

Bei der Poisson-Verteilung zum Parameter  $\lambda$  gilt  $\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k \in \mathbb{N}_0$ . Somit ist die erzeugende Funktion gegeben durch

$$g_X(t) = \sum_{k=0}^{\infty} t^k \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(t\lambda)^k}{k!} = e^{-\lambda} e^{t\lambda} = e^{\lambda(t-1)}.$$

 $\blacksquare$

Die Bedeutung dieser Funktionen liegt in einer Reihe besonderer Eigenschaften.

**Satz 13.1 (Wahrscheinlichkeitserzeugende Funktion und Verteilung)**

Sei  $X$  eine Zufallsvariable mit  $X(\Omega) \subseteq \mathbb{N}_0$  und wahrscheinlichkeitserzeugender Funktion  $g_X$ . Dann ist durch  $g_X$  die Verteilung von  $X$  festgelegt, und es gilt

$$\mathbb{P}(X = n) = \frac{g_X^{(n)}(0)}{n!}, \quad n \geq 0,$$

wobei  $g_X^{(n)}$  die  $n$ -te Ableitung von  $g_X$  ist ( $g_X^{(0)} = g_X$ ). Außerdem gilt  $g_X(1) = 1$ , und

$$\mathbb{E}[X] = g'_X(1).$$

**Beweis 13.1** Durch Einsetzen von  $t = 0$  sieht man  $g_X(0) = \sum_{k=0}^{\infty} 0^k \mathbb{P}(X = k) = \mathbb{P}(X = 0)$ , da  $0^0 = 1$  ist, und  $0^n = 0$  für alle anderen Werte von  $n$ . Allgemein gilt  $g_X(t) = \sum_{n=0}^{\infty} t^n \mathbb{P}(X = n)$ . Ableiten auf beiden Seiten (auf der rechten Seite darf die Ableitung unter die Summe gezogen werden) führt auf

$$g'_X(t) = \sum_{k=1}^{\infty} k t^{k-1} \mathbb{P}(X = k).$$

Setzt man  $t = 0$  ein, so sind alle Summanden auf der rechten Seite gleich 0, mit Ausnahme des Summanden für  $k = 1$ . Somit folgt  $g'_X(0) = \mathbb{P}(X = 1)$ . Für  $n \geq 2$  folgt die Aussage analog durch mehrfaches Ableiten und Umformen. Weiter gilt  $g_X(1) = \sum_{k=0}^{\infty} \mathbb{P}(X = k) = 1$ , und  $g'_X(1) = \sum_{k=1}^{\infty} k \mathbb{P}(X = k) = \mathbb{E}[X]$ .  $\square$

**Satz 13.2**

Sei  $X$  eine Zufallsvariable mit Werten in  $\mathbb{N}_0$ , und es gelte  $\mathbb{P}(X \geq 2) > 0$ . Dann gelten  $g'_X > 0$  und  $g''_X > 0$ .

**Beweis 13.2** Die Voraussetzung  $\mathbb{P}(X \geq 2) > 0$  impliziert, dass es mindestens ein  $k \geq 2$  gib, so dass  $\mathbb{P}(X = k) > 0$  ist. Somit folgt die Behauptung aus Satz 13.1.  $\square$

Dieser Satz impliziert, dass unter den getroffenen Annahmen die wahrscheinlichkeitserzeugende Funktion  $g_X$  auf  $]0, 1[$  monoton wachsend und *konvex* ist. Der Begriff konvex bedeutet dabei, dass die Tangente an den Graphen von  $g_X$  in jedem Punkt unterhalb der Funktion liegt, oder anders ausgedrückt, dass die Krümmung des Funktionsgraphen von  $g_X$  positiv ist.

## 13.3 Aussterbewahrscheinlichkeit im Galton-Watson Prozess

Mit Hilfe der Eigenschaften der wahrscheinlichkeitserzeugenden Funktion kann man die Wahrscheinlichkeit berechnen, dass der Galton-Watson Prozess irgendwann ausstirbt. Dies erfordert jedoch noch einige Vorüberlegungen.

Es sei  $(Z_i)_{i \in \mathbb{N}_0}$  ein Galton-Watson-Prozess mit Nachkommenverteilung  $q_k, k \in \mathbb{N}_0$ . Weiter sei

$$g_X(t) = \sum_{k=0}^{\infty} q_k t^k$$

die wahrscheinlichkeitserzeugende Funktion der Nachkommenverteilung.

### Satz 13.3

Es sei  $(Z_n)_{n \in \mathbb{N}_0}$  ein Galton-Watson Prozess mit  $Z_0 = 1$  und erzeugender Funktion  $g_X$  der Nachkommenverteilung. Die wahrscheinlichkeitserzeugende Funktion der Anzahl Knoten auf Ebene  $n$  ist gegeben durch

$$g_{Z_n}(t) = g_{Z_{n-1}}(g_X(t)) = \underbrace{(g_X \circ \dots \circ g_X)}_{n \text{ Mal}}(t),$$

wobei rechts die  $n$ -fache Verknüpfung von  $g_X$  mit sich selbst steht.

**Beweis 13.3** Nach Definition 13.2 ist  $Z_n = \sum_{i=1}^{Z_{n-1}} X^{(i)}$ , wobei die  $X^{(i)}$  unabhängig und mit Verteilung  $(q_k)_{k \in \mathbb{N}_0}$  sind, also erzeugende Funktion  $g_X$  haben. Es gilt nach Definition 13.4 und dem Satz von der Gesamtwahrscheinlichkeit

$$\begin{aligned} g_{Z_n}(t) &= \mathbb{E}[t^{Z_n}] = \mathbb{E} \left[ t^{\sum_{i=1}^{Z_{n-1}} X^{(i)}} \right] = \sum_{m=0}^{\infty} \mathbb{E} \left[ t^{\sum_{i=1}^m X^{(i)}} \mid Z_{n-1} = m \right] \mathbb{P}(Z_{n-1} = m) \\ &= \sum_{m=0}^{\infty} \mathbb{E} \left[ \prod_{i=1}^m t^{X^{(i)}} \mid Z_{n-1} = m \right] \mathbb{P}(Z_{n-1} = m) \\ &= \sum_{m=0}^{\infty} \prod_{i=1}^m \mathbb{E} \left[ t^{X^{(i)}} \right] \mathbb{P}(Z_{n-1} = m) = \sum_{m=0}^{\infty} \mathbb{E} \left[ t^{X^{(1)}} \right]^m \mathbb{P}(Z_{n-1} = m) \\ &= \sum_{m=0}^{\infty} g_X(t)^m \mathbb{P}(Z_{n-1} = m) = g_{Z_{n-1}}(g_X(t)). \end{aligned}$$

Dabei wurde in der dritten Zeile die Unabhängigkeit benutzt. Dies zeigt die erste Gleichung in der Behauptung. Die zweite folgt induktiv aus der Tatsache dass  $Z_1 = X$  und somit  $g_{Z_1} = g_X$  gilt.  $\square$

Allgemein zeigt dieser Beweis, dass die wahrscheinlichkeitserzeugende Funktion einer Summe von unabhängigen, identisch verteilten Zufallsvariablen  $X_i$  mit einer zufälligen Anzahl Summanden  $Z$  gegeben ist durch  $g_Z \circ g_{X_1}$ , vergleiche Aufgabe 13.3.

### Definition 13.5

Die **Aussterbewahrscheinlichkeit** des Galton-Watson Prozesses  $(Z_i)_{i \in \mathbb{N}_0}$  sei

$$P(A_Z) := \mathbb{P}(\exists n \in \mathbb{N} : Z_n = 0) = \lim_{n \rightarrow \infty} \mathbb{P}(Z_n = 0),$$

also die Wahrscheinlichkeit, dass es zu irgendeinem Zeitpunkt  $n$  keine Knoten mehr gibt.  $\blacklozenge$



Wegen Satz 13.1 und Lemma 13.3 gilt

$$\mathbb{P}(Z_n = 0) = g_{Z_n}(0) = \underbrace{(g_X \circ \dots \circ g_X)}_{n \text{ Mal}}(0).$$

Dies werden wir gleich im Beweis des Hauptresultates dieses Kapitels verwenden.

**Satz 13.4 (Aussterbewahrscheinlichkeit des Galton-Watson Prozesses)**

Es sei  $(Z_i)_{i \in \mathbb{N}_0}$  ein Galton-Watson-Prozess mit Nachkommenverteilung  $(q_k)_{k \in \mathbb{N}_0}$ , mit  $q_0 > 0$ . Es sei  $g_X(t)$  die wahrscheinlichkeitserzeugende Funktion der Nachkommenverteilung. Dann gelten:

- Falls  $\mathbb{E}[X] = g'(1) \leq 1$  ist, so ist die Aussterbewahrscheinlichkeit  $P(A_Z) = 1$ .
- Falls  $\mathbb{E}[X] = g'(1) > 1$  ist, so ist die Aussterbewahrscheinlichkeit  $(P(A_Z))$  strikt kleiner als 1, und zwar ist sie gegeben durch die kleinste Zahl  $s \in [0, 1]$ , welche die Gleichung

$$g_X(s) = s$$

löst.

Um die Aussterbewahrscheinlichkeit zu berechnen, muss also im ersten Schritt  $\mathbb{E}[X]$  bestimmt werden. Ist dieser Erwartungswert kleiner oder gleich 1, so ist man fertig, denn man weiß nun, dass der Prozess sicher irgendwann ausstirbt (man weiß aber nicht, wie lange das dauert). Ist der Erwartungswert strikt größer als 1, so muss die Gleichung  $g_X(s) = s$  gelöst werden. Die kleinste Lösung in  $[0, 1]$  gibt dann die Aussterbewahrscheinlichkeit. Dabei ist zu beachten, dass  $P(A_Z) < 1$  ist, aber nicht  $P(A_Z) = 0$ . Der Prozess stirbt also immer noch mit einer gewissen Wahrscheinlichkeit aus.

**Beweis 13.4** Der Beweis beruht auf einer genauen Analyse von  $g_X(t)$  der wahrscheinlichkeitserzeugenden Funktion der Nachkommenverteilung, und insbesondere ihrer Fixpunkte, d.h. Werten  $s \in [0, 1]$  mit  $g_X(s) = s$ . Wir geben die Argumente an Hand einer Skizze wieder, auf einen komplett formalen Beweis aller Schritte verzichten wir.

Aus der Definition der Wahrscheinlichkeitserzeugenden Funktion und den Sätzen 13.1 und 13.2 wissen wir, dass  $g_X$  die folgenden Eigenschaften hat:

- $g_X$  ist auf ganz  $[0, 1]$  definiert, und auf  $]0, 1[$  monoton wachsend und konvex, also  $g'_X(t) > 0$  und  $g''_X(t) > 0$  für alle  $t \in ]0, 1[$ .
- Es gelten  $g_X(0) = q_0$ ,  $g_X(1) = 1$ , und  $g'_X(1) = \mathbb{E}[X]$ .

Diese Eigenschaften kann man in einer Skizze des Funktionsgraphen von  $g_X$  zusammenfassen:

(Skizze)

Zusätzlich kann man in die Skizze den Graphen der Funktion  $f(t) = t$ , also der Identitätsfunktion, einzeichnen. Dieser bildet die Diagonale des Quadrats  $[0, 1] \times [0, 1]$ . Dann erkennt man, dass sich die Graphen von  $g_X$  und von  $f$  mindestens im Punkt  $(1, 1)$  berühren, da  $g_X(1) = f(1) = 1$  ist. Je nachdem wie schnell  $g_X$  wächst, gibt es noch einen zweiten Schnittpunkt der beiden Graphen, in einem Punkt  $s < 1$ .

(Skizze)

Diesen zweiten Schnittpunkt gibt es dann und nur dann, wenn  $g'_X(1) > 1$  ist, da die Existenz eines solchen Punktes impliziert, dass der Graph von  $g_X$  die Winkelhalbierende schneidet, und dann mit einer Steigung größer als 1 wachsen muss, um den Wert  $g_X(1) = 1$  erreichen zu können. Nun ist aber genau  $g'_X(1) = e[X]$  nach Satz 13.1, also existiert ein solcher zweiter Schnittpunkt genau im Fall dass der Galton-Watson Prozess superkritisch ist.

Dieser zweite Schnittpunkt  $s$  von  $g_X$  und  $f$  ist ein Fixpunkt von  $g_X$ , denn es gilt  $g_X(s) = f(s) = s$ . Man kann beweisen, dass die Folge  $g_X(0), g_X(g_X(0)), \dots$  gegen  $g_X(s) = s$  konvergiert, anschaulich ist das in der Skizze zu erkennen:

(Skizze).

Damit folgt für die Aussterbewahrscheinlichkeit

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n = 0) = g_{Z_n}(0) = \lim_{n \rightarrow \infty} (g_X \circ \dots \circ g_X)(0) = s.$$

Falls kein zweiter Schnittpunkt existiert, was genau dann der Fall ist, wenn  $\mathbb{E}[X] \leq 1$  ist, da dann der Graph von  $g_X$  immer oberhalb von des Graphen von  $f$  liegt, kann man ebenso beweisen, dass  $g_X(0), g_X(g_X(0)), \dots$  gegen 1 konvergiert.

(Skizze)

Dies liefert die Behauptung des Satzes. □

Streng genommen haben wir obigen Satz nur unter der Voraussetzung  $\mathbb{P}(X \geq 2) > 0$  bewiesen, die für Satz 13.2 nötig ist. Man kann sich jedoch leicht überlegen, dass er auch sonst gilt, vgl. Aufgabe 13.5.

### Beispiel 13.4 (Binärer Galton-Watson Prozess, Fortsetzung)

Im Beispiel 13.1 ist die wahrscheinlichkeitserzeugende Funktion

$$g_X(t) = p + (1-p)t^2.$$

Die Gleichung  $g_X(t) = t$  hat die Lösungen  $t = 1$  und  $t = \frac{p}{1-p}$ . Falls  $p < 1/2$  ist, also genau dann, wenn der Prozess superkritisch ist, ist die zweite Lösung  $t = \frac{p}{1-p}$  kleiner als 1, und somit ist in diesem Fall die Aussterbewahrscheinlichkeit

$$P(A_Z) = \lim_{n \rightarrow \infty} \mathbb{P}(Z_n = 0) = \frac{p}{1-p}.$$

Andernfalls ist  $P(A_Z) = 1$ . ■

## 13.4 Weitere Anwendungen der wahrscheinlichkeitserzeugenden Funktion

Nach Satz 13.1 legt die wahrscheinlichkeitserzeugende Funktion die Verteilung der zugehörigen Zufallsvariablen fest. Diese Eigenschaft hat weitreichende Anwendungen, eine einfache davon stellen wir hier vor.

