# ELEMENTARE WAHRSCHEINLICHKEITSTHEORIE UND STATISTIK

**Wolfgang König**

Vorlesungsskript

Universität zu Köln

Wintersemester 2002/03

# Inhaltsverzeichnis

# Vorwort

Dies ist das Vorlesungsskript einer vierstündigen elementaren einführenden Vorlesung über Stochastik. Hierbei werden die beiden Teilgebiete Wahrscheinlichkeitstheorie und Statistik zu gleichen Teilen behandelt. Es werden die grundlegenden Begriffe, Beispiele und Methoden motiviert und entwickelt, ohne Anspruch auf Allgemeinheit oder Vollständigkeit zu erheben.

Die Vorlesung ist elementar in dem Sinne, dass Vorkenntnisse nur in den Grundlagen der Analysis und Linearen Algebra voraus gesetzt werden und dass vollständig auf die Maßtheorie, insbesondere das Lebesgue-Maß, verzichtet wird. Es werden also nur diskrete Wahrscheinlichkeitsräume und Wahrscheinlichkeiten, die man über eine Riemann-integrierbare Dichte definieren kann, behandelt. Daher ist diese Vorlesung für Mathematikstudenten ab dem dritten (oder auch schon ab dem zweiten) Semester geeignet, und auch für Studenten der meisten anderen Richtungen, sofern sie eine Grundausbildung in Analysis oder Linearer Algebra genossen haben.

Auf der anderen Seite zwingt uns der Verzicht auf Maßtheorie dazu, die mathematische Fundierung gewisser Teilkonzepte, etwa die der Zufallsgrößen mit Dichten oder unendlich viele unabhängige Zufallsgrößen, außen vor zu lassen und uns mit einem intuitiven Verständnis zu begnügen. Die Behandlung der (allgemeinen) Wahrscheinlichkeitstheorie, insbesondere die der Wahrscheinlichkeiten auf überabzählbaren Mengen, ist Gegenstand der Vorlesung *Stochastik I.*

Die Stochastik befasst sich mit der mathematisch korrekten Behandlung der Wahrscheinlichkeiten zufälliger Ereignisse. Sie gliedert sich in die beiden Teilgebiete Wahrscheinlichkeitstheorie und Statistik. Die erstere bildet das Fundament, indem sie Konzepte definiert, in denen zufällige Prozesse durch stochastische Modelle beschrieben werden, innerhalb derer man mit mathematischen Methoden Lösungen suchen kann, und letztere sucht für einen real beobachteten zufälligen Prozess ein mathematisches Modell, das diesen Prozess möglichst geeignet beschreibt. In diesem Sinne sind diese beiden Teilgebiete dual zu einander: Die Wahrscheinlichkeitstheorie behandelt die Frage: 'Gegeben ein stochastisches Modell: Was werde ich beobachten?', während die Statistik Antworten sucht auf die Frage: 'Gegeben zufällige Beobachtungen: Welches stochastische Modell steckt dahinter?'

Die Frage, was der Zufall an sich sei, ist eine offene Frage von großer philosophischer Tiefe und wird von der mathematischen Stochastik nicht berührt. Wir begnügen uns mit einer intuitiven Vorstellung vom Zufall. Zum Beispiel werden wir immer wieder Fragen betrachten, die bei einem ideellen Glücksspiel (etwa das Werfen von Würfeln oder das Ziehen von Kugeln) auftauchen, aber nicht etwa Fragen wie: 'Mit welcher Wahrscheinlichkeit gibt es Leben auf dem Pluto?'. Das Erste ist ein wohldefiniertes zufälliges Experiment, das zweite betrifft eine Aussage, die richtig ist oder nicht, deren Wahrheitsgehalt wir aber nicht kennen.

Der Stochastiker macht bei der Bildung seiner Modelle fast immer eine Reihe von idealisierenden Annahmen, die die mathematische Behandlung erst ermöglichen. Er abstrahiert und

'glättet' die Situation, um sie mathematischen Methoden zugänglich zu machen, und ignoriert dabei oft zwangsläufig Details und/oder störende Einflüsse. Ob dieses idealisierte Modell genügend geeignet erscheint, um ein gegebenes reales Problem zu modellieren, bleibt der Intuition zur Entscheidung überlassen; oft aber ist man aus mathematischen oder auch praktischen Gründen zu einer starken Vereinfachung gezwungen, um überhaupt ein praktikables Modell aufstellen zu können.

Doch ist die Theorie der Stochastik durchaus so gut entwickelt, dass eine Fülle von Modellen bekannt sind, die von sich aus eine große Vielzahl von zufälligen Prozessen adäquat beschreiben können. Außerdem ist die Theorie so flexibel, dass auch für die Beschreibung vieler anderer Prozesse hilfreiche mathematische Ansätze vorhanden sind.

Das vorliegende Skript wurde natürlich aus mehreren verschiedenen Quellen gespeist, die man nicht mehr alle im Einzelnen zurück verfolgen kann. Aber die Lehrbücher, die den meisten Einfluss auf die Gestaltung dieses Skriptes hatten, sind der klassische Text [Kr02] von U. Krengel sowie das neue Lehrbuch [Ge02] von H.-O. Georgii, das sicherlich geeignet ist, in den nächsten Jahren ebenfalls einen klassischen Status zu erlangen. Beide sind jeweils so umfangreich, dass für eine vierstündige Vorlesung natürlich streng ausgewählt werden musste.

# Kapitel 1

# Diskrete Wahrscheinlichkeitsräume

In diesem Abschnitt führen wir die grundlegenden Begriffe der diskreten Wahrscheinlichkeitstheorie ein, das heißt der Theorie der Wahrscheinlichkeiten auf höchstens abzählbar unendlich großen Mengen. Wir geben eine Auswahl an Beispielen und sammeln Eigenschaften dieses Konzeptes.

## 1.1 Grundbegriffe

Wir beginnen mit einem einführenden Beispiel.

**Beispiel 1.1.1.** Wir würfeln mit zwei fairen[1] Würfeln. Wie groß ist die Wahrscheinlichkeit, dass die Augensumme nicht kleiner als zehn ist?

Es bieten sich (mindestens) zwei Möglichkeiten der Problemlösung an. Wir können zum Beispiel ansetzen, dass jedes Element aus $\Omega = \{1, 2, \ldots, 6\}^2$ (das ist die Menge aller Zahlenpaare mit Koeffizienten zwischen 1 und 6) die selbe Wahrscheinlichkeit besitzt (also 1/36), wobei wir ein solches Paar identifizieren als die Ergebnisse der beiden Würfel. (Wir behandeln also die beiden Würfel als unterscheidbar, obwohl davon sicher die gesuchte Wahrscheinlichkeit nicht abhängen wird.) Dann zählen wir die günstigen Elementarereignisse, also diejenigen Paare, die das gesuchte Ereignis realisieren. Wir kommen auf die sechs Paare $(4, 6), (5, 5), (5, 6), (6, 4), (6, 5)$ und $(6, 6)$. Also antworten wir, dass die gesuchte Wahrscheinlichkeit gleich 6/36 sei, also 1/6.

Eine zweite Möglichkeit ist, die Menge $\Omega = \{2, 3, 4, \ldots, 12\}$ von möglichen Augensummen zu betrachten. Allerdings müssen wir beachten, dass diese elf Elementarereignisse nicht gleich wahrscheinlich sind. Zum Beispiel hat die 2 die Wahrscheinlichkeit 1/36, und die 3 hat die Wahrscheinlichkeit 1/18. Nun müssen wir also die Wahrscheinlichkeiten der 10, 11 und 12 ermitteln und sie addieren, um die gesuchte Wahrscheinlichkeit zu erhalten. Die 10 hat die Wahrscheinlichkeit 1/12, denn es gibt drei Möglichkeiten, die Summe 10 zu würfeln. Die 11 und die 12 haben die Wahrscheinlichkeiten 2/36 bzw. 1/36. Eine Addition ergibt, dass die gesuchte Wahrscheinlichkeit gleich 1/6 beträgt, also das selbe Ergebnis wie oben. ◇

Dieses Beispiel macht mehrere Dinge deutlich:

[1] Mit einem 'fairen' Würfel meinen wir einen idealiserten, der jedes der sechs Ergebnisse mit der selben Wahrscheinlichkeit $\frac{1}{6}$ ausgibt.