**Satz 13.5 (Produktformel)**

Seien  $X$  und  $Y$  zwei unabhängige Zufallsvariablen mit wahrscheinlichkeitserzeugenden Funktionen  $g_X$  und  $g_Y$ . Dann hat die Zufallsvariable  $X + Y$  die wahrscheinlichkeitserzeugende Funktion

$$g_{X+Y}(t) = g_X(t) \cdot g_Y(t).$$

**Beweis 13.5** Unter Verwendung der Potenzgesetze und der Unabhängigkeit findet man

$$g_{X+Y}(t) = \mathbb{E}[t^{X+Y}] = \mathbb{E}[t^X t^Y] = \mathbb{E}[t^X] \mathbb{E}[t^Y] = g_X(t) g_Y(t).$$

□

**Beispiel 13.5 (Summe von Poisson-verteilten Zufallsvariablen)**

$X$  und  $Y$  seien unabhängige, Poisson-verteilte Zufallsvariablen mit Parametern  $\mu$  bzw.  $\lambda$ . In Beispiel 13.3 haben wir die entsprechende erzeugende Funktion berechnet. Nun folgt aus Satz 13.5

$$g_{X+Y}(t) = g_X(t) g_Y(t) = e^{\mu(t-1)} e^{\lambda(t-1)} = e^{(\mu+\lambda)(t-1)}.$$

Die erzeugende Funktion von  $X + Y$  hat also wieder dieselbe Form wie die erzeugende Funktion der einzelnen Zufallsvariablen  $X$  bzw.  $Y$ , was genau die erzeugende Funktion einer Poisson-verteilten Zufallsvariablen mit Parameter  $\mu + \lambda$  ist. Da dadurch die Verteilung festgelegt ist, muss  $X + Y$  also wieder Poisson-verteilt sein mit der Summe der Parameter der Summanden. ■

Dieses Ergebnis hatten wir bereits in Kapitel 4 mit Hilfe der Faltungsformel 4.4 gefunden, der Beweis mittels wahrscheinlichkeitserzeugenden Funktionen ist jedoch bedeutend einfacher.

**Zusammenfassung**

Galton-Watson-Prozesse beschreiben die zeitliche Entwicklung von idealisierten Populationen und können als Algorithmus zur Erzeugung zufälliger Bäume interpretiert werden. Ihr Langzeitverhalten kann mittels wahrscheinlichkeitserzeugender Funktionen analysiert werden. Es stellt sich heraus, dass ein Galton-Watson-Prozess mit einer erwarteten Nachkommenanzahl von weniger als einem Nachkommen sicher irgendwann ausstirbt, und er andernfalls mit positiver Wahrscheinlichkeit überlebt.

**Aufgaben**

**13.1** Es sei  $X$  eine Zufallsvariable mit wahrscheinlichkeitserzeugender Funktion  $g_X$ . Zeigen Sie:

$$\mathbb{E}[X(X-1)\dots(X-k+1)] = g_X^{(k)}(1).$$

Folgern Sie daraus

$$\mathbb{V}(X) = g_X''(1) + g_X'(1) - (g_X'(1))^2$$

**13.2** Berechnen Sie die Wahrscheinlichkeitserzeugende Funktion der Bernoulli- und der Binomialverteilung. *Hinweis:* Satz 13.5 verwenden.

**13.3** Es seien  $X_i, i \in \mathbb{N}$  unabhängige, identisch verteilte Zufallsvariablen, und  $Z$  unabhängig davon. Sei  $Y = \sum_{i=1}^Z X_i$ . Zeigen Sie:  $g_Y(t) = g_Z(g_{X_1}(t))$ . *Hinweis:* Beweis von Satz 13.3.

**13.4** Sei  $X$  eine Zufallsvariable mit  $X(\Omega) = \{0, 1, 2\}$  und Verteilung  $(q_k)_{k \in \{0,1,2\}}$  gegeben durch

$$q_k = \begin{cases} \frac{1}{2}, & k = 0, \\ \frac{1}{4}, & k = 1, \\ \frac{1}{4} & k = 2. \end{cases}$$

Weiter sei  $(Z_n)_{n \in \mathbb{N}_0}$  ein Galton-Watson-Verzweigungsprozess mit Nachkommenverteilung  $(q_k)_{k \in \{0,1,2\}}$ . (a) Berechnen Sie  $\mathbb{E}[X]$  sowie die erzeugende Funktion  $g_X(t) = \mathbb{E}[t^X]$ . (b) Berechnen Sie die Aussterbewahrscheinlichkeit.

**13.5** Es sei  $q_0 = p, q_1 = 1 - p$  für  $p \in ]0, 1[$ . Man zeige ohne die Verwendung von Satz 13.4, dass der zugehörige Galton-Watson Prozess mit Wahrscheinlichkeit 1 ausstirbt, und dass  $\min\{n : Z_n = 0\}$  geometrisch verteilt mit Parameter  $p$  ist.

# 14 Warteschlangenmodelle und Markov-Ketten in stetiger Zeit

---

## Übersicht

14.1 Warteschlangenmodelle in stetiger Zeit .....	192
14.2 Eigenschaften der Exponentialverteilung .....	194
14.3 Markov-Ketten in stetiger Zeit .....	197

---

Warteschlangen (englisch: queues) bezeichnen Systeme von Bedienern (servers) und Kunden (jobs). Dabei werden die Kunden von den Bedienern nach vorgegebenen Regeln abgearbeitet. In solchen Modellen sind Fragen der *Stabilität* von Interesse, d.h. ob die Server die Menge der eingehenden Jobs abarbeiten können. Wir betrachten eine Klasse von Warteschlangenmodellen in stetiger Zeit, bei denen gewisse spezielle Eigenschaften der Exponentialverteilung eine wichtige Rolle spielen. Diese führen dann allgemeiner auf die Konstruktion von Markov-Ketten in stetiger Zeit mit endlichem Zustandsraum.

### Lernziele

- Wichtige Warteschlangenmodelle in stetiger Zeit kennen, und auf Stabilität untersuchen können
- Wichtige Eigenschaften der Exponentialverteilung kennen
- Eine Konstruktion und wichtige Eigenschaften von Markov-Ketten in stetiger Zeit kennen

### Vorkenntnisse

Dabei werden Kenntnisse aus den Bereichen Zufallsvariablen, Unabhängigkeit, Markov-Ketten und Exponentialverteilung benötigt.

## 14.1 Warteschlangenmodelle in stetiger Zeit

Modelle für Warteschlangen oder queues existieren in verschiedener Form und von unterschiedlicher Komplexität. Die folgenden Annahmen werden dabei meistens getroffen:

- Warteschlangen bestehen aus einem System von mehreren Bedienern (Servern)
- Kunden kommen unabhängig voneinander zu zufälligen Zeiten im System an
- Kunden werden bedient, die Bediendauer jedes Kunden ist zufällig, Bedienreihenfolge folgt vorgegebenen Regeln (zufällig oder deterministisch)
- Nach Bedienung verlassen Kunden die Warteschlange

(Skizze)

In Beispiel 11.12 haben wir eine Warteschlange in diskreter Zeit behandelt. Dabei haben wir untersucht, unter welchen Bedingungen an die Parameter die eingehenden Jobs von den Servern in endlicher Zeit abgearbeitet werden können, ohne dass es zu einem Stau kommt. Allgemein können bei Warteschlangen folgende Fragen von Interesse sein:

- Stabilität des Systems?
- Wie müssen die Parameter gewählt werden, damit kein Stau entsteht?
- Wie viele Kunden befinden sich zu einer gegebenen Zeit in der Schlange?
- Wie lange wartet ein Kunde im Mittel?
- Weitere Besonderheiten im (Langzeit-)Verhalten?

In diesem Kapitel werden wir uns einer Klasse von Warteschlangenmodellen in stetiger Zeit widmen. Im Gegensatz zum bereits behandelten Modell in diskreter Zeit müssen dabei die Verteilungen der Ankunftszeiten der Kunden und der Bearbeitungsdauer spezifiziert werden. Wir wählen dabei die Exponentialverteilung, aus Gründen, die später diskutiert werden.

### Definition 14.1 (M/M/K-Warteschlange in stetiger Zeit)

Die  $M/M/K$ –**Warteschlange in stetiger Zeit** besteht aus  $K \in \mathbb{N}$  Bedienern. Kunden kommen unabhängig voneinander im System an, die Zeiten zwischen der Ankunft zweier Kunden sind unabhängig und exponentialverteilt mit Parameter  $\lambda > 0$ . Die Kunden werden von den Servern in der Reihenfolge ihres Eintreffens bedient, sind alle Server belegt, so warten Kunden bis einer frei wird. Die Bediendauer eines Kunden ist exponentialverteilt mit Parameter  $\mu > 0$ , unabhängig für verschiedene Kunden und unabhängig von der Ankunftszeit. Nach Ende der Bedienung verlässt ein Kunde das System. ♦

Die  $M$  in der Bezeichnung der Warteschlange stehen für “Markov” und werden später erläutert werden.

Zur formalen Konstruktion der Warteschlange kann folgendermaßen vorgegangen werden: Sei  $(T_i)_{i \in \mathbb{N}}$  eine Folge von unabhängigen, exponentialverteilten Zufallsvariablen mit Parameter  $\lambda$ , welche die Zeiten zwischen der Ankunft zweier Kunden bezeichnen. Sei weiter  $(S_i)_{i \in \mathbb{N}}$  eine Folge von unabhängigen, exponentialverteilten Zufallsvariablen mit

Parameter  $\mu$ , unabhängig von den  $T_i$ , die Bedienzeiten. Wir schreiben  $M_t$  für die (zufällige) Anzahl Kunden im System zur Zeit  $t \in [0, \infty[$ , und definieren

$$P_n(t) := \mathbb{P}(M_t = n).$$

### Definition 14.2 (Stabilität)

Wir sagen, die Warteschlange sei **stabil**, wenn

$$\pi_n := \lim_{t \rightarrow \infty} P_n(t)$$

für alle  $n \in \mathbb{N}_0$  existiert und in  $]0, 1[$  liegt. ♦

Stabilität bedeutet nach dieser Definition also, dass sich auf lange Sicht die Anzahl Kunden im System einem *Gleichgewichtszustand* annähert. Dieser Gleichgewichtszustand spielt die gleiche Rolle wie die invariante Verteilung einer Markov-Kette aus Kapitel 11. Instabilität würde hier bedeuten, dass auf lange Sicht mehr Kunden ins System strömen als abgearbeitet werden, so dass die  $P_n$  für kleine  $n$  auf lange Sicht gegen 0 konvergieren.

Zur Untersuchung der Stabilität ist die zeitliche Veränderung der Anzahl Kunden, von Interesse, d.h.  $P'_n(t)$ ,  $t \in [0, \infty[$ ,  $n \in \mathbb{N}$ . Wir betrachten zuerst nur den Fall  $K = 1$ .

### Satz 14.1

Für die  $M/M/1$ -Warteschlangen gelten, mit obiger Notation, für  $t \geq 0$  und  $n \in \mathbb{N}_0$ ,

$$P'_0(t) = -\lambda P_0(t) + \mu P_1(t)$$

und

$$P'_n(t) = -(\lambda + \mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t), \quad n \geq 1.$$

Wir verzichten auf einen formalen Beweis. Anschaulich beruht dieser darauf, dass innerhalb eines kurzen Zeitraums ein neuer Kunde dazukommt, oder ein Kunde das System verlässt, oder nichts passiert. Die Wahrscheinlichkeit dass mehrere Ereignisse innerhalb eines kurzen Zeitraums geschehen ist gering. Mit Hilfe der Gedächtnislosigkeit (Satz 14.3) sowie der Definition der Ableitung kann dies präzise gemacht werden, siehe auch ?, Abschnitt 8.6.1.

### Satz 14.2 (Stabilität der $M/M/1$ -Warteschlange)

Falls im  $M/M/1$ -Warteschlangenmodell  $\lambda < \mu$  ist, so existiert  $\pi_n := \lim_{t \rightarrow \infty} P_n(t)$  für alle  $n \in \mathbb{N}_0$ , und es gilt

$$\pi_0 = 1 - \frac{\lambda}{\mu} \quad \text{und} \quad \pi_n = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n, \quad n \geq 1.$$

**Beweis 14.1** Um ein Gleichgewicht zu erhalten, muss die zeitliche Veränderung von  $P_n(t)$  auf lange Sicht 0 werden. Somit ist eine Bedingung für Gleichgewicht  $\pi$  gegeben durch  $P'_n(t) = 0$  für alle  $n \in \mathbb{N}$ . Setzt man diese Bedingung in das Ergebnis von Lemma 14.1 ein, so erhält man das lineare Gleichungssystem für  $\pi_i$ ,  $i \in \mathbb{N}_0$ ,

$$\begin{cases} -\lambda\pi_0(t) + \mu\pi_1(t) = 0, \\ -(\lambda + \mu)\pi_n(t) + \lambda\pi_{n-1}(t) + \mu\pi_{n+1}(t) = 0, \quad n \geq 1. \end{cases} \quad (14.1)$$

Dieses besteht zwar aus unendlich vielen Gleichungen, es kann jedoch iterativ gelöst werden, unter der Zusatzbedingung  $\sum_{i=0}^{\infty} \pi_i = 1$ , und man erhält

$$\pi_0 = 1 - \frac{\lambda}{\mu} \quad \text{und} \quad \pi_n = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n, n \geq 1. \quad (14.2)$$

□

Intuitiv kann man die Bedingung, dass für ein Gleichgewicht  $\lambda < \mu$  sein muss, leicht verstehen als die Forderung, dass pro Zeiteinheit im Mittel mehr Kunden das System verlassen als betreten.

Bei genauerer Betrachtung fällt auf, dass die Gleichgewichtsverteilung  $\pi$  eine geometrische Verteilung ist. Damit kann man weitere interessante Größen berechnen.

### Beispiel 14.1 (Erwartete Anzahl Kunden)

Bezeichne mit  $M_t$  die Anzahl Kunden, die sich zu einem gegebenen Zeitpunkt  $t$  in der Warteschlange befinden. Ist die Schlange im Gleichgewicht, so gilt

$$\begin{aligned} \mathbb{E}[M_t] &= \sum_{k=0}^{\infty} k \cdot \pi_k = \sum_{k=0}^{\infty} k \cdot \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^k \\ &= \frac{\lambda}{\mu} \sum_{k=0}^{\infty} k \cdot \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^{k-1} = \frac{\lambda}{\mu} \frac{1}{1 - \lambda/\mu} \\ &= \frac{\lambda}{\mu - \lambda}, \end{aligned}$$

wobei wir benutzt haben, dass die letzte Summe der Erwartungswert einer geometrischen Verteilung mit Parameter  $1 - \lambda/\mu$  ist, von der wir den Erwartungswert  $1/(1 - \lambda/\mu)$  aus Kapitel 5 kennen. ■

Für mehr als einen Bediener ( $K \geq 2$ ) kann auf ähnliche Weise, jedoch mit mehr Aufwand, bewiesen werden, dass im Fall  $\lambda < K\mu$  ein Gleichgewicht existiert, dieses erfüllt

$$\pi_n = \begin{cases} \pi_0 \cdot \left(\frac{\lambda}{\mu}\right)^n \cdot \frac{1}{n!} & \text{falls } n \leq K, \\ \pi_0 \cdot \left(\frac{\lambda}{\mu}\right)^n \cdot \frac{1}{K!} K^{n-K} & \text{sonst.} \end{cases}$$

Dabei muss  $\pi_0$  so gewählt werden, dass  $\sum_{i=0}^{\infty} \pi_i = 1$  ist. Mehr dazu finden Sie z.B. in ?.