1. Es empfiehlt sich, eine Grundmenge von Elementarereignissen zu definieren, deren Wahrscheinlichkeiten einzeln bestimmt werden. Letzteres ist besonders einfach, wenn alle die gleiche Wahrscheinlichkeit besitzen.

2. Das Ereignis, nach dem gefragt ist, identifiziert man mit einer Teilmenge der Grundmenge. Ihre Wahrscheinlichkeit ist die Summe ihrer Einzelwahrscheinlichkeiten.

3. Es gibt im Allgemeinen mehrere Möglichkeiten, eine Grundmenge zu wählen.

Das zu Grunde liegende Konzept fassen wir wie folgt zusammen. Mit $\mathcal{P}(\Omega)$ bezeichnen wir die *Potenzmenge* einer Menge $\Omega$, also die Menge aller Teilmengen von $\Omega$.

**Definition 1.1.2 (diskreter Wahrscheinlichkeitsraum).** *Ein* diskreter Wahrscheinlichkeitsraum *ist ein Tupel $(\Omega, p)$, bestehend aus einer endlichen oder höchstens abzählbar unendlichen Menge $\Omega$ und einer Abbildung $p\colon \Omega \to [0,1]$ mit der Eigenschaft*

$$\sum_{\omega\in\Omega} p(\omega) = 1.$$

*Wir nennen $\Omega$ den* Ereignisraum*, seine Elemente die* Elementarereignisse*, seine Teilmengen die* Ereignisse *und die $p(\omega)$ die* Einzelwahrscheinlichkeiten*.*

*Die Abbildung $\mathbb{P}\colon \mathcal{P}(\Omega) \to [0,1]$, definiert durch*

$$\mathbb{P}(A) = \sum_{\omega\in A} p(\omega) \qquad \text{für alle } A \subset \Omega,$$

*heißt das von den Einzelwahrscheinlichkeiten induzierte* Wahrscheinlichkeitsmaß*. Man spricht auch von einer* Verteilung *$\mathbb{P}$ auf $\Omega$.*

Da alle Einzelwahrscheinlichkeiten nicht negativ sind, spielt in dem Ausdruck $\sum_{\omega\in\Omega} p(\omega)$ die Reihenfolge der Summanden keine Rolle. Genau genommen handelt es sich um den Grenzwert $\lim_{n\to\infty} \sum_{i=1}^{n} p(\omega_i)$, wobei $\omega_1, \omega_2, \ldots$ eine Abzählung von $\Omega$ ist.

Der ersten Lösung im Beispiel 1.1.1 lag also der Wahrscheinlichkeitsraum $(\Omega, p)$ mit $\Omega = \{1,2,\ldots,6\}^2$ und $p(\omega) = \frac{1}{36}$ für jedes $\omega \in \Omega$ zu Grunde.

Der Begriff des Wahrscheinlichkeitsraums in Definition 1.1.2 ist ein Spezialfall eines allgemeinen und fundamentalen Konzepts, das allerdings nicht in dieser Vorlesung behandelt werden wird, sondern in der Vorlesung *Stochastik I*. Die fundamentalen Eigenschaften des Wahrscheinlichkeitsmaßes $\mathbb{P}$ in Definition 1.1.2 sind die folgenden.

**Bemerkung 1.1.3 (Kolmogorovsche Axiome).** Sei $(\Omega, p)$ ein diskreter Wahrscheinlichkeitsraum. Das Wahrscheinlichkeitsmaß $\mathbb{P}$ hat die beiden Eigenschaften

(i) $\mathbb{P}(\Omega) = 1$ (Normierung),

(ii) für alle Folgen $(A_i)_{i\in\mathbb{N}}$ von paarweise disjunkten[2] Ereignissen gilt

$$\mathbb{P}\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i) \qquad \text{(abzählbare Additivität)}.$$

[2]Wir nennen Ereignisse $A_i$ mit $i \in I$ *paarweise disjunkt*, wenn $A_i \cap A_j = \emptyset$ für alle $i, j \in I$ mit $i \neq j$.

Die Gültigkeit dieser beiden Eigenschaften ist evident. ◇

Diese sogenannten Kolmogorovschen Axiome sind für uns also Folgerungen aus Definition 1.1.2. In allgemeinerem Rahmen (siehe die Vorlesung *Stochastik I*) fordert man sie üblicherweise als Axiome, muss aber zunächst definieren, was Ereignisse sein sollen. Es ist leicht zu sehen, dass jede endliche oder höchstens abzählbare Menge $\Omega$, auf deren Potenzmenge eine Abbildung $\mathbb{P}$ definiert ist, die den Kolmogorovschen Axiomen genügt, durch die Definition

$$p(\omega) = \mathbb{P}(\{\omega\}) \qquad \text{für } \omega \in \Omega$$

zu einem diskreten Wahrscheinlichkeitsraum $(\Omega, p)$ gemacht wird.

Wir sprechen also von Teilmengen von $\Omega$ als von Ereignissen. Wir listen die wichtigsten gängigen Sprechweisen für Ereignisse und ihre Entsprechung in der mathematischen Mengenschreibweise auf:

| **Sprache der Ereignisse** | **Mengenschreibweise** |
|---|---|
| $A$, $B$ und $C$ sind Ereignisse | $A, B, C \subset \Omega$ |
| $A$ und $B$ | $A \cap B$ |
| $A$ oder $B$ | $A \cup B$ |
| nicht $A$ | $A^{\mathrm{c}} = \Omega \setminus A$ |
| $A$ und $B$ schließen sich aus | $A \cap B = \emptyset$ |
| $A$ impliziert $B$ | $A \subset B$ |

Wahrscheinlichkeiten genügen einigen einfachen Regeln:

**Lemma 1.1.4.** *Sei $(\Omega, p)$ ein diskreter Wahrscheinlichkeitsraum, dann hat das zugehörige Wahrscheinlichkeitsmaß $\mathbb{P}$ die folgenden Eigenschaften:*

*(a)* $\mathbb{P}(\emptyset) = 0$,

*(b)* $A \subset B \Longrightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$ *für alle Ereignisse $A$ und $B$,*

*(c)* $\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)$ *für alle Ereignisse $A$ und $B$,*

*(d)* $\mathbb{P}(\bigcup_{i\in\mathbb{N}} A_i) \leq \sum_{i\in\mathbb{N}} \mathbb{P}(A_i)$ *für alle Folgen $(A_i)_{i\in\mathbb{N}}$ von Ereignissen,*

*(e) Falls für eine Folge $(A_i)_{i\in\mathbb{N}}$ von Ereignissen und ein Ereignis $A$ gilt:*[3] $A_i \downarrow A$ *oder* $A_i \uparrow A$, *so gilt* $\lim_{i\to\infty} \mathbb{P}(A_i) = \mathbb{P}(A)$.

**Beweis.** Die Beweise von (a) - (d) sind Übungsaufgaben. Der Beweis von (e) geht wie folgt. Falls $A_i \uparrow A$, so ist $A$ gleich der disjunkten Vereinigung der Mengen $A_i \setminus A_{i-1}$ mit $i \in \mathbb{N}$ (wobei wir $A_0 = \emptyset$ setzen), und daher gilt nach Bemerkung 1.1.3:

$$\mathbb{P}(A) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty}(A_i \setminus A_{i-1})\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i \setminus A_{i-1}) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(A_i \setminus A_{i-1}) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

Der Beweis der Aussage unter der Voraussetzung $A_i \downarrow A$ ist eine Übungsaufgabe. □

[3] Wir schreiben $A_i \downarrow A$ für $i \to \infty$, falls $A_{i+1} \subset A_i$ für jedes $i$ und $\bigcap_{i\in\mathbb{N}} A_i = A$. Analog ist $A_i \uparrow A$ definiert.

**Beispiel 1.1.5.** Wir betrachten ein Kartenspiel mit $2n$ Karten, darunter zwei Jokern. Wir bilden zwei gleich große Stapel. Wir groß ist die Wahrscheinlichkeit, dass sich beide Joker im selben Stapel befinden?

Wir benutzen den Wahrscheinlichkeitsraum $(\Omega, p)$ mit

$$\Omega = \big\{(i,j) \in \{1,2,\dots,2n\}^2 \colon i \neq j\big\} \qquad \text{und} \qquad p((i,j)) = \frac{1}{|\Omega|} = \frac{1}{2n(2n-1)}.$$

Wir interpretieren $i$ und $j$ als die Plätze der beiden Joker im Kartenspiel. Die Plätze 1 bis $n$ gehören zum ersten Stapel, die anderen zum zweiten. Also betrachten wir das Ereignis

$$A = \big\{(i,j) \in \Omega \colon i,j \leq n\big\} \cup \big\{(i,j) \in \Omega \colon i,j \geq n+1\big\}.$$

Man sieht leicht, dass $A$ genau $2n(n-1)$ Elemente besitzt. Also ist $\mathbb{P}(A) = \frac{n-1}{2n-1}$. ◇

## 1.2 Urnenmodelle

Wie aus Beispiel 1.1.1 klar wurde, ist einer der einfachsten Verteilungen durch die *Gleichverteilung* (auch *Laplace-Verteilung*) auf einer endlichen Menge $\Omega$. Hier gilt $p(\omega) = 1/|\Omega|$ für jedes $\omega \in \Omega$, wobei $|\Omega|$ die *Kardinalität* von $\Omega$ bezeichnet, also die Anzahl der Elemente in $\Omega$. Im Folgenden geben wir eine Liste von wichtigen Beispielen von Laplace-Räumen, die von Urnenmodellen her rühren. Dabei werden einige sehr wichtige Formeln der elementaren Kombinatorik motiviert und hergeleitet.

**Beispiel 1.2.1 (Urnenmodelle).** Es sei eine Urne gegeben, in der $N$ Kugeln mit den Aufschriften $1, 2, \dots, N$ liegen. Wir ziehen nun $n$ Kugeln aus der Urne. Es stellt sich die Frage, wieviele unterschiedliche Ergebnisse wir dabei erhalten können. Ein Ergebnis ist hierbei ein Tupel $(k_1, \dots, k_n)$ in $\{1, \dots, N\}^n$, wobei $k_i$ dafür stehen soll, dass in der $i$-ten Ziehung die Kugel mit der Nummer $k_i$ gezogen wird.

Die Anwort auf diese Frage hängt stark davon ab, ob wir die einzelnen Ziehungen mit oder ohne zwischenzeitlichem Zurücklegen durchführen und ob wir Ergebnisse als unterschiedlich ansehen wollen, wenn sie sich nur in der Reihenfolge unterscheiden. Mit $\mathcal{M}$ bezeichnen wir die Menge $\{1, \dots, N\}$.

*(I) mit Zurücklegen, mit Reihenfolge.* Wir legen also nach jeder Ziehung die gezogene Kugel jeweils wieder in die Urne zurück, und wir betrachten die Ziehung unterschiedlicher Kugeln in unterschiedlicher Reihenfolge als unterschiedlich. Als Modell bietet sich die Menge

$$\Omega_I = \mathcal{M}^n = \{(k_1, \dots, k_n) \colon k_1, \dots, k_n \in \mathcal{M}\},$$

die Menge aller $n$-Tupel mit Koeffizienten aus $\mathcal{M}$, an. Es ist klar, dass $\Omega_I$ genau $N^n$ Elemente besitzt.

*(II) ohne Zurücklegen, mit Reihenfolge.* Hier legen wir zwischendurch keine gezogenen Kugeln zurück, insbesondere müssen wir voraus setzen, dass $n \leq N$. Ein geeignetes Modell ist

$$\Omega_{II} = \{(k_1, \dots, k_n) \in \mathcal{M}^n \colon k_1, \dots, k_n \in \mathcal{M} \text{ paarweise verschieden}\}.$$

Die Anzahl der Elemente von $\Omega_{II}$ ist leicht aus der folgenden Argumentation als

$$|\Omega_{II}| = N(N-1)(N-2) \cdot \dots \cdot (N-n+1) = \frac{N!}{(N-n)!}$$

zu ermitteln:[4] Bei der ersten Ziehung haben wir $N$ Kugeln zur Auswahl, bei der zweiten nur noch $N-1$ und so weiter. Jede dieser insgesamt $n$ Ziehungen führt zu einem anderen $n$-Tupel.

*(III) ohne Zurücklegen, ohne Reihenfolge.* Hier legen wir keine gezogenen Kugeln zurück, und uns interessiert zum Schluss nicht, in welcher Reihenfolge wir die $n$ Kugeln gezogen haben. Also empfiehlt sich die Grundmenge

$$\Omega_{III} = \{A \subset \mathcal{M} \colon |A| = n\} = \{\{k_1, \dots, k_n\} \subset \mathcal{M} \colon k_1, \dots, k_n \text{ paarweise verschieden}\},$$

die Menge der $n$-elementigen Teilmengen von $\mathcal{M}$. Die Kardinalität von $\Omega_{II}$ kann man ermitteln, indem man sich überlegt, dass die Menge $\Omega_{II}$ alle Teilmengen aus $\Omega_{III}$ genau $n!$ Mal auflistet, nämlich als Tupel in sämtlichen möglichen Reihenfolgen. Also gilt

$$|\Omega_{III}| = \frac{|\Omega_{II}|}{n!} = \frac{N!}{(N-n)!n!} = \binom{N}{n}.$$

Den Ausdruck $\binom{N}{n}$ liest man '$N$ über $n$', er wird *Binomialkoeffizient* genannt. Eine seiner wichtigsten Bedeutungen ist, dass er die Anzahl der $n$-elementigen Teilmengen einer $N$-elementigen Menge angibt.

*(IV) mit Zurücklegen, ohne Reihenfolge.* Wir legen also die jeweils gezogene Kugel nach jeder Ziehung zurück, und am Ende ordnen wir die Aufschriften der gezogenen Kugeln und zählen, wie oft wir welche gezogen haben. In einer gewissen Abwandlung bedeutet also $k_i$ nicht mehr die Aufschrift der Kugel, die wir als $i$-te gezogen haben, sondern die $i$-t größte Aufschrift, die wir gezogen haben. Dies legt die Grundmenge

$$\Omega_{IV} = \{(k_1, \dots, k_n) \in \mathcal{M}^n \colon k_1 \le k_2 \le \dots \le k_n\},$$

nahe, die Menge der $n$-Tupel in nichtabsteigender Reihenfolge. Zur Ermittlung der Kardinalität verwenden wir einen kleinen Trick. Wir betrachten die Abbildung $(k_1, \dots, k_n) \mapsto (k'_1, \dots, k'_n)$ mit $k'_i = k_i + i - 1$. Diese Abbildung ist bijektiv von $\Omega_{IV}$ in die Menge

$$\Omega = \big\{(k'_1, \dots, k'_n) \in \mathcal{M}' \colon k'_1, \dots, k'_n \text{ paarweise verschieden}\big\},$$

wobei $\mathcal{M}' = \{1, \dots, N+n-1\}$. Offensichtlich besitzt diese Menge $\Omega$ die selbe Kardinalität wie die oben behandelte Menge $\Omega'_{III}$, wobei $\Omega'_{III}$ allerdings nicht mit $\mathcal{M} = \{1, \dots, N\}$ gebildet wird, sondern mit $\mathcal{M}'$. Nun können wir die obige Formel für die Kardinalität von $\Omega_{III}$ benutzen, nachdem wir $N$ durch $N+n-1$ ersetzt haben, und wir erhalten

$$|\Omega_{IV}| = |\Omega| = |\Omega'_{III}| = \binom{N+n-1}{n}.$$

◇

Es stellt sich heraus, dass eine Fülle von weiteren Modellen auf die obigen Urnenmodelle zurück geführt und mit Hilfe der Formeln in Beispiel 1.2.1 behandelt werden können.