## 14.2 Eigenschaften der Exponentialverteilung

Bei der Definition der  $M/M/K$ -Warteschlange haben wir als Verteilung der Ankunfts- und Bedienzeiten die Exponentialverteilung gewählt. Der Grund dafür ist eine spezielle Eigenschaft dieser Verteilung, der Gedächtnislosigkeit. Wir erinnern zunächst daran (vgl. Kapitel 6), dass die Dichte der Exponentialverteilung mit Parameter  $\lambda$  gegeben ist durch

$$f(t) = \lambda \cdot e^{-\lambda t}, t \geq 0.$$



Der Erwartungswert einer  $\text{Exp}(\lambda)$ -verteilten Zufallsvariablen ist  $1/\lambda$ . Eine wichtige Eigenschaft der Exponentialverteilung ist folgende (vergleiche dazu Aufgabe 6.8):

**Satz 14.3 (Gedächtnislosigkeit)**

Sei  $X \sim \text{Exp}(\lambda)$ . Dann gilt für alle  $s, t \geq 0$

$$\mathbb{P}(X \geq t + s \mid X \geq t) = \mathbb{P}(X \geq s).$$

Im Kontext von Warteschlangen kann man diese Aussage folgendermaßen interpretieren:  $X$  bezeichne die Wartezeit bis zum Eintreffen des ersten Kunden. Nach Voraussetzung gilt also  $X \sim \text{Exp}(\lambda)$ . Die Gedächtnislosigkeit besagt nun Folgendes: Ist bis zum Zeitpunkt  $t$  kein Kunde gekommen, so hat die Wartezeit *ab diesem Zeitpunkt* wiederum die Verteilung  $\text{Exp}(\lambda)$ . Die Exponentialverteilung “vergisst” also, wie lange man schon gewartet hat. In anderen Worten ist die Verteilung der “Zukunft” (wie lange muss man noch warten) *unabhängig* von der “Vergangenheit” (wie lange hat man schon gewartet). Das entspricht nun aber genau der Eigenschaft, die man von Markov-Ketten fordert (vgl. Kapitel 11). Mit Hilfe der Exponentialverteilung können deshalb Markov-Ketten konstruiert werden, bei denen die Sprünge nicht zu festen Zeiten  $n = 1, 2, \dots$ , sondern zu zufälligen Zeiten stattfinden. Eine Konstruktion wird weiter unten skizziert werden. Die  $M/M/K$ -Warteschlange ist ein Beispiel einer solchen Markov-Kette. Der Gleichgewichtszustand aus Satz 14.2 entspricht dabei genau der invarianten Verteilung aus Kapitel 11, vgl. Satz 11.7.

**Beispiel 14.2 (Erwartete Zeit, die ein Kunde im System verbringt)**

Wir betrachten eine  $M/M/1$ -Warteschlange, und bezeichnen mit  $Z$  die Zeit, die ein zufällig ausgewählter Kunde im System verbringt. Dabei interessieren wir uns für den Erwartungswert von  $Z$ .

Bezeichne mit  $A_k$  das Ereignis, dass sich bei Ankunft des betreffenden Kunden genau  $k$  andere Kunden bereits im System befinden. Es gilt nach der Formel von der Gesamtwahrscheinlichkeit

$$\mathbb{E}[Z] = \sum_{k=0}^{\infty} \mathbb{E}[Z \mid A_k] \cdot \mathbb{P}(A_k).$$

Falls die Warteschlange im Gleichgewichtszustand ist, gilt  $\mathbb{P}(A_k) = \pi_k$ , und  $\pi_k$  kennen wir aus Satz 14.2. Für den bedingten Erwartungswert gilt

$$\mathbb{E}[Z \mid A_k] = \frac{k+1}{\mu},$$

denn bevor der Kunde das System verlassen kann, müssen die  $k$  Kunden vor ihm und er selbst bedient werden. Jeder dieser Bedienvorgänge dauert im Mittel  $1/\mu$ , auch derjenige des Kunden, der bereits bedient wird wenn der neue Kunde ins System kommt. Dies liegt an der Gedächtnislosigkeit. Somit erhalten wir

$$\begin{aligned} \mathbb{E}[Z] &= \sum_{k=0}^{\infty} \mathbb{E}[Z \mid A_k] \cdot \mathbb{P}(A_k) = \sum_{k=0}^{\infty} \frac{k+1}{\mu} \pi_k = \frac{1}{\mu} \sum_{k=0}^{\infty} \pi_k + \frac{1}{\mu} \sum_{k=0}^{\infty} k \cdot \pi_k \\ &= \frac{1}{\mu} + \frac{1}{\mu} \sum_{k=0}^{\infty} k \cdot \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^k = \frac{1}{\mu} + \frac{1}{\mu} \frac{\lambda}{\mu - \lambda} = \frac{1}{\mu - \lambda}. \end{aligned}$$

Dabei haben wir denselben Trick wie in Beispiel 14.1 benutzt. ■

### Beispiel 14.3 (Wartezeitenparadox)

■

Zur Simulation von Warteschlangen in stetiger Zeit sowie allgemein bei der Konstruktion von Markov-Ketten in stetiger Zeit wird eine weitere Eigenschaft der Exponentialverteilung benötigt.

### Satz 14.4 (Minimum von exponentialverteilten Zufallsvariablen)

Seien  $X_1, \dots, X_n$  unabhängige, exponentialverteilte Zufallsvariablen mit Parametern  $\lambda_1, \dots, \lambda_n$ . Dann ist die Zufallsvariable  $Z := \min(X_1, \dots, X_n)$  ebenfalls exponentialverteilt, und hat Parameter  $\sum_{i=1}^n \lambda_i$ . Außerdem gilt

$$\mathbb{P}(Z = X_j) = \frac{\lambda_j}{\sum_{i=1}^n \lambda_i}.$$

Für einen Beweis siehe z.B. ? Lemma 8.5.

### Beispiel 14.4 (Zeit bis zum ersten Ereignis)

In der  $M/M/K$ -Warteschlange untersuche folgende Frage: Wenn man zu einem beliebigen Zeitpunkt  $t$  das System betrachtet, wie lange muss man warten, bis “etwas geschieht”, d.h. bis entweder ein neuer Kunde ins System kommt, oder einer der vorhandenen das System verlässt? Dies hängt natürlich davon ab, wie viele Kunden zur Zeit  $t$  im System sind. Sind dies  $i$  Kunden, welche alle gleichzeitig bedient werden (also  $i < K$ ), dann gehört zu jedem dieser  $i$  Kunden eine exponentialverteilte Zufallsvariable  $S_j, j = 1, \dots, i$  mit Rate  $\mu$ , welche angibt, wie lange es noch dauert, bis er das System verlässt. Wegen der Gedächtnislosigkeit ist dies zu jedem Zeitpunkt  $t$  der Fall. Ausserdem gibt es eine weitere exponentialverteilte Zufallsvariable  $T$  mit Parameter  $\lambda$ , welche angibt, wann der nächste Kunde ins System eintritt. Die Zufallsvariable  $W$  bezeichne nun den ersten Zeitpunkt ab dem aktuellen Moment  $t$ , zu dem etwas passiert, also ein Kunde das System verlässt oder ein neuer reinkommt. Damit ist

$$W = \min\{S_1, \dots, S_i, T\}$$

das Minimum dieser Wartezeiten. Somit ist nach Satz 14.4  $W$  wiederum exponentialverteilt mit Parameter  $i\mu + \lambda$ , und die Wahrscheinlichkeit dafür, dass zum Zeitpunkt  $W$  ein Kunde neu in das System kommt, ist nach demselben Satz gegeben durch

$$\mathbb{P}(W = T) = \lambda/(i\mu + \lambda).$$

Im Fall dass  $i > K$  ist, geht man analog vor, wobei man berücksichtigt, dass nur einer der  $K$  Kunden, welche gerade bedient werden, das System verlassen können, man hat also die Wartezeiten  $T_1, \dots, T_K$  und  $S$  wie oben.  $W$  ist dann exponentialverteilt mit Parameter  $K\mu + \lambda$ , und  $\mathbb{P}(W = T) = \lambda/(K\mu + \lambda)$ . ■

Wie bei vielen komplexen stochastischen Systemen kann eine Simulation aufschlussreich sein. Dank der Eigenschaften der Exponentialverteilung kann beispielsweise die Anzahl Kunden  $M_t$  im System zum Zeitpunkt  $t$  simuliert werden. Die Überlegungen aus Beispiel 14.4 ergeben den folgenden Algorithmus.

**Algorithmus: M/M/K-Warteschlange**

Vorgegeben: Parameter  $\lambda > 0, \mu > 0$ . Startwert  $M_0$  (Anzahl Kunden am Anfang, z. B.  $M_0 = 0$ , Endzeitpunkt  $T_{max}$ )

1. Falls aktuell  $M_t = i, t < T_{max}$ :

- Falls  $i = 0$  : Simuliere  $W \sim \text{Exp}(\mu)$ . Setze  $M_s = i$  für  $s \in [t, t + W[$  und  $M_{t+W} = 1$ .
- Falls  $1 \leq i \leq K$  : Simuliere  $W \sim \text{Exp}(i \cdot \mu + \lambda)$  sowie  $V$  mit  $\mathbb{P}(V = 1) = \frac{\lambda}{\lambda + i \cdot \mu}, \mathbb{P}(V = -1) = 1 - \frac{\lambda}{\lambda + i \cdot \mu}$ . Setze

$$M_s = i \quad \text{für } s \in [t, t + W[, \quad \text{und} \quad M_{t+W} = X_t + V$$

- Falls  $i > K$  : Simuliere  $W \sim \text{Exp}(K \cdot \mu + \lambda)$  sowie  $V$  mit  $\mathbb{P}(V = 1) = \frac{\lambda}{\lambda + K \cdot \mu}, \mathbb{P}(V = -1) = 1 - \frac{\lambda}{\lambda + K \cdot \mu}$ . Setze

$$M_s = i \quad \text{für } s \in [t, t + W[, \quad \text{und} \quad M_{t+W} = M_t + V.$$

2. Wiederhole mit  $t = t + W$ .

3. Falls  $t \geq T_{max}$  : Stop.

Dabei ist in der Simulation  $W$  wie oben die Wartezeit bis zum nächsten Ereignis, und  $V$  entspricht der Veränderung in der Anzahl Kunden zur Zeit  $W$ .

In diesem Algorithmus müssen wiederholt Zufallsvariablen mit einer vorgegebenen Verteilung simuliert werden. Wie man dies machen kann, werden wir in Kapitel 15 sehen. Programme wie R haben diese Möglichkeit bereits eingebaut.

## 14.3 Markov-Ketten in stetiger Zeit

In diesem Abschnitt erläutern wir die Verallgemeinerung von Algorithmus ?? zur Konstruktion von Markov-Ketten in stetiger Zeit (auf endlichem Zustandsraum). Wir verzichten hier auf Beweise, und verweisen stattdessen auf ?. Zur Erinnerung: Eine Markov-Kette in diskreter Zeit konnte man sich so vorstellen, dass zu festen Zeitpunkten  $1, 2, 3, \dots \in \mathbb{N}$  die Kette  $X_n, n \in \mathbb{N}$  mit Wahrscheinlichkeit  $p_{i,j}$  aus dem Zustand  $X_n = i$  in den Zustand  $X_{n+1} = j$  springt. Nun soll die Kette nicht zu festen, sondern zu zufälligen Zeiten springen, wobei die Wartezeit bis zum nächsten Sprung exponentialverteilt sein soll, mit einem Parameter, welcher abhängig vom aktuellen Zustand der Kette ist. Wohin die Kette springt, ist weiterhin durch eine stochastische Matrix gegeben. Konkret:

### Markov-Kette in stetiger Zeit

Sei  $S$  eine endliche Menge, sei  $P = (p_{i,j})_{i,j \in S}$  eine stochastische Matrix mit  $p_{i,i} = 0$  für alle  $i \in S$ , und  $\nu_i, i \in S$  eine Startverteilung auf  $S$ . Seien weiter  $\lambda_i > 0, i \in S$  gegeben.

1. Start in  $X_0$ , mit Verteilung  $\nu$ .
2. Falls  $X_0 = i$  ist, warte eine exponentialverteilte Zeit  $T$  mit Parameter  $\lambda_i$ , und springe dann in den nächsten Zustand  $j$  mit Wahrscheinlichkeit  $p_{i,j}$ .
3. Wiederhole 2. ausgehend vom aktuellen Zustand  $X_T$ .

Dieses Vorgehen liefert einen stochastischen Prozess  $(X_t)_{t \geq 0}$ , wobei der Zeitindex  $t$  aus den nichtnegativen reellen Zahlen stammt.  $(X_t)_{t \geq 0}$  ist eine **Markov-Kette in stetiger Zeit**, mit Anfangsverteilung  $\nu$ , Sprungraten  $\lambda_i, i \in S$ , und Übergangsmatrix  $P$ .

#### Satz 14.5 (Markov-Eigenschaft)

Für eine Markov-Kette  $(X_t)_{t \geq 0}$  in stetiger Zeit, konstruiert wie in obigem Vorgehen gilt für alle  $n \in \mathbb{N}, 0 \leq t_1 \leq \dots \leq t_n < \infty$  und  $i_1, \dots, i_n \in S$

$$\mathbb{P}(X_{t_n} = i_n \mid X_{t_1} = i_1, \dots, X_{t_{n-1}} = i_{n-1}) = \mathbb{P}(X_{t_n} = i_n \mid X_{t_{n-1}} = i_{n-1}).$$

Wir führen den Beweis nicht durch, er beruht jedoch entscheidend auf der Gedächtnislosigkeit der Exponentialverteilung. Diese impliziert nämlich, dass es keine Rolle spielt, wie lange der Prozess bereits in einem bestimmten Zustand  $i$  wartet, die Zeit bis zum nächsten Sprung ist in jedem Fall einfach durch eine exponentialverteilte Zufallsvariable mit Parameter  $\lambda_i$  gegeben. Im Allgemeinen nimmt man die Eigenschaft 14.5 als Definition einer Markov-Kette, was etwas allgemeiner ist als die obige Konstruktion und unendliche Zustandsräume zulässt. In der Praxis ist der Algorithmus jedoch handlicher.

Im Gegensatz zu Markov-Ketten in diskreter Zeit wird also zur Beschreibung neben der Startverteilung  $\nu$  und der Übergangsmatrix  $P$  auch die Kenntnis der Sprungraten  $\lambda_i, i \in S$  für jeden Zustand benötigt. Die Sprungraten und die Übergangsmatrix können in der sogenannten Q-Matrix zusammengefasst werden.

#### Definition 14.3 (Q-Matrix)

Sei  $(X_t)_{t \geq 0}$  eine Markov-Kette in stetiger Zeit auf endlichem Zustandsraum  $S$ , mit Übergangsmatrix  $P$  und Sprungraten  $\lambda_i > 0, i \in S$ . Dann heißt die Matrix  $Q = (q_{i,j})_{i,j \in S}$  definiert durch

$$q_{i,j} = \begin{cases} \lambda_i p_{i,j} & \text{falls } i \neq j, \\ \lambda_i (p_{i,i} - 1) & \text{falls } i = j \end{cases}$$

**Q-Matrix** (oder **Generator**) der Markov-Kette. ◆

Auf Grund der Eigenschaften einer Übergangsmatrix hat eine  $Q$ -Matrix auf der Diagonalen negative und ansonsten positive Einträge, welche sich zeilenweise zu 0 summieren. Die  $Q$ -Matrix enthält viele relevante Informationen über die Markov-Kette. Beispielsweise kann die Verteilung der Kette zu einem festen Zeitpunkt daraus berechnet werden.

#### Satz 14.6

Sei  $(X_t)_{t \geq 0}$  eine Markov-Kette in stetiger Zeit auf endlichem Zustandsraum  $S$  mit  $Q$ -Matrix  $Q = (q_{i,j})_{i,j \in S}$ . Dann gilt

$$\mathbb{P}(X_t = j \mid X_0 = i) = (e^{tQ})_{i,j}$$

für alle  $i, j \in S, t \geq 0$ . Dabei ist  $e^{tQ} := \sum_{k=0}^{\infty} Q^k \frac{t^k}{k!}$  das Matrix-Exponential.