**Beispiel 1.2.2.** Wir würfeln mit vier Würfeln. Wie groß ist die Wahrscheinlichkeit dafür, dass wir vier verschiedene Augenzahlen erhalten? (Übungsaufgabe) ◇

[4] Mit $n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 2 \cdot 1$ (lies '$n$ Fakultät') bezeichnet man das Produkt der ersten $n$ natürlichen Zahlen. Man sieht leicht, dass es genau $n!$ verschiedene Permutationen von $n$ unterscheidbaren Objekten gibt.

**Beispiel 1.2.3.** Wie groß ist beim Zahlenlotto '6 aus 49' die Wahrscheinlichkeit für genau sechs bzw. für genau vier Richtige? (Übungsaufgabe) ◇

**Beispiel 1.2.4 (Geburtstagszwillinge).** Wie groß ist die Wahrscheinlichkeit $p$, dass unter $n \in \mathbb{N}$ Personen keine zwei Personen am selben Tag Geburtstag haben? (Natürlich setzen wir voraus, dass das Jahr 365 Tage hat, dass $n \leq 365$ gilt und dass alle Geburtstage unabhängig von einander die gleiche Wahrscheinlichkeit haben.)

Die Menge aller Geburtstagskombinationen der $n$ Personen ist $\Omega_I$ mit $N = 365$ aus Beispiel 1.2.1, und die Menge von Kombinationen, die das gesuchte Ereignis realisieren, ist die Menge $\Omega_{II}$. Also ist die Anwort, dass die gesuchte Wahrscheinlichkeit gleich

$$p = \frac{|\Omega_{II}|}{|\Omega_I|} = 1\Big(1 - \frac{1}{N}\Big)\Big(1 - \frac{2}{N}\Big) \dots \Big(1 - \frac{n-1}{N}\Big) = \exp\Big\{\sum_{i=1}^{n-1} \log\Big(1 - \frac{i}{N}\Big)\Big\}$$

ist. Bei $N = 365$ und $n = 25$ ist dieser Wert etwa gleich $0,432$. Für allgemeine $n$ und $N$ können wir eine Approximation dieses unhandlichen Ausdrucks vornehmen, wenn $n$ sehr klein im Verhältnis zu $N$ ist. In disem Fall benutzen wir die Näherung $\log(1 + x) \approx x$ für kleine $|x|$ und erhalten

$$p \approx \exp\Big\{-\sum_{i=1}^{n-1} \frac{i}{N}\Big\} = \exp\Big\{-\frac{n(n-1)}{2N}\Big\},$$

was sich viel leichter berechnen lässt. ◇

**Beispiel 1.2.5.** Wie viele Möglichkeiten gibt es, $n$ ununterscheidbare Murmeln auf $N$ Zellen zu verteilen?

Diese Frage führt auf die Frage der Kardinalität von $\Omega_{IV}$ in Beispiel 1.2.1, und wir benutzen die Gelegenheit, eine alternative Lösung anzubieten. Wir denken uns die $n$ Murmeln in eine Reihe gelegt. Sie in $N$ Zellen einzuteilen, ist äquivalent dazu, $N - 1$ Trennwände zwischen die $n$ Murmeln zu setzen, denn dadurch werden ja die $N$ Zellen definiert. (Links von der ersten Trennwand ist der Inhalt der ersten Zelle, zwischen der $(i-1)$-ten und $i$-ten Trennwand ist der Inhalt der $i$-ten Zelle, und rechts von der $(N-1)$-ten Trennwand ist der Inhalt der letzten Zelle.) Dadurch erhalten wir eine Reihe von $N + n - 1$ Objekten: $n$ Murmeln und $N - 1$ Trennwände. Jede der $\binom{N+n-1}{n}$ möglichen Anordnungen (hier benutzen wir die Formel für $\Omega_{II}$) korrespondiert mit genau einer Möglichkeit, die Murmeln in $N$ Zellen einzuteilen. Also ist diese Anzahl gleich $\binom{N+n-1}{n}$. ◇

## 1.3 Weitere Beispiele von Verteilungen

Bei Ziehungen aus einer Urne, in der zwei verschiedene Typen von Kugeln liegen, bietet sich das folgende Modell an.

**Beispiel 1.3.1 (Hypergeometrische Verteilung).** Aus einer Urne mit $S$ schwarzen und $W$ weißen Kugeln ziehen wir $n$ Kugeln ohne Zurücklegen. Wir setzen dabei voraus, dass $S, W, n \in \mathbb{N}_0$ mit $n \leq S + W$. Wir nehmen wie immer an, dass die Kugeln in der Urne gut gemischt sind. Uns interessiert die Wahrscheinlichkeit, dass dabei genau $s$ schwarze und $w = n - s$ weiße Kugeln gezogen wurden, wobei $s, w \in \mathbb{N}_0$ mit $s \leq S$ und $w \leq W$. Das heißt, dass $s$ die Bedingungen

$\max\{0, n-W\} \le s \le \min\{S, n\}$ erfüllen muss. Die gesuchte Wahrscheinlichkeit ist gleich

$$\mathrm{Hyp}_{n,S,W}(s) = \frac{\binom{S}{s}\binom{W}{n-s}}{\binom{S+W}{n}}, \qquad s \in \big\{\max\{0, n-W\}, \dots, \min\{S, n\}\big\}.$$

Dies ergibt sich aus einer mehrfachen Anwendung der Formel für $|\Omega_{II}|$ in Beispiel 1.2.1: Der Nenner ist die Anzahl der Möglichkeiten, $n$ Kugeln aus $S+W$ auszuwählen, und im Zähler steht die Anzahl der Möglichkeiten, $s$ Kugeln aus $S$ auszuwählen und $n-s$ aus $W$. Wir können die Menge $\Omega = \{\max\{0, n-W\}, \dots, \min\{S, n\}\}$ zusammen mit $\mathrm{Hyp}_{n,S,W}$ als einen Wahrscheinlichkeitsraum auffassen. Der Nachweis, dass $\mathrm{Hyp}_{n,S,W}$ wirklich auf $\Omega$ normiert ist, ist etwas unangenehm; wir begnügen uns mit der probabilistischen Argumentation, dass wir die Summe der Wahrscheinlichkeiten aller möglichen Ereignisse erhalten, indem wir $\mathrm{Hyp}_{n,S,W}(s)$ über $s \in \Omega$ summieren. Man nennt $\mathrm{Hyp}_{n,S,W}$ die *hypergeometrische Verteilung* auf $\Omega$ mit Parametern $n$, $S$ und $W$. ◇

**Beispiel 1.3.2.** Wie groß ist die Wahrscheinlichkeit, dass nach dem Geben beim Skatspiel Spieler $A$ genau drei Asse besitzt? (Wir gehen davon aus, dass 32 Karten an drei Spieler, die je zehn Karten erhalten, und den Skat verteilt werden, und dass genau vier Asse unter den 32 Karten sind. Natürlich nehmen wir wieder an, dass das Spiel gut gemischt ist.)

Um diese Frage zu beantworten, benutzen wir die hypergeometrische Verteilung mit $S = 4$, $W = 28$ und $n = 10$ und $s = 3$, denn Spieler $A$ erhält zehn von 32 Karten, und es gibt vier Asse und 28 andere Karten im Spiel. Die gesuchte Wahrscheinlichkeit ist also

$$\mathrm{Hyp}_{10,4,28}(3) = \frac{\binom{4}{3}\binom{28}{7}}{\binom{32}{10}} = \frac{66}{899}.$$

◇

Eine der wichtigsten Verteilungen, die uns in dieser Vorlesung immer wieder begegnen wird, taucht auf bei der Frage nach Erfolgszahlen bei Serien von Glücksspielen.

**Beispiel 1.3.3 (Bernoulliverteilung).** Wir spielen ein Glücksspiel, bei dem es mit Wahrscheinlichkeit $p \in [0, 1]$ einen Erfolg gibt und sonst keinen. Insgesamt spielen wir es $n$ Mal, und alle $n$ Spiele sind unabhängig voneinander. (Zur Präzisierung dieses Begriffs siehe Abschnitt 2.2, insbesondere Beispiel 2.2.3.) Man sagt, wir führen ein *Bernoulli-Experiment* der Länge $n$ mit Erfolgswahrscheinlichkeit $p$ durch. Dieses Experiment können wir mit dem Wahrscheinlichkeitsraum $(\Omega, q)$ gut beschreiben, wobei $\Omega = \{0, 1\}^n$ (die Menge aller Vektoren der Länge $n$ mit Koeffizienten 0 oder 1, wobei '1' für 'Erfolg' steht und '0' für 'Misserfolg'), und die Einzelwahrscheinlichkeiten sind gegeben durch

$$q(\omega) = p^{\sum_{i=1}^n \omega_i}(1-p)^{n-\sum_{i=1}^n \omega_i}, \qquad \omega = (\omega_1, \dots, \omega_n) \in \Omega.$$

(Man beachte, dass $\sum_{i=1}^n \omega_i$ geich der Anzahl der Einsen im Vektor $\omega$ ist und $n - \sum_{i=1}^n \omega_i$ die Anzahl der Nullen.) Den Nachweis, dass $q$ tatsächlich eine Wahrscheinlichkeitsverteilung auf $\Omega$ induziert (d. h., dass die $q(\omega)$ sich zu Eins aufsummieren), führt man am besten über eine vollständige Induktion. ◇

**Beispiel 1.3.4 (Binomialverteilung).** Wir führen ein Bernoulli-Experiment der Länge $n$ mit Erfolgswahrscheinlichkeit $p \in [0, 1]$ durch wie in Beispiel 1.3.3. Wir groß ist die Wahrscheinlichkeit, genau $k$ Erfolge zu haben, wobei $k \in \{0, \dots, n\}$? Formaler gefragt, wie groß ist die Wahrscheinlichkeit des Ereignisses $A_k = \{\omega \in \Omega \colon \sum_{i=1}^n \omega_i = k\}$?

Die Lösung anzugeben fällt uns nicht schwer. Es gibt $\binom{n}{k}$ Anzahlen von Spielverläufen, in denen es genau $k$ Erfolge gibt, und die $k$ Erfolge geschehen mit Wahrscheinlichkeit $p^k$, und die $n-k$ Misserfolge mit Wahrscheinlichkeit $(1-p)^{n-k}$. Also ist die gesuchte Wahrscheinlichkeit gegeben durch die Formel

$$\mathrm{Bi}_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k \in \{0, \dots, n\}.$$

Die Verteilung auf $\Omega = \{0, \dots, n\}$ mit den Einzelwahrscheinlichkeiten $\mathrm{Bi}_{n,p}$ heißt die *Binomialverteilung* mit Parametern $p \in [0,1]$ und $n \in \mathbb{N}$. Wir erinnern an den Binomischen Lehrsatz

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}, \qquad x, y \in \mathbb{R}, n \in \mathbb{N}, \tag{1.3.1}$$

der eine analytische Rechtfertigung dafür liefert, dass $\mathrm{Bi}_{n,p}$ tatsächlich eine Wahrscheinlichkeitsverteilung induziert. Ohne Probleme leitet man die Beziehungen

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n, \qquad \sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0, \qquad \sum_{k=0}^{n} k \binom{n}{k} = n 2^{n-1}, \qquad \binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1} \tag{1.3.2}$$

her (Übungsaufgabe). ◇

In den folgenden beiden Beispielen führen wir die wohl wichtigsten Verteilungen auf einem abzählbar *unendlichen* Raum ein.

**Beispiel 1.3.5 (Geometrische Verteilung).** Wir stellen uns vor, dass wir das Glücksspiel aus Beispiel 1.3.4 so lange spielen, bis wir zum ersten Male einen Erfolg verbuchen können. Mit welcher Wahrscheinlichkeit passiert dies beim $k$-ten Spiel (wobei $k \in \mathbb{N}$)?

Auch diese Frage beantworten wir leicht. Das gesuchte Ereignis ist dadurch charakterisiert, dass wir genau $k-1$ Mal hinter einander keinen Erfolg haben und im $k$-ten Spiel endlich einen. Also ist die gesuchte Wahrscheinlichkeit gleich

$$\mathrm{Geo}_p(k) = p(1-p)^{k-1}, \qquad k \in \mathbb{N}. \tag{1.3.3}$$

Die Verteilung auf $\Omega = \mathbb{N}$ mit den Einzelwahrscheinlichkeiten $\mathrm{Geo}_p$ heißt *geometrische Verteilung* mit Parameter $p \in [0,1]$. Sie ist also die Verteilung der Wartezeit auf den ersten Erfolg in einer Serie von Glücksspielen, die jeweils mit Erfolgswahrscheinlichkeit $p$ ablaufen. Manche Autoren definieren die geometrische Verteilung auf der Menge $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ mittels der Formel $\mathrm{Geo}_p(k) = p(1-p)^k$ für $k \in \mathbb{N}_0$.

Die Interpretation als Wartezeit auf den ersten Erfolg in einer möglicherweise unendlich langen Serie von Glücksspielen scheint den überabzählbaren Wahrscheinlichkeitsraum $\Omega = \{0,1\}^{\mathbb{N}}$ zu erfordern, die Menge aller unendlich langen 0-1-Folgen. Das ist allerdings tatsächlich nicht nötig, da die Betrachtung des Ereignisses, dass das $k$-te Spiel den ersten Erfolg bringt, nur den endlichen Wahrscheinlichkeitsraum $\Omega = \{0,1\}^k$ erfordert. ◇

**Beispiel 1.3.6 (Poisson-Verteilung).** Mit einem Parameter $\alpha \in (0, \infty)$ betrachten wir die Verteilung auf $\Omega = \mathbb{N}_0 = \{0, 1, 2, \dots\}$, die gegeben ist durch

$$\mathrm{Po}_\alpha(k) = e^{-\alpha} \frac{\alpha^k}{k!}, \qquad k \in \mathbb{N}_0. \tag{1.3.4}$$