Bei der Berechnung von  $e^{tQ}$  müssen die Regeln der Matrizenrechnung beachtet werden:  $Q^k$  ist die  $k$ -te Potenz der Matrix  $Q$ , wird also durch  $k$ -fache Matrixmultiplikation berechnet. Die Einträge  $(e^{tQ})_{i,j}$  sind dann gegeben durch  $(e^{tQ})_{i,j} = \sum_{k=0}^{\infty} Q^k(i,j) \frac{t^k}{k!}$ . Beachte, dass im Allgemeinen  $(e^{tQ})_{i,j} \neq e^{tq_{i,j}}$  gilt.

Da die Berechnung des Matrix-Exponential für praktische Zwecke schwierig ist, wird zur Simulation von Markov-Ketten die Konstruktion über die Wartezeiten verwendet, und der Algorithmus für Warteschlangen entsprechend verallgemeinert (vgl. Aufgabe 14.2). Um den Algorithmus anwenden zu können, müssen die Sprungraten sowie die Übergangsmatrix bekannt sein. Kennt man stattdessen die  $Q$ -Matrix, so gibt es meistens mehrere Kombinationen von Sprungraten und Übergangsmatrix, die kompatibel mit der  $Q$ -Matrix sind. In den Übungsaufgaben wird diese Fragestellung untersucht.

Die  $Q$ -Matrix ist insbesondere bei der Bestimmung von invarianten Verteilungen nützlich.

#### Definition 14.4 (Invariante Verteilung)

Sei  $(X_t)_{t \geq 0}$  eine Markov-Kette in stetiger Zeit auf endlichem Zustandsraum  $S$  mit  $Q$ -Matrix  $Q$ . Eine *invariante Verteilung* von  $(X_t)_{t \geq 0}$  ist ein (Spalten-)Vektor  $\pi = (\pi_i)_{i \in S}$  mit  $\pi_i \geq 0$  für alle  $i \in S$ ,  $\sum_{i \in S} \pi_i = 1$ , und

$$\pi^T Q = 0.$$



Auf einem endlichen Zustandsraum existiert immer mindestens eine invariante Verteilung. Falls die Übergangsmatrix  $P$  irreduzibel im Sinn von Kapitel 11 ist, so ist die invariante Verteilung eindeutig. Sie beschreibt wie im Fall von Markov-Ketten in diskreter Zeit das Langzeitverhalten:

#### Satz 14.7

Sei  $(X_t)_{t \geq 0}$  eine Markov-Kette in stetiger Zeit auf endlichem Zustandsraum  $S$  mit eindeutiger invarianter Verteilung  $\pi$ . Dann gilt für jede Startverteilung  $\nu$

$$\lim_{t \rightarrow \infty} \mathbb{P}(X_t = i) = \pi_i, \quad i \in S.$$

### Zusammenfassung

Mit Hilfe der speziellen Eigenschaften der Exponentialverteilung, insbesondere der Gedächtnislosigkeit, können Warteschlangenmodelle in stetiger Zeit aufgestellt und auf Stabilität untersucht werden. Allgemeiner können Markov-Ketten in stetiger Zeit konstruiert und analysiert werden. Diese werden entweder durch die Angabe von Übergangsmatrix und Sprungraten charakterisiert, oder durch ihrer Q-Matrix, welche auch die Berechnung von invarianten Verteilungen ermöglicht.

### Aufgaben

**14.1** Überzeugen Sie sich durch Nachrechnen davon, dass (14.2) das Gleichungssystem (14.1) löst, und  $\sum_{i=1}^{\infty} \pi_i = 1$  erfüllt.

**14.2** Verallgemeinern Sie den Algorithmus für Warteschlangen zu einem Simulationsalgorithmus für Markov-Ketten in stetiger Zeit auf endlichem Zustandsraum.

**14.3** Simulieren Sie eine Markov-Kette mit Zustandsraum  $S = \{1, \dots, 5\}$ , Startverteilung  $\nu(1) = 1$ , Sprungraten  $\lambda_1 = \lambda_2 = \lambda_3 = 1, \lambda_4 = 2, \lambda_5 = 3$  und Übergangsmatrix

$$P = \begin{bmatrix} 0 & 0.2 & 0.5 & 0 & 0.3 \\ 0.1 & 0 & 0.1 & 0.1 & 0.7 \\ 0.8 & 0 & 0 & 0.2 & 0 \\ 0.5 & 0.2 & 0.2 & 0 & 0.1 \\ 0.4 & 0.3 & 0.2 & 0.1 & 0 \end{bmatrix}.$$

**14.4** Berechnen Sie zur Markov-Kette aus Aufgabe 15.1 die Q-Matrix.

**14.5** Sei eine Q-Matrix

$$Q = \begin{bmatrix} -2 & 1 & 1 \\ 4 & -4 & 0 \\ 2 & 1 & -3 \end{bmatrix}$$

gegeben. Finden Sie sämtliche möglichen Kombinationen von Sprungraten  $\lambda_i > 0, i \in \{1, 2, 3\}$  und stochastischen Matrizen  $P$  auf  $\{1, 2, 3\}$ , welche auf die angegebene Q-Matrix führen.

**14.6** Es sei  $S$  eine endliche Menge, und  $Q = (q_{i,j})_{i,j \in S}$  eine Matrix, für welche  $q_{ii} < 0, q_{i,j} \geq 0$  für  $j \neq i$ , und  $\sum_{j \in S} q_{i,j} = 0, i \in S$  gelten. Zeigen Sie: Dann ist  $\bar{P}$  mit  $\bar{p}_{i,j} = \frac{q_{i,j}}{-q_{i,i}}, j \neq i$ , und  $\bar{p}_{i,i} = 0$  für alle  $i \in S$  eine stochastische Matrix. Insbesondere kann man also von einer solchen Q-Matrix ausgehend  $\lambda_i = -q_{i,i}$  und  $P = \bar{P}$  wählen, und erhält so eine Markov-Kette in stetiger Zeit.

**14.7** In der Situation von Aufgabe 14.6, geben Sie eine allgemeine Formel für sämtliche möglichen Kombinationen von Sprungraten  $\lambda_i > 0, i \in S$  und Übergangsmatrizen  $P$  auf  $S$  an, welche auf die gebene  $Q$ -Matrix führen.

**14.8** Berechnen Sie die invariante Verteilung für die Markov-Kette aus 14.5.

**14.9** Sei  $(X_t)_{t \geq 0}$  eine Markov-Kette in stetiger Zeit auf endlichem Zustandsraum  $S$ , mit Übergangsmatrix  $P$   $p_{i,i} = 0$  für alle  $i \in S$ , und Sprungraten  $\lambda_i > 0, i \in S$ . Sei  $Q$  die zugehörige  $Q$ -Matrix. Zeigen Sie: Die Menge der invarianten Verteilungen von  $(X_t)_{t \geq 0}$  stimmt überein mit der Menge der invarianten Verteilungen der Markov-Kette in diskreter Zeit, welche durch die Übergangsmatrix  $P$  beschrieben wird.





# 15 Simulation von Zufallsvariablen, Monte-Carlo Methode

---

## Übersicht

15.1 Simulation von Zufallszahlen, Pseudozufallszahlen .....	203
15.2 Monte-Carlo Simulation .....	206

---

In diesem Kapitel werden einige Aspekte der Simulation von Zufallsvariablen beleuchtet. Dabei wird gezeigt, wie mit elementaren Mitteln aus gleichverteilten Zufallsvariablen auf  $[0, 1]$  wichtige vorgegebene Verteilungen erzeugt werden können. Weiter werden Grundprinzipien der Monte-Carlo Simulation eingeführt und an Beispielen erläutert.

### Lernziele

- Simulationsmethoden für Zufallsvariablen kennen und anwenden können
- Die Grundprinzipien von Monte Carlo Methoden kennen

### Vorkenntnisse

Es werden Kenntnisse aus den Bereichen Zufallsvariablen, Verteilungen und dem Gesetz der großen Zahlen benötigt.

## 15.1 Simulation von Zufallszahlen, Pseudozufallszahlen

Das Ziel bei der Simulation von Zufallsvariablen besteht darin, eine Folge von Zahlen zu erzeugen, welche sich in einem geeigneten Sinne verhalten wie Realisierungen von (unabhängigen) Zufallsvariablen einer bestimmten Verteilung. In anderen Worten, das

Ergebnis der Simulation sollte sich z.B. mit statistischen Methoden nicht von einer Folge von tatsächlich zufälligen Zahlen unterscheiden lassen.

Wir beginnen mit einem theoretisch relevanten Ergebnis.

**Satz 15.1 (Zufallsvariable mit vorgegebener Verteilung)**

Sei  $U$  auf  $[0, 1]$  gleichverteilt. Sei  $F$  eine invertierbare Verteilungsfunktion. Dann ist

$$Y := F^{-1}(U)$$

eine Zufallsvariable mit Verteilungsfunktion  $F$ .

**Beweis 15.1** Es gilt

$$\mathbb{P}(Y \leq y) = \mathbb{P}(F^{-1}(U) \leq y) = \mathbb{P}(U \leq F(y)) = F(y),$$

nach Beispiel 6.3. Somit ist  $F$  die Verteilungsfunktion von  $Y$ . □

Falls  $F$  nicht invertierbar ist, kann man stattdessen mit der sogenannten *verallgemeinerten Inversen*  $F^{-1}(x) := \inf\{y : F(y) \geq x\}$  arbeiten. Wegen der Monotonie und der Rechtsstetigkeit der Verteilungsfunktion gilt die Aussage genauso.

**Beispiel 15.1 (Exponentialverteilung)**

Die Exponentialverteilung mit Parameter  $\lambda$  hat die Verteilungsfunktion  $F(x) = 1 - e^{-\lambda x}$ ,  $x \geq 0$ . Invertieren führt auf  $F^{-1}(y) = \frac{-\ln(1-y)}{\lambda}$ ,  $0 \leq y \leq 1$ . Somit ist  $\frac{-\ln(1-U)}{\lambda}$  exponentialverteilt zum Parameter  $\lambda$ , falls  $U$  auf  $[0, 1]$  gleichverteilt ist. ■

Auch diskrete Verteilungen können aus der stetigen Gleichverteilung erzeugt werden.

**Beispiel 15.2 (Bernoulli- und Binomialverteilung)**

Sei  $U$  gleichverteilt auf  $[0, 1]$  und  $p \in [0, 1]$ . Dann ist  $X = 1_{\{U \leq p\}}$  Bernoulli-verteilt mit Parameter  $p$ , denn es gilt

$$\mathbb{P}(X = 1) = \mathbb{P}(U \leq p) = p.$$

Sind  $U_1, \dots, U_n$  unabhängige gleichverteilte Zufallsvariablen, so sind somit  $X_i = 1_{\{U_i \leq p\}}$  unabhängig und Bernoulli-verteilt zum Parameter  $p$ . Somit ist

$$Y = \sum_{i=1}^n X_i$$

binomialverteilt mit Parametern  $n$  und  $p$ . ■

Allgemeiner wurde dieser Ansatz zur Simulation von Verteilungen mit endlichem Wertebereich in Beispiel 3.9 verwendet.

**Beispiel 15.3 (Normalverteilung)**

Die Verteilungsfunktion der Normalverteilung ist nicht explizit anzugeben, und somit nicht invertierbar. Dennoch kann die Normalverteilung aus der uniformen Verteilung erzeugt werden, z.B. mit folgendem Verfahren: Seien  $U$  und  $V$  unabhängige, auf  $[0, 1]$  gleichverteilte Zufallsvariablen. Dann sind

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

zwei unabhängige, standardnormalverteilte Zufallsvariablen. Wir verzichten hier auf einen Beweis. Eine normalverteilte Zufallsvariable  $Z$  mit beliebigen Parametern  $\mu$  und  $\sigma^2$  erhält man mittels der Transformation  $Z = \sigma X + \mu$  aus einer standardnormalverteilten Zufallsvariablen  $X$ . Dies folgt aus Satz 6.7. ■

Man kann also die wichtigsten Zufallsvariable simulieren, sobald man eine oder mehrere auf  $[0, 1]$  gleichverteilte Zufallsvariable simulieren kann. Somit konzentrieren wir uns nun auf die Gleichverteilung.

### Definition 15.1

Ein **Pseudozufallszahlengenerator** ist ein Algorithmus, welcher ausgehend von einem Startwert eine deterministische Folge von Zahlen erzeugt, die sich wie eine Realisierung einer Folge von (auf  $[0, 1]$  gleichverteilten) Zufallsvariablen verhält. ♦

Die Überprüfung des zufälligen Verhaltens erfolgt dabei mit Hilfe von statistischen Tests, z.B.  $\chi^2$ -Test auf Gleichverteilung (vgl. Kapitel 9). Pseudozufallszahlen sind also nicht tatsächlich zufällig, sondern deterministisch, im Gegensatz zu „echten“ Zufallszahlen, welche mit Hilfe physikalischer Generatoren, z.B. durch Beobachtung radioaktiver Zerfälle, kosmisches Rauschen etc. erhalten werden. Auf solche Verfahren wird hier nicht eingegangen.

### Beispiel 15.4 (Linearer Kongruenzgenerator)

Der lineare Kongruenzgenerator ist ein Beispiel für einen Algorithmus, welcher Pseudozufallszahlen erzeugt. Algorithmus:

Wähle Parameter  $m \in \mathbb{N}$ ,  $a \in \mathbb{Z} \setminus \{0\}$ ,  $c \in \mathbb{Z}$ , und einen Startwert  $x_0 \in \{0, \dots, m\}$ .

■ Definiere rekursiv

$$x_n := (ax_{n-1} + c) \bmod m$$

■ Definiere für  $n \in \mathbb{N}$

$$u_n := \frac{x_n}{m}$$

Die Bezeichnung  $\bmod m$  bedeutet dabei, dass „modulo  $m$ “ gerechnet wird, d.h.  $ax_{n-1} + c \bmod m$  ist gleich dem Rest, der bei der Division von  $ax_{n-1} + c$  durch  $m$  bleibt. Insbesondere ist  $ax_{n-1} + c \bmod m$  genau dann gleich dem tatsächlichen Wert  $ax_{n-1} + c$ , wenn dieser kleiner als  $m$  ist.

Man kann zeigen, dass die so erzeugte Folge  $(u_n)_{n \in \mathbb{N}_0}$  eine Folge von Pseudozufallszahlen (uniform auf  $[0, 1]$ ) ist. Nach Konstruktion können so höchstens  $m$  verschiedene Zahlen erzeugt werden, und die erzeugte Folge ist periodisch. Entsprechend sollte  $m$  möglichst groß gewählt werden, auf jeden Fall deutlich größer als die Anzahl zu erzeugender Zufallszahlen. Die Wahl von  $a$  nach kann mit zahlentheoretischen Überlegungen optimiert werden, auf die wir hier nicht eingehen. Die Wahl des Startwerts kann große Auswirkung auf die erzeugte Folge haben, deshalb ist die Methode geeignet, um Zufallszahlen zu erzeugen. Reproduzierbarkeit der Methode ist durch Wahl des Startwerts gegeben. ■

Das Software-Paket R bietet die Auswahl aus verschiedenen Pseudozufallszahlen-Generatoren, der Standard ist der sogenannte *Mersenne-Twister* (welcher ebenfalls ein rekursives Verfahren benutzt).