Die Verteilung $\mathrm{Po}_\alpha$ heißt die *Poisson-Verteilung* zum Parameter $\alpha$. Eine ihrer wichtigsten Bedeutungen und Anwendungen kommt von der folgenden Überlegung her. Wir denken uns, dass im Zeitintervall $(0,t]$ eine zufällige Anzahl von Zeitpunkten zufällig verteilt sind. Um dies zu präzisieren, müssen wir ein paar zusätzliche, plausible Annahmen machen. Wir teilen das Intervall in $n$ gleiche Stücke der Länge $t/n$ ein und nehmen an, dass jedes dieser Teilintervalle für großes $n$ höchstens einen dieser zufälligen Punkte enthält. Außerdem nehmen wir an, dass die Entscheidung, ob die $n$ Teilintervalle jeweils einen solchen Zeitpunkt enthalten, *unabhängig* fällt (ein Begriff, den wir in Abschnitt 2.2 präzisieren werden), und zwar mit Wahrscheinlichkeit $p_n$. Mit anderen Worten, wir nehmen an, dass wir es mit einem Bernoulli-Experiment der Länge $n$ mit Erfolgswahrscheinlichkeit $p_n$ zu tun haben. Den Parameter $p = p_n$ legen wir so fest, dass für großes $n$ die erwartete Gesamtzahl der zufälligen Punkte im Intervall $(0,t]$ (dies ist die Zahl $np_n$, was in Abschnitt 3.3 präzisiert werden wird) ungefähr ein fester Wert $\alpha \in (0,\infty)$ ist. Der folgende Satz sagt, dass die Verteilung der Anzahl der zufälligen Punkte in $(0,t]$ dann etwa durch die Poisson-Verteilung gegeben ist:

**Satz 1.3.7 (Poissonapproximation der Binomialverteilung; Poissonscher Grenzwertsatz).** *Falls* $\lim_{n\to\infty} np_n = \alpha$*, so gilt für jedes* $k \in \mathbb{N}_0$

$$\lim_{n\to\infty} \mathrm{Bi}_{n,p_n}(k) = \mathrm{Po}_\alpha(k).$$

**Beweis.** Wir benutzen die (gebräuchliche) Notation $a_n \sim b_n$, wenn $\lim_{n\to\infty} a_n/b_n = 1$. Also sieht man, dass

$$\binom{n}{k} = \frac{n^k}{k!}\frac{n(n-1)(n-2)\cdot\dots\cdot(n-k+1)}{n^k} \sim \frac{n^k}{k!}, \tag{1.3.5}$$

da wir den zweiten Bruch als Produkt von $k$ Brüchen schreiben können, die jeweils gegen Eins konvergieren. Nun benutzen wir die Tatsache $\lim_{t\to 0}(1+t)^{1/t} = e$, um zu schlussfolgern, dass

$$\begin{aligned}\mathrm{Bi}_{n,p_n}(k) &= \binom{n}{k} p_n^k (1-p_n)^n (1-p_n)^{-k} \sim \frac{n^k}{k!} p_n^k \big[(1-p_n)^{1/p_n}\big]^{np_n} \sim \frac{(np_n)^k}{k!} e^{-np_n} \\ &\to \frac{\alpha^k}{k!} e^{-\alpha} = \mathrm{Po}_\alpha(k).\end{aligned}$$

□◇

**Beispiel 1.3.8 (Verallgemeinerte hypergeometrische Verteilung und Multinomialverteilung).** In einem Teich gibt es $N$ Fische von $k$ unterschiedlichen Sorten, und zwar genau $N_i$ Stück von Sorte $i \in \{1,\dots,k\}$, wobei $N = N_1 + \dots + N_k$. Wir fischen $n \in \{0,\dots,N\}$ Fische zufällig aus dem Teich heraus. Mit welcher Wahrscheinlichkeit fangen wir dabei genau $n_i \in \mathbb{N}_0$ Fische von der Sorte $i$ für jedes $i \in \{1,\dots,k\}$? Wir setzen natürlich voraus, dass $n = n_1 + \dots + n_k$ und $n_i \in \{0,\dots,N_i\}$ für jedes $i$.

Die Antwort ist offensichtlich (bei 'guter Mischung' der Fische im Teich) gegeben durch die Formel

$$\mathrm{Hyp}_{n;N_1,\dots,N_k}(n_1,\dots,n_k) = \frac{\prod_{i=1}^k \binom{N_i}{n_i}}{\binom{N}{n}}.$$

Die Verteilung $\mathrm{Hyp}_{n;N_1,\dots,N_k}$ auf der Menge

$$\Omega_n = \big\{(n_1,\dots,n_k) \in \mathbb{N}_0^k \colon n_1 + \dots + n_k = n\big\}$$

heißt die *verallgemeinerte hypergeometrische Verteilung* mit Parametern $n$ und $N_1, \dots, N_k$. Man kann $\Omega_n$ auffassen als die Menge aller Zerlegungen einer Menge von $n$ ununterscheidbaren Objekten in $k$ Teilmengen, siehe auch Beispiel 1.2.5.

Falls man annehmen kann, dass sich von jeder Sorte sehr viele Fische im Teich befinden, dann kann man die verallgemeinerte hypergeometrische Verteilung auch approximieren mit einer leichter handhabbaren Verteilung. Wir nehmen dabei an, dass $N_1, \dots, N_k$ so groß sind, dass die obige Wahrscheinlichkeit im Wesentlichen nur noch von den relativen Anteilen $N_i/N$ abhängt. Dann kann man in etwa davon ausgehen, bei jedem Fang eines einzelnen Fisches die Sorte $i$ mit Wahrscheinlichkeit $p_i = N_i/N$ zu erwischen, unabhängig davon, wieviele und welche Fische man schon vorher gefischt hat. Dies wird im folgenden Satz präzisiert.

**Satz 1.3.9 (Multinomialapproximation).** *Für jedes $n \in \mathbb{N}$ und alle $n_1, \dots, n_k \in \mathbb{N}$ mit $n_1 + \dots + n_k = n$ und für alle $p_1, \dots, p_k \in [0,1]$ mit $p_1 + \dots + p_k = 1$ gilt*

$$\lim_{\substack{N, N_1, \dots, N_k \to \infty \\ N_i/N \to p_i \forall i}} \mathrm{Hyp}_{n;N_1,\dots,N_k}(n_1, \dots, n_k) = \mathrm{Mul}_{n;p_1,\dots,p_k}(n_1, \dots, n_k) = \binom{n}{n_1, \dots, n_k} \prod_{i=1}^{k} p_i^{n_i}.$$

Der *Multinomialkoeffizient*

$$\binom{n}{n_1, \dots, n_k} = \frac{n!}{n_1!\, n_2! \dots n_k!}, \qquad \text{für } n_1 + \dots + n_k = n,$$

gibt offensichtlich die Anzahl der Möglichkeiten an, eine Menge von $n$ unterscheidbaren Objekten in $k$ Teilmengen mit $n_1, n_2, \dots, n_k$ Elementen zu zerlegen. Die in Satz 1.3.9 eingeführte Verteilung $\mathrm{Mul}_{n;p_1,\dots,p_k}$ auf $\Omega_n$ heißt die *Multinomialverteilung* mit Parametern $p_1, \dots, p_k$. Sie beschreibt die Wahrscheinlichkeit, genau $n_i$ Mal das Ergebnis $i$ zu erhalten für jedes $i \in \{1, \dots, k\}$, wenn $n$ Mal hintereinander ein Zufallsexperiment mit den möglichen Ausgängen $1, \dots, k$, die jeweils die Wahrscheinlichkeit $p_1, \dots, p_k$ haben, durchgeführt wird. Daher ist die Multinomialverteilung eine Verallgemeinerung der Binomialverteilung auf mehr als zwei mögliche Ausgänge des Experiments. Der *Multinomialsatz*

$$(x_1 + \dots + x_k)^n = \sum_{\substack{n_1, \dots, n_k \in \mathbb{N}_0 \\ n_1 + \dots + n_k = n}} \binom{n}{n_1, \dots, n_k} \prod_{i=1}^{k} x_i^{n_i} \qquad \text{für alle } x_1, \dots, x_k \in \mathbb{R},$$

liefert die Begründung, dass $\mathrm{Mul}_{n;p_1,\dots,p_k}$ tatsächlich eine Wahrscheinlichkeitsverteilung auf $\Omega_n$ induziert.