## 15.2 Monte-Carlo Simulation

Beispiel 1.2 aus dem ersten Kapitel betrachtete die Frage, mit welcher Wahrscheinlichkeit ein zufällig ausgewählter Punkt aus einem Quadrat im Innern einer dort eingezeichneten Figur liegt. Dies wurde später formalisiert: Wenn  $Q$  das Quadrat bezeichnet, soll also  $X \sim \text{Unif}(Q)$  sein, d.h.

$$\mathbb{P}(X \in A) = \frac{\text{Fläche von } A}{\text{Fläche von } Q}$$

Wählt man mit einer geeigneten Methode eine zufällige Folge  $X_1, X_2, X_3, \dots$  von Punkten aus  $Q$ , d.h.  $X_i, i \in \mathbb{N}$  sind auf  $Q$  gleichverteilt und unabhängig, so kann nach dem Gesetz der großen Zahlen (Theorem 7.1) die Wahrscheinlichkeit für  $X \in A$  durch

$$\mathbb{P}(X \in A) \approx \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \in A\}}$$

approximiert werden. Das Vorgehen zur Auswahl der Punkte kann also als Zufallszahlen-generator verstanden werden. Das wiederholte Ausführen eines Zufallsexperiments zur Bestimmung der entsprechenden Wahrscheinlichkeiten ist das Grundprinzip für sogenannte *Monte-Carlo-Methoden*, bzw. *Monte-Carlo-Simulation*.

### Definition 15.2 (Monte-Carlo Schätzer für eine Wahrscheinlichkeit)

Sei  $U_n, n \in \mathbb{N}$  eine Folge von unabhängigen, auf  $[0, 1]$  gleichverteilten Zufallsvariablen. Sei  $A \subset [0, 1]$ . Dann ist

$$\hat{\mathbb{P}}(A) := \frac{1}{n} \sum_{i=1}^n 1_{\{U_i \in A\}}$$

ein **Monte-Carlo-Schätzer** für die Wahrscheinlichkeit  $\mathbb{P}(U_1 \in A)$ . Analog ist

$$\hat{\mathbb{E}}[h(U)] := \frac{1}{n} \sum_{i=1}^n h(U_i)$$

ein Monte-Carlo-Schätzer für  $\mathbb{E}[h(U_1)]$ . ♦

Zur konkreten Berechnung setzt man für die  $U_i$  beispielsweise geeignete Pseudozufallszahlen  $u_i, i \in \mathbb{N}$ , ein. Die Methode kann verallgemeinert werden, indem man mit  $A \subset [0, 1]^n, A \subset [a, b], a, b \in \mathbb{R}$  oder  $A \subset [a_1, b_1] \times \dots \times [a_n, b_n] \subset \mathbb{R}^n$  arbeitet.

### Satz 15.2

Der Monte-Carlo Schätzer  $\hat{\mathbb{P}}(A)$  (bzw.  $\hat{\mathbb{E}}[h(U)]$ ) ist ein erwartungstreuer und konsistenter Schätzer für  $\mathbb{P}(A)$  (bzw.  $\mathbb{E}[h(U_1)]$ ).

**Beweis 15.2** Es gilt

$$\mathbb{E}[\hat{\mathbb{P}}(A)] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n 1_{\{U_i \in A\}}\right] = \frac{1}{n} \sum_{i=1}^n (1 \cdot \mathbb{P}(U_i \in A) + 0 \cdot \mathbb{P}(U_i \notin A)) = \mathbb{P}(U_1 \in A).$$

Dies zeigt die Erwartungstreue. Die Konsistenz folgt aus der Tatsache, dass mit dem Gesetz der großen Zahlen  $\frac{1}{n} \sum_{i=1}^n 1_{\{U_i \in A\}}$  gegen  $\mathbb{E}[1_{\{U_1 \in A\}}]$  konvergiert. Der Beweis für  $\hat{\mathbb{E}}[h(U)]$  verläuft analog.  $\square$

Ein Anwendungsbereich dieser Methode ist die näherungsweise Berechnung von Integralen. Der Einfachheit halber betrachten wir zuerst den Fall, dass  $h$  eine integrierbare Funktion vom Intervall  $[0, 1]$  nach  $\mathbb{R}$  ist. Sei  $U$  eine gleichverteilte Zufallsvariable auf  $[0, 1]$ , also mit Dichte  $f_U(x) = 1_{\{x \in [0, 1]\}}$ . Somit gilt

$$\int_0^1 h(x) dx = \int_{-\infty}^{\infty} h(x) f_U(x) dx = \mathbb{E}[h(U)],$$

wobei die letzte Gleichung aus Satz 6.4 folgt. Somit kann  $\hat{\mathbb{E}}[h(U)]$  aus Definition 15.2 als Näherungswert für  $\int_0^1 h(x) dx$  verwendet werden. Diese Überlegungen können wiederum verallgemeinert werden.

**Satz 15.3**

Sei  $h : [a, b] \rightarrow \mathbb{R}$  eine integrierbare Funktion. Seien  $U_i, i \in \mathbb{N}$  unabhängige, auf  $[a, b]$  gleichverteilte Zufallsvariablen. Dann ist

$$\hat{I}(h) := \frac{(b-a)}{n} \sum_{i=1}^n h(U_i)$$

ein erwartungstreuer Schätzer für das Integral  $\int_a^b h(x) dx$ . Entsprechend nennt man  $\hat{I}(h)$  einen **Monte-Carlo Schätzer** für das Integral.

**Beweis 15.3** Ist  $X$  gleichverteilt auf  $[a, b]$ , so hat  $X$  die Dichte  $f_X(t) = \frac{1}{b-a}$  für  $t \in [a, b]$  und  $f_X(t) = 0$  sonst. Somit ist

$$\int_a^b h(x) dx = (b-a) \int_a^b \frac{1}{b-a} h(x) dx = (b-a) \int_{-\infty}^{\infty} f_X(x) h(x) dx = (b-a) \mathbb{E}[h(X)].$$

Einsetzen des Monte Carlo Schätzers für  $\mathbb{E}[h(X)]$  liefert die Aussage.  $\square$

Der Beweis folgt analog zur obigen Herleitung für den Spezialfall des Intervalls  $[0, 1]$ , unter Berücksichtigung der Formel für die Dichte der mehrdimensionalen Gleichverteilung, vgl. Beispiel 6.10. Zur praktischen Verwendung dieses Satzes simuliert man eine ausreichende Anzahl unabhängiger gleichverteilter Zufallsvariablen  $U_i$ , und berechnet daraus  $\hat{I}(h)$ .

Auch mehrdimensionale Integrale können mit Hilfe dieser Methode angenähert werden.

**Satz 15.4**

Sei  $h : [a_1, b_1] \times \dots \times [a_n, b_n] \rightarrow \mathbb{R}$  eine integrierbare Funktion. Seien  $U_i, i \in \mathbb{N}$  unabhängige, auf  $[a_1, b_1] \times \dots \times [a_n, b_n]$  gleichverteilte Zufallsvariablen. Dann ist

$$\hat{I}(h) := \frac{1}{m} \prod_{i=1}^n (b_i - a_i) \sum_{i=1}^m h(U_i)$$

ein erwartungstreuer Schätzer für das Integral  $\int_{[a_1, b_1] \times \dots \times [a_n, b_n]} h(x_1, \dots, x_n) dx_1 \dots dx_n$ . Entsprechend nennt man  $\hat{I}(h)$  einen **Monte-Carlo Schätzer** für das Integral.

**Beispiel 15.5 (Monte-Carlo Schätzer für Integral)**

Zur Illustration der Methode berechnen wir den obigen Schätzer für ein einfaches Integral, welches auch explizit bestimmt werden kann. Natürlich liegt der Nutzen der Methode aber insbesondere darin, dass auch nicht explizit berechenbare Integrale damit angenähert werden kann. Sei hier  $h(x) = 3x^2$ , und somit  $\int_0^1 h(x) dx = 1$ . Mit Hilfe eines geeigneten Programms können 100 auf  $[0, 1]$  gleichverteilte Pseudozufallszahlen  $u_1, \dots, u_{100}$  generiert werden. Damit erhält man für den Monte-Carlo Schätzer

$$h(U) = \frac{1}{100} \sum_{i=1}^{100} 3u_i^2.$$

■

Ganz allgemein bezieht sich der Begriff *Monte-Carlo Schätzer*, *Monte-Carlo Simulation* oder *Monte-Carlo Methode* auf die numerische Berechnung einer unbekannten Größe mittels einer großen Zahl von unabhängigen, identisch verteilten (meist gleichverteilten) Zufallsvariablen. Über die obige Berechnung von Integralen hinaus soll dies hier an einem weiteren Beispiel illustriert werden.

**Beispiel 15.6 (MC-Schätzer im Galton-Watson-Prozess)**

Vorgegeben sei ein zufälliger Baum, entstanden aus einem Galton-Watson-Prozess. Die Anzahl der *Blätter* (=Endknoten) sei dabei unbekannt und soll geschätzt werden, ohne den ganzen, möglicherweise sehr großen, Baum untersuchen zu müssen, sondern nur eine Anzahl Teilbäume.

**Algorithmus (D. Knuth)**

Starte bei der Wurzel  $v_0$ . Bezeichne mit  $Y_1$  die Anzahl Knoten, welche direkt mit  $v_0$  verbunden sind (Nachfolgeknoten). Für  $i = 0, 1, \dots$  wiederhole

1. Von  $v_i$  aus wähle zufällig (=gleichverteilt) einen der  $Y_{i+1}$  Nachfolgeknoten aus. Dieser ist  $v_{i+1}$ .
2. Ist  $v_{i+1}$  ein Blatt: Stopp,  $T = i + 1$ . Andernfalls  $i \rightarrow i + 1$ , weiter bei 1.

Ergebnis: Berechne  $M := \prod_{i=1}^T Y_i$ .

Zur Berechnung des Monte-Carlo Schätzers wird der Algorithmus  $n$  Mal mit unabhängiger Knotenwahl ausgeführt, und die Ergebnisse  $M_i, i = 1, \dots, n$  berechnet. Dann ist

$$\frac{1}{n} \sum_{i=1}^n M_i$$

Schätzer für die Anzahl Blätter. Fasst man nämlich  $M$  als Zufallsvariable auf (als Funktion der  $Y_i$ , welche mit Ausnahme von  $Y_1$  zufällig sind, da sie von den zufällig gewählten Knoten  $v_i$  abhängen), so kann man zeigen, dass  $\mathbb{E}[M]$  die Anzahl der Blätter des Baumes ist.

Zuerst ergänzen wir den Baum durch Anhängen einzelner Äste so, so, dass alle Äste gleich lang sind, die Anzahl Blätter aber gleich bleibt, indem wir jedes Blatt, welches nicht auf der letzten Ebene des Baums liegt, mittels einzelnen Nachkommen auf die letzte Ebene verschieben.

Dies hat den Vorteil, dass  $T$  bei jedem Versuch gleich ist,  $M$  jedoch nicht verändert wird (Multiplikation mit 1 entlang der zusätzlichen Kanten). Damit können wir nun eine vollständige Induktion über die Anzahl Ebenen  $T$  durchführen. Bezeichne mit  $n_1$  die Anzahl Nachkommen der Wurzel  $v_0$ . Damit unterteilt man den Baum gleichzeitig in  $Z_1$  Teilbäume  $B_1, \dots, B_{Z_1}$  der Länge  $T - 1$ , nämlich einen Teilbaum für jeden Nachkommen der Wurzel.

Die Induktion läuft dann wie folgt:

$T = 1$  In diesem Fall ist die Anzahl Blätter gleich  $n_1 = Y_1 = M$ .

$T > 1$ : **Induktionsannahme:** Die Anzahl Blätter eines Baums der Länge  $T - 1$  ist gleich  $\mathbb{E}[M]$ , wenn  $M$  das Ergebnis des Algorithmus ist.

$T > 1$ : **Induktionsschluss:** Im ersten Schritt des Algorithmus wählt man einen der  $Z_1$  Knoten und damit einen der  $Z_1$  Teilbäume. Jeder Teilbaum wird mit Wahrscheinlichkeit  $1/Z_1$  gewählt. Bezeichne mit  $M_j$  das Ergebnis des Algorithmus, wenn dieser auf Teilbaum  $j$  abläuft. Somit ist das Ergebnis des Algorithmus auf dem ganzen Baum

$$M = Z_1 \cdot M_j,$$

und entsprechend

$$\mathbb{E}[M \mid \text{Teilbaum } j \text{ gewählt}] = N_1 \cdot \mathbb{E}[M_j].$$

Somit folgt mit der Formel von der Gesamtwahrscheinlichkeit

$$\begin{aligned} \mathbb{E}[M] &= \sum_{j=1}^{Z_1} \mathbb{P}(\text{Teilbaum } j \text{ gewählt}) \mathbb{E}[M \mid \text{Teilbaum } j \text{ gewählt}] \\ &= \sum_{j=1}^{Z_1} \frac{1}{Z_1} \cdot N_1 \cdot \mathbb{E}[M_j] = \sum_{j=1}^{Z_1} \mathbb{E}[M_j]. \end{aligned}$$

Da der im ersten Schritt gewählte Teilbaum nur  $T - 1$  Ebenen hat, kann die Induktionsannahme angewandt werden, und er hat  $\mathbb{E}[M_j]$  Blätter, und somit ist die rechte Seite der letzten Gleichungskette genau die Anzahl Blätter im ganzen Baum.



### Zusammenfassung

Zufallsvariablen mit bekannter Verteilung können in vielen wichtigen Fällen aus gleichverteilten Zufallsvariablen auf  $[0, 1]$  erzeugt werden. Letztere werden mit Hilfe von Pseudozufallszahlengeneratoren simuliert, welche Folgen von Zahlen erzeugen, die sich statistisch so verhalten, als wären sie Realisierungen von gleichverteilten Zufallsvariablen. Die Monte Carlo Methode dient dazu, schwer zu berechnende Größen mit stochastischen Überlegungen anzunähern, indem beispielsweise ein Integral als Erwartungswert aufgefasst wird, und das Gesetz der großen Zahlen angewandt. Sie können aber in unterschiedlicher Spielart auch in komplexeren Situationen angewandt werden.

### Aufgaben

**15.1** Es seien  $U_i, i \in \mathbb{N}$  unabhängige, auf  $[0, 1]$  gleichverteilte Zufallsvariablen. Finden Sie eine Funktion  $g$ , so dass  $Y_i := g(U_i)$  Pareto-verteilt zum Parameter  $\alpha > 1$  ist, d.h. die Verteilungsfunktion  $F$  von  $Y_i$  gegeben ist durch

$$F(x) = \begin{cases} 1 - x^{-\alpha}, & \text{falls } x \geq 1, \\ 0 & \text{sonst.} \end{cases}$$

**15.2** Es seien  $U_i, i \in \mathbb{N}$  unabhängige, auf  $[0, 1]$  gleichverteilte Zufallsvariablen. Es seien  $V_i := 1_{\{U_i \leq 1/4\}}$  und  $X := \min\{i : V_i = 1\}$ . Welche Verteilung haben die  $V_i$ , welche Verteilung hat  $X$ ?

**15.3** Es seien  $U_i, i \in \mathbb{N}$  unabhängige, auf  $[0, 1]$  gleichverteilte Zufallsvariablen. Geben Sie mit Hilfe der Zufallsvariablen  $U_i$  einen Monte-Carlo-Schätzer für  $\int_0^1 \pi \sin(x) dx$  an.

**15.4** Sei  $U$  auf  $[a, b]$  gleichverteilt. Zeigen Sie: Dann ist  $V := U(b - a) + a$  auf  $[0, 1]$  gleichverteilt.

**15.5** Verwenden Sie Aufgabe 15.1, um eine Formel für den MC-Schätzer  $\hat{I}$  aus Satz 15.3 mit Hilfe von gleichverteilten Zufallsvariablen  $V_i, i \in \mathbb{N}$  auf  $[0, 1]$  herzuleiten.



# 16 Markov-Ketten Monte Carlo und Konvergenzgeschwindigkeit

---

## Übersicht

16.1 Grundideen .....	212
16.2 Metropolis-Algorithmus .....	212
16.3 Konvergenzgeschwindigkeit und Mischzeiten .....	215

---

Als Grundprinzip der Monte-Carlo Schätzung wurde die wiederholte Durchführung eines Zufallsexperiment bezeichnet, um aus den Realisierungen mit Hilfe des Gesetzes der großen Zahlen einen Schätzer zu erhalten. Im Falle von Markov-Ketten Monte Carlo besteht dies im wiederholten Aufrufen einer Markov-Kette, bis diese annähernd ihre stationäre Verteilung erreicht hat. Von praktischer Relevanz ist dabei auch die Frage, wie schnell eine Markov-Kette sich ihrer stationären Verteilung annähert.

### Lernziele

- Die Grundprinzipien von Markov Ketten Monte Carlo Methoden kennen
- Den Metropolis-Algorithmus kennen
- Einen Zusammenhang zwischen Konvergenzgeschwindigkeit und Spektrallücke kennen

### Vorkenntnisse

Es werden gute Kenntnisse aus dem Kapitel über Markov-Ketten, insbesondere über invariante Verteilungen, benötigt.

## 16.1 Grundideen

Die hier vorgestellten Methoden beruhen auf Kapitel 11, insbesondere Theorem 11.7. Dort wurde gezeigt, dass eine irreduzible, aperiodische Markov-Kette  $(X_n)_{n \in \mathbb{N}}$  auf einem endlichen Zustandsraum  $S$  gegen ihre invariante Verteilung  $\pi$  konvergiert (falls  $\pi$  existiert), d.h. dass für alle  $k \in S$ , unabhängig von der Startverteilung,

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n = k) = \pi_k$$

gilt.

Dies verwendet man bei der Markov-Ketten Monte Carlo Methode (MCMC-Methode). Dabei ist das Ziel, aus einer vorgegebenen Verteilung  $v = (v_x)_{x \in S}$  zu simulieren, wobei  $S$  eine (endliche) Menge ist. Die Idee ist nun, eine Markov-Kette  $X = (X_n)_{n \in \mathbb{N}}$  mit Zustandsraum  $S$  zu konstruieren, für welche  $v$  die invariante Verteilung ist, d.h.  $v = \pi$ . Dann lässt man die Markov-Kette einfach lang genug laufen, so dass nach Theorem 11.7 die Verteilung ungefähr gleich  $\pi$  ist.

Dabei stellen sich zwei grundsätzliche Fragen: Wie kann eine solche Kette konstruiert werden, und wie lange muss sie laufen, um hinreichend nah an die gewünschte Verteilung zu kommen?

Im nächsten Abschnitt wird eine oft verwendete Methode vorgestellt, und im übernächsten wird die Frage der Konvergenzgeschwindigkeit diskutiert.

## 16.2 Metropolis-Algorithmus

Eine wichtige Methode zur Konstruktion einer geeigneten Markov-Kette benutzt ein zweistufiges Verfahren.

1. Konstruktion einer einfachen *Referenzkette*, welche zulässige Übergänge „vorschlägt“
2. Bei jedem vorgeschlagenen Übergang wird entschieden, ob er tatsächlich ausgeführt wird oder nicht. Dabei werden bevorzugt diejenigen Übergänge ausgeführt, welche auf einen neuen Zustand mit einem großen Wahrscheinlichkeitsgewicht bezüglich der invarianten Verteilung führen.

### Metropolis-Algorithmus

Geben sei eine endliche Menge  $S$ , und eine Verteilung  $\pi = (\pi_k)_{k \in S}$  auf  $S$ .

1. Wähle eine *beliebige* Übergangsmatrix  $Q = (q_{k,l})_{k,l \in S}$  auf  $S$ . Die zu  $Q$  geörrige Markov-Kette ist die *Referenzkette*.

2. Definiere eine neue Übergangsmatrix  $P = (p_{k,l})_{k,l \in S}$  mit Hilfe von  $Q$  als

$$p_{k,l} = \begin{cases} q_{k,l} \cdot \min\left(1, \frac{\pi(l)q_{lk}}{\pi(k)q_{kl}}\right), & k \neq l \\ 1 - \sum_{l \neq k} p_{k,l}, & k = l. \end{cases}$$

2. Starte eine Markov-Kette  $(X_n)_{n \in \mathbb{N}}$  in einem beliebigen Punkt von  $S$ , und führe Übergänge entsprechend der Matrix  $P$  aus.

3. Stop nach einer vorgegebenen Anzahl Übergänge  $T$ . Notiere  $X_T$ .

Ergebnis: Nach wiederholtem Aufrufen des Algorithmus bildet die empirische Verteilung der Einzelergebnisse  $X_T$  ein Schätzer für die gewünschte Verteilung.

Als erstes überzeugen wir uns davon, dass die Matrix  $P$  aus diesem Algorithmus tatsächlich die invariante Verteilung  $\pi$  hat.

### Satz 16.1

Im Metropolis-Algorithmus ist  $\pi$  eine invariante Verteilung für  $P$ . Falls  $P$  irreduzibel und aperiodisch ist, konvergiert also die Markov-Kette gegen die gesuchte Verteilung  $\pi$ , d.h.

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n = k) = \pi_k, \quad k \in S.$$

**Beweis 16.1** Es ist zu zeigen, dass  $\sum_{k \in S} \pi_k p_{k,l} = \pi_l$  für alle  $l \in S$  gilt. Setze dabei die Definition von  $P$  ein.

$$\begin{aligned} \sum_{k \in S} \pi_k p_{k,l} &= \sum_{k \in S, k \neq l} \pi_k p_{k,l} + \pi_l p_{l,l} \\ &= \sum_{\substack{k \in S, k \neq l \\ \pi_l q_{l,k} \geq \pi_k q_{k,l}}} \pi_k p_{k,l} + \sum_{\substack{k \in S, k \neq l \\ \pi_l q_{l,k} < \pi_k q_{k,l}}} \pi_k p_{k,l} + \pi_l p_{l,l} \\ &= \sum_{\substack{k \in S, k \neq l \\ \pi_l q_{l,k} \geq \pi_k \cdot 1 \cdot q_{k,l}}} \pi_k q_{k,l} + \sum_{\substack{k \in S, k \neq l \\ \pi_l q_{l,k} < \pi_k q_{k,l}}} \pi_k \frac{\pi_l q_{l,k}}{\pi_k q_{k,l}} q_{k,l} + \pi_l p_{l,l} \\ &= \sum_{\substack{k \in S, k \neq l \\ \pi_l q_{l,k} \geq \pi_k q_{k,l}}} \pi_k \frac{\pi_l q_{l,k}}{\pi_l q_{l,k}} q_{k,l} + \sum_{\substack{k \in S, k \neq l \\ \pi_l q_{l,k} < \pi_k q_{k,l}}} \pi_l q_{l,k} + \pi_l p_{l,l} \\ &= \sum_{\substack{k \in S, k \neq l \\ \pi_l q_{l,k} \geq \pi_k q_{k,l}}} \pi_l p_{l,k} + \sum_{\substack{k \in S, k \neq l \\ \pi_l q_{l,k} < \pi_k q_{k,l}}} \pi_l p_{l,k} + \pi_l p_{l,l} \\ &= \pi_l \sum_{k \in S} p_{l,k} = \pi_l. \end{aligned}$$

Somit handelt es sich tatsächlich um eine invariante Verteilung. Die Konvergenzaussage folgt aus Satz 11.7.  $\square$

Das Vorgehen des Algorithmus kann so interpretiert werden, dass  $Q$  Übergänge vorschlägt, welche mit Wahrscheinlichkeit  $\min\left(1, \frac{\pi(l)q_{lk}}{\pi(k)q_{kl}}\right)$  tatsächlich ausgeführt werden. Insbesondere werden also vorgeschlagene Übergänge immer ausgeführt, falls  $\frac{\pi(l)q_{lk}}{\pi(k)q_{kl}} \geq 1$  ist,

falls also der Übergang die Kette in einen Zustand mit höherem Gewicht bezüglich  $\pi$  bringt. Jedoch werden auch Übergänge die auf ein kleineres Wahrscheinlichkeitsgewicht führen, mit einer gewissen Wahrscheinlichkeit ausgeführt. Dies hilft zu vermeiden, dass der Algorithmus in eine „Sackgasse“ gerät. Für die konkrete Anwendung ist hilfreich, dass man nur die Verhältnisse  $\pi(l)q_{lk}/\pi(k)q_{kl}$  zu kennen braucht. Durch geschickte Wahl von  $Q$  kann der Algorithmus optimiert oder vereinfacht werden. Eine hilfreiche Vereinfachung entsteht, wenn  $Q$  so gewählt wird, dass  $q_{k,l} = q_{l,k}$  ist, da dann die Bedingung zu  $\frac{\pi(l)}{\pi(k)} \geq 1$  wird.

Zur Illustration der Methode nun ein Beispiel.

### Beispiel 16.1 (Simulation aus einer bedingten Verteilung)

Betrachte unabhängige, identisch verteilte, diskrete Zufallsvariablen  $Y_1, \dots, Y_N$  mit bekannter Verteilung  $r_i, i \in \mathbb{N}_0$ . Es gilt also  $\mathbb{P}(Y_j = i) = r_i$ . Gesucht ist die bedingte Verteilung des Vektors  $(Y_1, \dots, Y_N)$  gegeben  $Y_1 + \dots + Y_N = M$ , für ein vorgegebenes  $M \in \mathbb{N}$ , d.h.

$$\pi_y := \mathbb{P}(Y_1 = y_1, \dots, Y_N = y_N \mid Y_1 + \dots + Y_N = M), \quad y = (y_1, \dots, y_N) \in \mathbb{N}^N.$$

Diese Verteilung ist abhängig von  $N$  und  $M$ , und natürlich von der Verteilung  $r_i, i \in \mathbb{N}_0$ . Sie ist insbesondere im Allgemeinen auch für explizit bekanntes  $(r_i)_{i \in \mathbb{N}_0}$  nicht unbedingt einfach zu berechnen. Deshalb können Simulationsverfahren hilfreich sein.

Die gesuchte bedingte Verteilung von  $(Y_1, \dots, Y_N)$  gegeben  $Y_1 + \dots + Y_N = M$  hat die Form

$$\pi_k = \frac{\mathbb{P}(Y_1 = y_1, \dots, Y_N = y_N, Y_1 + \dots + Y_N = M)}{\mathbb{P}(Y_1 + \dots + Y_N = M)}$$

Die rechte Seite dieser Gleichung ist gleich 0 falls  $y_1 + \dots + y_N \neq M$ , und gleich

$$\frac{1}{C(N, M)} \cdot r_{y_1} \cdot \dots \cdot r_{y_N},$$

falls  $y_1 + \dots + y_N = M$  ist, wobei  $C(N, M) = \mathbb{P}(Y_1 + \dots + Y_N = M)$  gilt. Da im Metropolis-Algorithmus nur die Verhältnisse

$$\frac{\pi_k}{\pi_l} = \frac{r_{k_1} \cdot \dots \cdot r_{k_N}}{r_{l_1} \cdot \dots \cdot r_{l_N}}$$

eine Rolle spielen, braucht die Konstante  $C(N, M)$  nicht berechnet zu werden.

Wir konstruieren nun eine geeignete Übergangsmatrix  $Q$  für die Referenzkette. Zur Vereinfachung wollen wir nur Übergänge von  $k$  nach  $l$  zulassen, wenn sich die Vektoren  $k$  und  $l$  an genau zwei Einträgen unterscheiden. Wir arbeiten auf dem Zustandsraum  $S := \{k = (k_1, \dots, k_N) \in \mathbb{N}_0^N, k_1 + \dots + k_N = M\}$  und gehen folgendermaßen vor: Ist der aktuelle Zustand der Markov-Kette  $k = (k_1, \dots, k_N)$ , so springt die Kette in den neuen Zustand  $l = (l_1, \dots, l_N)$ , der sich wie folgt konstruiert:

- Wähle gleichverteilt ein Paar von Indizes  $(i, j) \in \{1, \dots, N\}, i \neq j$ .
- Wähle gleichverteilt einen Wert  $u \in \{0, \dots, k_i + k_j\}$ .
- Setze  $l_i = u, l_j = k_i + k_j - u$ , und  $l_m = k_m$  für  $m \notin \{i, j\}$ .

Die Kette macht also zufällige Übergänge, wobei der neue Zustand sich nur an zwei Einträgen vom alten unterscheidet, sowohl die Koordinaten als auch der Wert des neuen Eintrags wird uniform ausgewürfelt. Zu beachten ist, dass mit obigem Vorgehen auf jeden Fall  $k_1 + \dots + k_N = k_1 + \dots + k'_N$  gilt, was eine notwendige Bedingung ist. Es ist möglich, dass  $k = k'$  gilt (nämlich wenn  $u = k_i$  ist). Da es für die Wahl von  $(i, j)$  genau  $\binom{N}{2}$  Möglichkeiten gibt, und für die Wahl von  $u$  genau  $k_i + k_j + 1$ , folgt

$$q_{k,l} = \begin{cases} \left(\frac{1}{\binom{N}{2}}\right) \cdot \frac{1}{k_i + k_j + 1}, & \text{falls sich } k \text{ und } l \text{ an genau den Stellen } i \text{ und } j \text{ unterscheiden,} \\ 0, & \text{falls sich } k \text{ und } l \text{ an mehr als 2 Stellen unterscheiden,} \\ 1 - \sum_{k' \neq k} q_{k,k'}, & \text{falls } k = l. \end{cases}$$

Auf diese Referenzkette kann nun der Metropolis-Algorithmus angewandt werden. Dabei gilt in diesem Beispiel  $q_{k,l} = q_{l,k}$ , so dass die Übergangsmatrix der Metropolis-Kette gegeben ist durch

$$p_{k,l} = \begin{cases} \left(\frac{1}{\binom{N}{2}(k_i + k_j + 1)}\right) \cdot \min\left(1, \frac{r_{k_1} \dots r_{k_N}}{r_{l_1} \dots r_{l_N}}\right), & \text{falls sich } k \text{ und } l \text{ an } i \text{ und } j \text{ unterscheiden,} \\ 0, & \text{falls sich } k \text{ und } l \text{ an } > 2 \text{ Stellen unterscheiden,} \\ 1 - \sum_{l \neq k} p_{k,l}, & k = l. \end{cases}$$

Die Berechnung von  $Q$  und der Verhältnisse  $\pi_{k,l}/\pi_{l,k}$  ist einfach, wie wir gesehen haben. Insbesondere braucht man auch nicht die Kardinalität von  $S$  zu bestimmen. Deshalb ist der Metropolis-Algorithmus in diesem Beispiel leicht anwendbar, um eine potentiell komplizierte, und nicht explizit gegebene, Verteilung zu simulieren. ■

## 16.3 Konvergenzgeschwindigkeit und Mischzeiten

Es wurde bereits mehrfach die Tatsache benutzt, dass eine irreduzible und aperiodische Markov-Kette auf einem endlichen Zustandsraum gegen ihre invariante Verteilung konvergiert. Um beispielsweise den Metropolis-Algorithmus anwenden zu können, muss ein geeignetes Abbruchkriterium, also einen endlichen Zeitpunkt  $T$ , gefunden werden, zu dem nach Möglichkeit die Markov-Kette  $X_T$  hinreichend nah (in einem geeigneten Sinn) an der invarianten Verteilung angekommen sein sollte. Wir möchten mit anderen Worten die Geschwindigkeit der Konvergenz quantifizieren. Dafür werden einige weitere Begriffe benötigt.