**Beweis von Satz 1.3.9.** Man sieht leicht wie in (1.3.5), dass im Grenzwert $N_i \to \infty$ gilt

$$\binom{N_i}{n_i} \sim \frac{N_i^{n_i}}{n_i!},$$

wobei wir (wie allgemein üblich) $a \sim b$ schreiben, wenn der Quotient von $a$ und $b$ gegen 1 strebt. Also erhalten wir für den Grenzwert $N, N_1, \dots, N_k \to \infty$ mit $N_i/N \to p_i$ für alle $i$:

$$\mathrm{Hyp}_{n;N_1,\dots,N_k}(n_1, \dots, n_k) \sim \frac{\prod_{i=1}^{k} \frac{N_i^{n_i}}{n_i!}}{\frac{N^n}{n!}} = \binom{n}{n_1, \dots, n_k} \prod_{i=1}^{k} \Big(\frac{N_i}{N}\Big)^{n_i} \sim \binom{n}{n_1, \dots, n_k} \prod_{i=1}^{k} p_i^{n_i}.$$

□◇

# Kapitel 2

# Bedingte Wahrscheinlichkeit und Unabhängigkeit

In vielen Situationen liegt schon eine gewisse Information vor, wenn man die Wahrscheinlichkeit eines Ereignisses bestimmen möchte. Bei Abschluss einer Lebensversicherung kennt man schon das aktuelle Alter des Antragstellers, beim Skatspiel kennt man schon die eigenen Karten, bei Meinungsumfragen hat man vor der eigentlichen Befragung schon die Bevölkerungsgruppe des oder der Befragten festgestellt. Das heißt, dass man über das Eintreten oder Nichteintreten eines Ereignisses $B$ schon informiert ist, wenn man die Wahrscheinlichkeit eines Ereignisses $A$ bestimmen will. In diesem Abschnitt wollen wir mathematisch präzisieren, was eine bedingte Wahrscheinlichkeit von $A$ gegeben $B$ ist, und was es heißen soll, dass die Ereignisse $A$ und $B$ unabhängig sind.

Dem gesamten Kapitel legen wir einen diskreten Wahrscheinlichkeitsraum $(\Omega, p)$ zu Grunde.

## 2.1 Bedingte Wahrscheinlichkeiten

Wir beginnen wieder mit einem einführenden Beispiel.

**Beispiel 2.1.1 (Umfrage).** In einer Meinungsumfrage soll der Anzahl der Raucher an der Bevölkerung fest gestellt werden, das heißt die Wahrscheinlichkeit des Ereignisses $A$, dass eine zufällig heraus gegriffene Person Raucher ist. Hierbei möchte man allerdings mehrere Bevölkerungsgruppen unterscheiden, zum Beispiel die Gruppe der 21- bis 30jährigen Frauen, was das Ereignis $B$ darstellen soll. Man möchte also die bedingte Wahrscheinlichkeit von $A$ unter der Voraussetzung, dass $B$ eintritt, feststellen. Dazu wird man normalerweise die Anzahl der rauchenden 21- bis 30jährigen Frauen durch die Anzahl der 21- bis 30jährigen Frauen teilen, d.h., den Anteil der Raucherinnen unter den 21- bis 30jährigen Frauen bestimmen. Dies führt auf die plausible Formel

$$\mathbb{P}(\text{Raucher(in)} \mid \text{21- bis 30jährige Frau}) = \frac{|\{\text{21- bis 30jährige Raucherinnen}\}|}{|\{\text{21- bis 30jährige Frauen}\}|} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

◇

Dieses Beispiel nehmen wir nun als Anlass zu einer allgemeinen Definition. Wir müssen allerdings beachten, dass die obige Formel Gebrauch davon macht, dass wir es mit einer Gleichverteilung zu tun haben.

**Definition 2.1.2 (bedingte Wahrscheinlichkeit).** *Es seien $A$ und $B$ zwei Ereignisse mit $\mathbb{P}(B) > 0$. Mit*

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

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

**Beispiel 2.1.3.** Es sei $p_k$ die Wahrscheinlichkeit, dass man während des $k$-ten Lebensjahrs stirbt. (Hier machen wir natürlich wie üblich eine ganze Reihe an stark vereinfachenden Annahmen, z. B. dass diese Wahrscheinlichkeit nicht abhängt vom Geschlecht, dem Geburtsjahr, dem Wohnort, dem Beruf etc.) Dann ist $s_k = p_k + p_{k+1} + p_{k+2} + \dots$ die Wahrscheinlichkeit, dass man das Alter $k$ erreicht. Wenn man für einen Zeitgenossen die Wahrscheinlichkeit, im $k$-ten Lebensjahr zu sterben, ermitteln möchte, sollte man beachten, dass dieser gerade lebt und das Alter $l$ hat (mit $l \leq k$). Also sollte man tatsächlich die *bedingte* Wahrscheinlichkeit gegeben, dass er gerade $l$ Jahre alt ist, angeben, und dies ist der Quotient $p_k/s_l$. ◇

Wir sammeln ein paar einfache, aber wichtige Eigenschaften der bedingten Wahrscheinlichkeit.

**Lemma 2.1.4 (totale Wahrscheinlichkeit, Bayessche Formel).** *Es sei $B$ ein Ereignis mit $\mathbb{P}(B) > 0$.*

*(i)* $\mathbb{P}(\cdot \mid B)$ *erfüllt die Kolmogorovschen Axiome (siehe Bemerkung 1.1.3), d. h. es gilt* $\mathbb{P}(\Omega \mid B) = 1$, *und für alle Folgen* $(A_i)_{i\in\mathbb{N}}$ *von paarweise disjunkten Ereignissen gilt* $\mathbb{P}(\bigcup_{i\in\mathbb{N}} A_i \mid B) = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i \mid B)$.

*(ii) Es gilt die* Formel von der totalen Wahrscheinlichkeit*: Für jedes Ereignis $A$ und jede Folge $(B_i)_{i\in\mathbb{N}}$ von paarweise disjunkten Ereignissen mit $B = \bigcup_{i\in\mathbb{N}} B_i$ und $\mathbb{P}(B_i) > 0$ für alle $i \in \mathbb{N}$ gilt*

$$\mathbb{P}(A \cap B) = \sum_{i\in\mathbb{N}} \mathbb{P}(B_i)\mathbb{P}(A \mid B_i).$$

*(iii) Es gilt die* Formel von Bayes*: Für jedes Ereignis $A$ mit $\mathbb{P}(A) > 0$ und jede Folge $(B_i)_{i\in\mathbb{N}}$ von paarweise disjunkten Ereignissen mit $\Omega = \bigcup_{i\in\mathbb{N}} B_i$ und $\mathbb{P}(B_i) > 0$ für alle $i \in \mathbb{N}$ gilt*

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(B_i)\mathbb{P}(A \mid B_i)}{\sum_{j\in\mathbb{N}} \mathbb{P}(B_j)\mathbb{P}(A \mid B_j)}, \qquad i \in \mathbb{N}.$$

**Beweis.** einfache Übungsaufgabe. □

Die folgende Anwendung der Bayesschen Formel zeigt, dass man nicht zu alarmiert sein muss, wenn ein nicht ganz hundertprozentiger Test auf eine seltene Krankheit anspricht.

**Beispiel 2.1.5 (Test auf eine seltene Krankheit).** Eine seltene Krankheit liegt bei ungefähr 0,5 Prozent der Bevölkerung vor. Es gibt einen (recht guten) Test auf diese Krankheit, der bei 99 Prozent der Kranken anspricht, aber auch bei 2 Prozent der Gesunden. Mit welcher Wahrscheinlichkeit ist eine getestete Person tatsächlich krank, wenn der Test bei ihr angesprochen hat?