### Definition 16.1 (Mischzeit)

Sei  $(X_n)_{n \in \mathbb{N}_0}$  eine Markov-Kette mit invarianter Verteilung  $\{\pi_a\}_{a \in S}$ . Sei  $\varepsilon > 0$ . Die  $\varepsilon$ -Mischzeit  $\tau(\varepsilon)$  der Kette ist definiert als das kleinste  $n > 0$ , so dass für alle  $a \in S$  und für beliebige Startverteilung gilt:

$$|\mathbb{P}(X_n = a) - \pi(a)| \leq \varepsilon.$$



Das bedeutet: Für  $n = \tau(\varepsilon)$  ist der Abstand von der invarianten Verteilung höchstens  $\varepsilon$ . Man kann sogar beweisen, dass für alle  $n \geq \tau(\varepsilon)$  der Abstand der Kette von der invarianten Verteilung höchstens  $\varepsilon$  ist, d.h.  $|\mathbb{P}(X_n = a) - \pi(a)| \leq \varepsilon$  für alle  $n \geq \tau(\varepsilon)$ . Wird nur von der Mischzeit gesprochen, ohne  $\varepsilon$  anzugeben, so wird üblicherweise  $\varepsilon = 1/4$  gewählt.

Natürlich stellt sich nun die Frage, wie die Mischzeit in konkreten Fällen berechnet werden kann. Dies geschieht mit Hilfe der Eigenwerte der stochastischen Matrix  $P$  der Markov-Kette  $(X_n)_{n \in \mathbb{N}}$ .

**Satz 16.2 (Eigenwerte stochastischer Matrizen)**

Sei  $S$  eine endliche Menge mit  $|S| = N$ . Sei  $P$  die Übergangsmatrix einer irreduziblen, aperiodischen Markov-Kette auf  $S$ . Dann sind alle Eigenwerte dieser Matrix dem Betrag nach kleiner oder gleich 1.

**Beweis 16.2** Siehe Levin et al. (2017), Kapitel 12. □

**Definition 16.2 (Spektrallücke)**

Sei  $\lambda^*$  der dem Betrage nach größte Eigenwert von  $P$ , welcher  $\neq 1$  ist. Dann heißt

$$\gamma^* := 1 - |\lambda^*|$$

(absolute) Spektrallücke von  $P$  (bzw. der zugehörigen Markov-Kette). ◆

Das Hauptergebnis dieses Abschnitts kann nun formuliert werden. Wir verzichten auf einen Beweis. Dieser, und viel weiterführendes Material im Umfeld unserer Fragestellung findet man beispielsweise in Levin et al. (2017), Kapitel 12.

**Satz 16.3 (Spektrallücke und Mischzeit)**

Sei  $(X_n)_{n \in \mathbb{N}}$  eine Markov-Kette auf einem endlichen Zustandsraum  $S$ , mit Übergangsmatrix  $P$  und invarianter Verteilung  $\{\pi_k\}_{k \in S}$ , welche zusätzlich  $\pi_k p_{k,l} = \pi_l p_{l,k}$  für alle  $k, l \in S$  erfüllt. Sei  $\bar{\pi}$  das kleinste  $\pi_l, l \in S$ . Für die  $\varepsilon$ -Mischzeit gilt dann

$$\tau(\varepsilon) \leq \ln \left( \frac{1}{\bar{\pi} \varepsilon} \right) \cdot \frac{1}{\gamma^*}.$$

Daraus folgt insbesondere für  $k \in S, n \in \mathbb{N}$

$$|P(X_n = k) - \pi_k| \leq \frac{1}{\bar{\pi}} e^{-\gamma^* n}$$

Das bedeutet, je größer die Spektrallücke, desto schneller konvergiert die Markov-Kette gegen die invariante Verteilung.

Eine Verteilung  $\{\pi_k\}_{k \in S}$ , welche  $\pi_k p_{k,l} = \pi_l p_{l,k}$  für alle  $k, l \in S$  erfüllt, heißt auch *reversible Verteilung* von  $P$ . Man kann zeigen, dass eine reversible Verteilung automatisch invariant ist, die Umkehrung gilt jedoch nicht. Diese Bedingung ist in obigem Satz deswegen bedeutsam, weil sie sicherstellt, dass die Eigenwerte von  $P$  alle reell sind. Dies beweisen wir hier jedoch nicht.

Um die Aussage  $\mathbb{P}(X_n = k) \approx \pi(k)$  präzise zu machen, kann also folgendermaßen vorgegangen werden:

1. Wähle die Präzision  $\varepsilon$ .
2. Berechne  $\pi$  und untersuche ob die zusätzliche Bedingung erfüllt ist.
3. Berechne die Eigenwerte von  $P$ .
4. Berechne  $\bar{\pi}$  und die Spektrallücke  $\gamma^*$ .
5. Dann gilt für alle  $n \geq \log(1/(\bar{\pi}\varepsilon)) \frac{1}{\gamma^*}$ , dass

$$|\mathbb{P}(X_n = k) - \pi(k)| \leq \varepsilon$$

für alle  $k$  und jede Startverteilung gilt.

Die Bestimmung von  $\pi$  und  $\gamma^*$  ist ein im Prinzip einfaches Problem der linearen Algebra.

### Zusammenfassung

Markov-Ketten Monte Carlo Methode benutzen den Konvergenzsatz für (irreduzible, aperiodische, positiv rekurrente) Markov-Ketten, um eine vorgegebene Verteilung zu simulieren. Die Konvergenzgeschwindigkeit wird durch die Mischzeit bestimmt, und kann zumindest in gewissen Fällen mit Hilfe der Spektrallücke abgeschätzt werden.

## Aufgaben

**16.1** Man zeige, dass aus der Eigenschaft  $\pi_k p_{k,l} = \pi_l p_{l,k}$  für alle  $k, l \in S$  folgt, dass  $\pi$  eine invariante Verteilung von  $P$  ist.

**16.2** Man zeige, dass die Matrix  $P$  aus dem Metropolis-Algorithmus sogar  $\pi_k p_{k,l} = \pi_l p_{l,k}$  für alle  $k, l \in S$  erfüllt.

**16.3** Sei  $S = \{1, 2, 3\}$  und

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 \end{bmatrix}.$$

(a) Berechnen Sie die invariante Verteilung  $\pi$ . (b) Überprüfen Sie, ob die Voraussetzungen von Satz 16.3 erfüllt sind, und berechnen Sie die Spektrallücke. Geben Sie falls möglich eine obere Schranke für die  $\varepsilon$ -Mischzeit an.





# Anhang A Tabellen

$x$	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Tab. A.1: Die Verteilungsfunktion  $\Phi_{0,1}(x) = \mathbb{P}(X \leq x)$  für eine standardnormalverteilte Zufallsvariable  $X$ . Beachte  $\Phi_{0,1}(-x) = 1 - \Phi_{0,1}(x)$  für negative Werte von  $x$ . Ablesebeispiel:  $x = 0.94$ . Suche zuerst die Zeile mit  $x = 0.9$ , und lese dann den Wert in der Spalte zu  $x = 0.04$  ab, also  $\Phi_{0,1}(0.94) = 0,8264$ .

$\beta$	0.9	0.95	0.96	0.975	0.98	0.99
$z_\beta$	1.282	1.645	1.751	1.960	2.054	2.326

Tab. A.2: Quantile  $z_\beta$  mit  $\mathbb{P}(X \leq z_\beta) = \beta$  für  $X \sim \mathcal{N}(0, 1)$ . Zur Berechnung beachte außerdem die Symmetrie  $z_{1-\beta} = -z_\beta$ , also z.B.  $z_{0.05} = -z_{0.95} = -1.645$ .

$n \backslash \beta$	0.9	0.95	0.975	0.99
1	3,078	6,314	12,706	31,821
2	1,886	2,920	4,303	6,965
3	1,638	2,353	3,182	4,541
4	1,533	2,132	2,776	3,747
5	1,476	2,015	2,571	3,365
6	1,440	1,943	2,447	3,143
7	1,415	1,895	2,365	2,998
8	1,397	1,860	2,306	2,896
9	1,383	1,833	2,262	2,821
10	1,372	1,812	2,228	2,764
11	1,363	1,796	2,201	2,718
12	1,356	1,782	2,179	2,681
13	1,350	1,771	2,160	2,650
14	1,345	1,761	2,145	2,624
15	1,341	1,753	2,131	2,602
16	1,337	1,746	2,120	2,583
17	1,333	1,740	2,110	2,567
18	1,330	1,734	2,101	2,552
19	1,328	1,729	2,093	2,539
20	1,325	1,725	2,086	2,528
25	1,316	1,708	2,060	2,485
30	1,310	1,697	2,042	2,457
40	1,303	1,684	2,021	2,423
50	1,299	1,676	2,009	2,403
100	1,290	1,660	1,984	2,364
1000	1,282	1,646	1,962	2,330

Tab. A.3: Die Tabelle gibt  $t_{n,\beta}$  mit  $\mathbb{P}(X \leq t_{n,\beta}) = \beta$  für eine  $t$ -verteilte Zufallsvariable  $X$  mit Freiheitsgrad  $n$  an. Symmetrie:  $t_{n,1-\beta} = -t_{n,\beta}$ .

$\beta \backslash n$	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
1	0,00	0,00	0,00	0,02	2,71	3,84	5,02	6,63
2	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21
3	0,11	0,22	0,35	0,58	6,25	7,81	9,35	11,34
4	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28
5	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09
6	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48
8	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09
9	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67
10	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21
11	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,73
12	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22
13	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69
14	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14
15	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58
16	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00
17	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41
18	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,81
19	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19
20	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57
25	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31
30	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89
40	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69
50	29,71	32,36	34,76	37,69	63,17	67,50	71,42	76,15
100	70,06	74,22	77,93	82,36	118,50	124,34	129,56	135,81

Tab. A.4: Die Tabelle gibt  $\chi^2_{n,\beta}$  mit  $\mathbb{P}(X \leq \chi^2_{n,\beta}) = \beta$  für eine  $\chi^2$ -verteilte Zufallsvariable  $X$  mit Freiheitsgrad  $n$  an. Vorsicht: Nicht symmetrisch.



# Anhang B Lösungen der Aufgaben

---

---

## Lösungen zu Kapitel 1

**1.1** a)  $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8\}$ , b)  $A \cap B = \{5, 6, 7\}$ , c)  $B \cup C = \{1, 3, 5, 7\}$ , d)  $(A^c \cup B^c) \cap C = \{1, 3\}$

**1.2** a)  $\cup_{i=1}^n D_i$ , b)  $\cup_{i=1}^n (D_i \cap \cap_{j \neq i} D_j^c)$ , c)  $\cup_{i=1}^n \cap_{j \neq i} D_j^c$ , d)  $\cup_{i=1}^n \cap_{j \neq i} D_j$

**1.3** a)  $A = [0, 400)$ , b)  $B = [380, 780]$ , c)  $A \cap B = [380, 400)$  d)  $A^c \cup B = [380, \infty)$

**1.4** a)  $8^3$ , b) 21, c)  $\frac{1}{64}$ , d)  $\frac{21}{32}$

**1.5**  $\mathbb{P}(A \Delta B) = \mathbb{P}((A \cap B^c) \cup (A^c \cap B)) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A^c \cap B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(B \cap A)$ .

**1.6** Wiederholte Anwendung von Satz ??

**1.7**  $\mathbb{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{i \neq j} \mathbb{P}(A_i \cap A_j) + \sum_{i \neq j \neq k} \mathbb{P}(A_i \cap A_j \cap A_k) - \dots + (-1)^{n+1} \mathbb{P}(A_1 \cap \dots \cap A_n)$ .

**1.8** a)  $\frac{32}{81}$ , b)  $\frac{65}{81}$

**1.9** a) 64, b) 24, c)  $\frac{37}{64}$  d)  $\frac{3}{4}$

## Lösungen zu Kapitel 2

**2.1** a)  $\frac{1}{2}$ , b)  $\frac{1}{2}$ , c)  $\frac{1}{6}$ , d)  $\frac{1}{3}$

**2.2** a)  $\frac{1}{21}$ , b)  $\frac{7}{9}$ , c)  $\frac{1}{3}$ , d) 0

**2.3** a) 0.096, b) 0.168, c)  $\frac{12}{21}$

**2.4** a) nein, b) ja, c) ja, d) nein

**2.5** a) 0.04, b) 0.25, c)  $\approx 0.24$

**2.6**  $\mathbb{P}(A \cap B^c) = \mathbb{P}(A \setminus (A \cap B)) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A) \cap \mathbb{P}(B)$  da  $A$  und  $B$  unabhängig sind. Somit folgt  $\mathbb{P}(A \cap B^c) = \mathbb{P}(A)(1 - \mathbb{P}(B)) = \mathbb{P}(A)\mathbb{P}(B^c)$  was zu zeigen ist. Die anderen Fälle folgen ähnlich.

**2.7** a) Folgt durch direktes Nachrechnen,  $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(a) = 1/27$ . b) Nachrechnen zeigt z.B.  $\mathbb{P}(A \cap B) \neq \mathbb{P}(A)\mathbb{P}(B)$ .

**2.8** a)  $\approx 0.931$  b)  $\approx 0.636$

### Lösungen zu Kapitel 3

**3.1** a)  $7/36$ , b)  $1/9$ , c)  $7/12$

**3.2**  $X(\Omega) = \{1, 2, 3, 4, 5, 6\}$ ,  $\mathbb{P}(X = 1) = \frac{1}{6}$ ,  $\mathbb{P}(X = 2) = \frac{1}{6}$ ,  $\mathbb{P}(X = 3) = \frac{1}{6}$ ,  $\mathbb{P}(X = 4) = \frac{1}{6}$ ,  $\mathbb{P}(X = 5) = \frac{1}{6}$ ,  $\mathbb{P}(X = 6) = \frac{1}{6}$ .  $W(\Omega) = \{1, 2, 3, 4, 5, 6\}$ ,  $\mathbb{P}(W = 1) = \frac{11}{36}$ ,  $\mathbb{P}(W = 2) = \frac{1}{4}$ ,  $\mathbb{P}(W = 3) = \frac{7}{36}$ ,  $\mathbb{P}(W = 4) = \frac{5}{36}$ ,  $\mathbb{P}(W = 5) = \frac{1}{12}$ ,  $\mathbb{P}(W = 6) = \frac{1}{36}$ .  $Z(\Omega) = \{2, 4, 8, 16, 32, 64\}$ ,  $\mathbb{P}(Z = k) = \frac{1}{6}$  für alle  $k \in Z(\Omega)$ .

**3.3** a)  $\frac{54}{125}$  b)  $\frac{9}{40}$  c) 0 d)  $\frac{243}{500}$

**3.4** a)  $\approx 0.758$  b)  $\approx 0.028$  c)  $\approx 0.031$

**3.5**  $\approx 0.74$

**3.6**  $\frac{1}{1024}$

**3.7**  $\approx 0.202$

**3.9** a)  $\approx 0.11$  b)  $\approx 0.27$  c)  $\approx 0.00018$

**3.9** a) Das Ereignis  $\{X \geq k\}$  tritt genau dann ein, wenn mindestens die ersten  $k - 1$  Versuche im Bernoulli-Experiment Misserfolge waren, was genau mit Wahrscheinlichkeit  $\mathbb{P}(X > k - 1) = (1 - p)^{k-1}$  geschieht. b) Nach a) gilt  $\mathbb{P}(X > k + n | X > k) = \frac{\mathbb{P}(X > k + n)}{\mathbb{P}(X > k)} = \frac{(1-p)^{n+k}}{(1-p)^k} = (1-p)^n = \mathbb{P}(X > n)$ .

**3.10**  $\lim_{n \rightarrow \infty} \mathbb{P}(X > bn) = \lim_{n \rightarrow \infty} (1 - \frac{a}{n})^{bn} = e^{-ab}$ .

### Lösungen zu Kapitel 4

**4.1** a) Siehe Tabelle B.1 b)  $\mathbb{P}(X = 2) = 1/6$  c)  $\mathbb{P}(Y = 5) = 1/48$

**4.2**  $\mathbb{P}(X = x | Y = 5) = 0$  für  $x = 1, \dots, 4$ ,  $\mathbb{P}(X = 5 | Y = 5) = 1/4$ ,  $\mathbb{P}(X = 6 | Y = 5) = 3/4$

**4.3** a)  $\mathbb{P}(Y = -1 | X = -1) = 1/7$ ,  $\mathbb{P}(Y = 0 | X = -1) = 5/7$ ,  $\mathbb{P}(Y = 1 | X = -1) = 1/7$   
b)  $X^2(\Omega) = Y^2(\Omega) = \{0, 1\}$ ,  $\mathbb{P}(X^2 = 0, Y^2 = 0) = 1/4$ ,  $\mathbb{P}(X^2 = 0, Y^2 = 1) = 3/20$ ,  $\mathbb{P}(X^2 = 1, Y^2 = 0) = 7/20$ ,  $\mathbb{P}(X^2 = 1, Y^2 = 1) = 1/4$ . c) nein, z.B. ist  $\mathbb{P}(X^2 = 0, Y^2 = 0) = 1/4 \neq 2/5 \cdot 3/5 = \mathbb{P}(X^2 = 0)\mathbb{P}(Y^2 = 0)$ .

**4.4**  $Y(\Omega) = \{1, 2, 4, 6\}$ ,  $Z(\Omega) = \{0, 1\}$ ,  $\mathbb{P}(Y = 1, Z = 0) = 1/6$ ,  $\mathbb{P}(Y = 1, Z = 1) = 1/3$ ,  $\mathbb{P}(Y = 2, Z = 0) = 0$ ,  $\mathbb{P}(Y = 2, Z = 1) = 1/6$ ,  $\mathbb{P}(Y = 4, Z = 0) = 0$ ,  $\mathbb{P}(Y = 4, Z = 1) = 1/6$ ,  $\mathbb{P}(Y = 6, Z = 0) = 1/6$ ,  $\mathbb{P}(Y = 6, Z = 1) = 0$ .

$Y \backslash X$	1	2	3	4	5	6
0	1/12	1/24	1/48	1/96	1/192	1/384
1	1/12	1/12	1/16	1/24	5/192	1/64
2	0	1/24	1/16	1/16	10/192	5/128
3	0	0	1/48	1/24	10/192	5/96
4	0	0	0	1/96	5/192	5/128
5	0	0	0	0	1/192	1/64
6	0	0	0	0	0	1/384

Tab. B.1: Zu Aufgabe 4.1

- 4.5**  $X^2(\Omega) = \{0, 1\}$ ,  $\mathbb{P}(X^2 = 0) = \mathbb{P}(X^2 = 1) = 1/2$ ,  $\mathbb{P}(X = 1|X^2 = 1) = \mathbb{P}(X = -1|X^2 = 1) = 1/2$ ,  $\mathbb{P}(X = 0|X^2 = 0) = 1$ , alle anderen bedingten Wahrscheinlichkeiten sind 0.
- 4.6** a) Aus der Faltungsformel:  $\mathbb{P}(X + Y = 2k) = \sum_{m=0}^2 k \mathbb{P}(X = m) \mathbb{P}(Y = 2k - m) \geq \mathbb{P}(X = k) \mathbb{P}(Y = k)$  b) es gilt  $X + Z = X - X = 0$ , somit ist  $\mathbb{P}(X + Z = 2k) = 0$  für  $k \neq 0$ . c) Da  $X + Z = 0$  ist, ist die linke Seite stets 0 außer für  $k = 0$ , die rechte im Allgemeinen jedoch nicht.

## Lösungen zu Kapitel 5

- 5.1** a)  $\mathbb{E}[X] = \frac{5}{4}$ , b)  $\mathbb{E}[Y] = \frac{8}{5}$ , c)  $\mathbb{V}(X) = \frac{15}{16}$ , d)  $\mathbb{V}(Y) = \frac{824}{25}$
- 5.2** a)  $\mathbb{E}[X] = 1$ ,  $\mathbb{E}[Y] = 0.4$  b)  $\mathbb{V}(X) = 0.6$ ,  $\mathbb{V}(Y) = 0.24$  c)  $\text{cov}(X, Y) = 0.2$
- 5.3** a) 12, b) 1.2, c) 6.2, c) 2.3
- 5.4**  $\mathbb{E}[Y] = 0$ ,  $\text{cov}(X, Y) = \frac{32}{5}$ .
- 5.5** a)  $a \in [0, 1/2]$  b)  $\mathbb{E}[X] = \frac{5}{4} - 3a$ ,  $\mathbb{E}[Y] = \frac{3}{4} - a$ ,  $\mathbb{E}[XY] = \frac{3}{4} - 2a$ . c)  $\text{cov}(X, Y) = -\frac{3}{16} + \frac{3}{2}a - 3a^2$  d) unkorreliert für  $a = 1/4$ , negativ korreliert für alle anderen zulässigen Werte von  $a$ .
- 5.6** Mehrfaches Anwenden des Tricks aus Beispiel 5.3 und einige Umformungen führen auf  $\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1} - \frac{1}{p^2} = p \cdot \frac{d^2}{dp^2} \frac{(1-p)^2}{p} + p \frac{d}{dp} \frac{(1-p)}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}$ .
- 5.7**  $\mathbb{E}[X] = \sum_{k=1}^{\infty} k \cdot \frac{\lambda^k e^{-\lambda}}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda$ , wobei im letzten Schritt eine Indexverschiebung und die e-Reihe benutzt wurde.
- 5.8**  $\mathbb{E}[e^X] = p e^1 + (1-p) e^0 = 1 + p(e-1)$ .  $\mathbb{E}[e^Y] = \mathbb{E}[e^{\sum_{i=1}^n X_i}]$ , mit  $X_i, i = 1, \dots, n$  unabhängige Bernoulli-verteilte Zufallsvariablen mit Parameter  $p$ . Aus der Unabhängigkeit und dem ersten Resultat folgt  $\mathbb{E}[e^Y] = \prod_{i=1}^n \mathbb{E}[e^{X_i}] = (1 + p(e-1))^n$ .

## Lösungen zu Kapitel 6

**6.1** a) Dichte b) Dichte c) keine Dichte da  $f(1) < 0$ , d) Dichte da  $\sum_{k=1}^{\infty} \frac{1}{2^k} = 1$

**6.2** a)  $c = 2$  b)  $2$  c)  $\infty$  d)  $F_X(t) = \frac{t^2-1}{t^2}$

**6.3**  $\mathbb{E}[X] = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{b^2-a^2}{2} \frac{1}{b-a} = \frac{b+a}{2}$ ,  $\mathbb{V}(X) = \frac{(b-a)^2}{12}$ .

**6.4** a)  $a = \frac{2}{13}, b = \frac{11}{26}$  b)  $\frac{11}{13}$  c)  $\frac{11}{26}$  d)  $0$

**6.6** a)  $\mathbb{E}[Y^2] = \frac{9}{2}$  b)  $\mathbb{P}(1 \leq Y < 5/2) \approx 0.3245$  c)  $\mathbb{P}(Y^2 \geq 4) \approx 0.2636$  d)  $\mathbb{P}(2Y < 1) \approx 0.2835$

**6.8**  $\mathbb{P}(X > a + b | X > a) = \frac{\mathbb{P}(X > a+b)}{\mathbb{P}(X > a)} = \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b} = \mathbb{P}(X > b)$ .

**6.9** Nach Satz 6.8 ist  $\sum_{i=1}^n$  normalverteilt mit Erwartungswert  $\sum_{i=1}^n \mu = n\mu$  und Varianz  $n\sigma^2$ . Somit folgt die Behauptung aus Satz 6.7.

**6.11** a)  $\mathbb{P}(X \geq 0) \approx 0.6915$  b)  $\mathbb{P}(-4 \leq X \leq -3) \approx 0.0166$  c)  $\mathbb{P}(|X| < 2) \approx 0.6247$  d)  $\mathbb{P}(3X < 4) \approx 0.5675$

**6.12** a)  $x = 0$  b)  $x \approx 0.67$

## Lösungen zu Kapitel 7

**7.2** (todo)

**7.3** Mit dem Gesetz der großen Zahlen:  $X_1 + \dots + X_{10'000} \approx \mathbb{E}[X_1] \cdot 10'000 = 10'000\lambda^{-1}$ .

Mit dem Zentralen Grenzwertsatz:  $\mathbb{P}(X_1 + \dots + X_{10'000} \leq 8600) \approx \mathbb{P}(Y \leq \frac{8600 - 10'000\mathbb{E}[X_1]}{\sqrt{10'000\mathbb{V}(X_1)}}) = \mathbb{P}(Y \leq \frac{8600 - 10'000 \cdot \lambda^{-1}}{100 \cdot \lambda^{-2}})$ .

**7.4** a)  $\approx 0.0548$  b)  $\approx 0.1711$  c)  $\approx 0.8968$  d)  $\approx 0.4801$

**7.5**  $\approx 0.0022$

## Lösungen zu Kapitel 8

**8.1** a)  $51.7$  b)  $50.5$  c)  $\approx 67.57$ .

**8.5** a) negativ korreliert b)  $\approx -0.96$  c)  $y = -1.03x + 11.35$

## Lösungen zu Kapitel 9

**9.1** a)  $\approx 12.72$  b)  $\approx 12.44$  c)  $\approx 10.34$

**9.3** a)  $\approx 44.18$  b)  $[245.68, 253.92]$

**9.4**  $[20.45, 22.97]$



## Lösungen zu Kapitel 10

**10.2**  $t_{9,0.95} = 1.833, t_{9,0.975} = 2.262$ . Testwert  $t = -2.077$ . Annahme von  $H_0$  für  $\alpha = 0.05$ , Ablehnung für  $\alpha = 0.1$ .

**10.3**  $H_0 : \mu \geq \mu_0$ .  $t_{9,0.05} = -1.833, t_{9,0.1} = -1.383, t = -2.077$ . Verwerfen in beiden Fällen.

## Lösungen zu Kapitel 11

**11.2** 
$$P = \begin{bmatrix} 0.7 & 0 & 0.3 \\ 0 & 0.4 & 0.6 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}.$$

Die Kette ist irreduzibel und aperiodisch.  $\mathbb{P}(X_3 = 1 | X_0 = 1) = 0.457, \mathbb{P}(X_3 = 2 | X_0 = 1) = 0.144, \mathbb{P}(X_3 = 3 | X_0 = 1) = 0.399$ .

**11.3** a) irreduzibel, periodisch, b) irreduzibel, aperiodisch, c) nicht irreduzibel

**11.4** a)  $\pi \approx (0.35, 0.17, 0.36, 0.12)$ , b)  $\pi = (\frac{4}{13}, \frac{3}{13}, \frac{6}{13})$ .

**11.5**  $\frac{28}{13}$

**11.6** a) Nicht irreduzibel b)  $\pi = (0, 0, 1)$  (eindeutig) c) Wegen b) gilt  $\lim_{n \rightarrow \infty} \mathbb{P}(X_n = 3 | X_0 = 1) = 1$ , und  $\lim_{n \rightarrow \infty} \mathbb{P}(X_n = i | X_0 = 1) = 0$  für  $i \in \{1, 2\}$ .

**11.7** a) nicht irreduzibel b)  $\pi = (0, a, 1 - a), a \in [0, 1]$ , c)  $\lim_{n \rightarrow \infty} \mathbb{P}_\nu(X_n = 1) = 0, \lim_{n \rightarrow \infty} \mathbb{P}_\nu(X_n = 2) = 1, \lim_{n \rightarrow \infty} \mathbb{P}_\nu(X_n = 3) = 0$ .

**11.9** Nach Satz 11.9 ist  $\pi_i = (1 - \frac{\lambda}{\mu}) \left(\frac{\lambda}{\mu}\right)^i, i \in \mathbb{N}_0$ . Die erwartete Anzahl Kunden auf lange Sicht ist  $\sum_{i=1}^{\infty} i \cdot \pi_i = \frac{\lambda}{\mu - \lambda}$ .

## Lösungen zu Kapitel 12

(todo)

## Lösungen zu Kapitel 13

(todo)

## Lösungen zu Kapitel 14

### 14.4

$$Q = \begin{bmatrix} -1 & 0.2 & 0.5 & 0 & 0.3 \\ 0.1 & -1 & 0.1 & 0.1 & 0.7 \\ 0.8 & 0 & -1 & 0.2 & 0 \\ 0.5 & 0.2 & 0.2 & -2 & 0.1 \\ 0.4 & 0.3 & 0.2 & 0.1 & -5 \end{bmatrix}.$$

**14.5**  $\lambda_1 \geq 2, \lambda_2 \geq 4, \lambda_3 \geq 3, p_{i,i} = 1 + q_{i,i}/\lambda_i, p_{i,j} = q_{i,j}/\lambda_i$ .

**14.6** Nachrechnen zeigt  $0 \leq \bar{p}_{i,j} \leq 1$  für alle  $i, j \in S$ , und  $\sum_{j \in S} \bar{p}_{i,j} = 1$  für alle  $i \in S$ .

**14.7**  $\lambda_i \geq -q_{i,i}, i \in S$  und  $p_{i,i} = 1 + q_{i,i}/\lambda_i, p_{i,j} = q_{i,j}/\lambda_i$ .

**14.8**  $\pi = (2/5, 1/5, 2/5)$

**14.9** Nachrechnen zeigt  $\pi^T Q = 0 \Leftrightarrow \pi^T P = \pi^T$ , was die Behauptung ist.

## Lösungen zu Kapitel 15

(todo)

## Lösungen zu Kapitel 16

(todo)

# Literaturverzeichnis

- Brin, S. und Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, Band 30 107–117.
- Dümbgen, L. (2003) Stochastik für Informatiker. Springer Verlag, Berlin Heidelberg.
- Geiger, J. (2006) Algorithmen und Zufall. [docplayer.org/33362275-Algorithmen-und-zufall.html](http://docplayer.org/33362275-Algorithmen-und-zufall.html).
- Georgii, H.-O. (2009) Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik. De Gruyter
- Henze, N. (2013) Stochastik für Einsteiger. Springer Spektrum
- Hübner, G. (2009) Stochastik. Eine anwendungsorientierte Einführung für Informatiker, Ingenieure und Mathematiker. Vieweg+Teubner, Wiesbaden.
- Klenke, A. (2008) Wahrscheinlichkeitstheorie. Springer Verlag, Berlin Heidelberg.
- König, W. (2003) Stochastische Algorithmen. <https://www.wias-berlin.de/people/koenig/www/AlgStoch.pdf>.
- Levin, D. A., Peres, Y. und Wilmer, E. L. (2017) Markov chains and mixing times. American Mathematical Society, Providence, RI.
- Mitzenmacher, M. und Upfal, E. (2017) Probability and Computing. Cambridge University Press, Cambridge.
- Motwani, R. und Raghavan, P. (1995) Randomized Algorithms. Cambridge University Press, Cambridge.
- Scheutzwow, M. (2018) Stochastische Modelle. <http://page.math.tu-berlin.de/scheutzwow/main.pdf>.
- Storrer, H. H. (1995) Einführung in die mathematische Behandlung der Naturwissenschaften, Band II. Birkhäuser Skripten