Wir teilen also die Gesamtmenge $\Omega$ aller getesteten Personen ein in die Ereignisse $B_1$ der kranken und $B_2$ der gesunden getesteten Personen. Es sei $A$ das Ereignis, dass die getestete Person auf den Test anspricht. Gesucht ist die bedingte Wahrscheinlichkeit $\mathbb{P}(B_1 \mid A)$. Dem Aufgabentext entnimmt man die Angaben

$$\mathbb{P}(B_1) = 0,005, \qquad \mathbb{P}(A \mid B_1) = 0,99, \qquad \mathbb{P}(A \mid B_2) = 0,02.$$

Insbesondere ist natürlich $\mathbb{P}(B_2) = 0,995$. Nun können wir die Bayessche Formel benutzen, um die Aufgabe zu lösen:

$$\mathbb{P}(B_1 \mid A) = \frac{\mathbb{P}(B_1)\mathbb{P}(A \mid B_1)}{\mathbb{P}(B_1)\mathbb{P}(A \mid B_1) + \mathbb{P}(B_2)\mathbb{P}(A \mid B_2)} = \frac{0,005 \cdot 0,99}{0,005 \cdot 0,99 + 0,995 \cdot 0,02} = \frac{495}{2485} \approx 0,2.$$

Also ist trotz positiven Testergebnisses die Wahrscheinlichkeit, diese Krankheit zu haben, nicht alarmierend hoch. Man sollte unter Beobachtung bleiben. ◇

In manchen Fällen erweist sich die folgend vorgestellte Formel als nützlich.

**Lemma 2.1.6 (Multiplikationsformel).** *Für jedes $n \in \mathbb{N}$ und alle Ereignisse $A_1, \ldots, A_n \subset \Omega$ mit $\mathbb{P}(A_1 \cap \cdots \cap A_{n-1}) \neq 0$ gilt*

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2) \ldots \mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

**Beweis.** Klar. □

**Beispiel 2.1.7.** Mit welcher Wahrscheinlichkeit besitzt beim Skat jeder der drei Spieler nach dem Geben genau ein Ass?

Wir verteilen also zufällig 32 Karten, darunter vier Asse, an drei Spieler, die je zehn Karten erhalten, und den Skat. Es sei $A_i$ das Ereignis, dass der $i$-te Spieler genau ein Ass erhält. Mit Hilfe der Multiplikationsformel aus Lemma 2.1.6 errechnen wir:

$$\begin{aligned}\mathbb{P}(A_1 \cap A_2 \cap A_3) &= \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2) \\ &= \frac{\binom{4}{1}\binom{28}{9}}{\binom{32}{10}} \times \frac{\binom{3}{1}\binom{19}{9}}{\binom{22}{10}} \times \frac{\binom{2}{1}\binom{10}{9}}{\binom{12}{10}} = 10^3 \frac{2 \cdot 4!}{32 \cdot 31 \cdot 30 \cdot 29} \approx 0,0556.\end{aligned}$$

◇

## 2.2 Unabhängigkeit von Ereignissen

In diesem Abschnitt präzisieren wir, was es heißt, dass zwei oder mehr Ereignisse unabhängig sind. Intuitiv bedeutet Unabhängigkeit der Ereignisse $A$ und $B$, dass das Eintreffen oder Nichteintreffen von $A$ nicht beeinflusst wird vom Eintreffen oder Nichteintreffen von $B$. Ein elementares Beispiel, in dem man den Unterschied zwischen Abhängigkeit und Unabhängigkeit gut sehen kann, ist das Ziehen zweier Kugeln aus einer Urne ohne Zurücklegen im Vergleich zum Ziehen mit Zurücklegen: Das Ziehen der ersten Kugel sollte das Ergebnis der zweiten Ziehung beeinflussen, wenn man die erste nicht wieder zurück legt, aber nicht, wenn man sie zurück legt und damit den Zustand, der vor der ersten Ziehung vorlag, wieder her stellt.

Wie formalisiert man aber Unabhängigkeit zweier Ereignisse $A$ und $B$? Eine gute Idee ist es, zu fordern, dass die Wahrscheinlichkeit von $A$ überein stimmen muss mit der bedingten

Wahrscheinlichkeit von $A$ gegeben $B$ (zumindest, falls $\mathbb{P}(B) > 0$), also $\mathbb{P}(A) = \mathbb{P}(A \mid B)$. Dies ist auch tatsächlich eine plausible Formel, doch sollte man in der Definition für Symmetrie sorgen. Ferner brauchen wir einen tragfähigen Begriff der Unabhängigkeit mehrerer Ereignisse:

**Definition 2.2.1 (Unabhängigkeit).** *(i) Zwei Ereignisse $A$ und $B$ heißen* unabhängig, *falls gilt:* $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

*(ii) Eine Familie $(A_i)_{i\in I}$ (wobei $I$ eine beliebige Indexmenge ist) heißt* unabhängig, *falls für jede endliche Teilmenge $J$ von $I$ die folgende* Produktformel *gilt:*

$$\mathbb{P}\Big(\bigcap_{i\in J} A_i\Big) = \prod_{i\in J} \mathbb{P}(A_i). \tag{2.2.1}$$

**Beispiel 2.2.2.** Beim $n$-fachen Wurf mit einem fairen Würfel (also $\Omega = \{1,\dots,6\}^n$ mit der Gleichverteilung) sind die $n$ Ereignisse '$i$-ter Wurf zeigt $a_i$' für $i = 1,\dots,n$ bei beliebigen $a_1,\dots,a_n \in \{1,\dots,6\}$ unabhängig, denn ihre jeweilige Wahrscheinlichkeit ist $\frac{1}{6}$, und die Wahrscheinlichkeit eines beliebigen Schnittes dieser Ereignisse ist $\frac{1}{6}$ hoch die Anzahl der geschnittenen Mengen. ◇

**Beispiel 2.2.3 (Bernoulli-Experiment).** Das in Beispiel 1.3.3 eingeführte Bernoulli-Experiment besteht tatsächlich aus $n$ unabhängigen Experimenten, die jeweils Erfolgswahrscheinlichkeit $p$ haben. Mit anderen Worten, die Ereignisse $A_i = \{\omega \in \Omega\colon \omega_i = 1\}$ ('$i$-tes Spiel hat Erfolg') sind unabhängig für $i = 1,\dots,n$. Der Nachweis der Unabhängigkeit ist eine Übungsaufgabe. ◇

**Beispiel 2.2.4.** Beim zweimaligen Ziehen je einer Kugel aus einer gut gemischten Urne mit $s$ schwarzen und $w$ weißen Kugeln sind die Ereignisse 'erste Kugel ist weiß' und 'zweite Kugel ist weiß' unabhängig, wenn nach dem ersten Ziehen die gezogene Kugel zurück gelegt wurde, aber nicht unabhängig, wenn dies nicht geschah. (Übungsaufgabe: Man beweise dies.) Dies unterstreicht, dass Unabhängigkeit nicht nur eine Eigenschaft der Ereignisse ist, sondern auch entscheidend vom Wahrscheinlichkeitsmaß abhängt. ◇

**Bemerkung 2.2.5.** (a) Es ist klar, dass jede Teilfamilie einer Familie von unabhängigen Ereignissen unabhängig ist.

(b) Die Forderung, dass die Produktformel in (2.2.1) für jede endliche Teilauswahl $J$ gelten soll (und nicht nur für $J = I$ oder nur für alle zweielementigen Teilauswahlen) ist sehr wesentlich. Der (weitaus weniger wichtige) Begriff der *paarweisen Unabhängigkeit* fordert die Produktformel nur für die zweielementigen Teilmengen $J$ von $I$. Im folgenden Beispiel haben wir drei paarweise unabhängige Ereignisse, die nicht unabhängig sind: In $\Omega = \{1,2,3,4\}$ mit der Gleichverteilung betrachten wir die Ereignisse $A_1 = \{1,2\}$, $A_2 = \{2,3\}$ und $A_3 = \{3,1\}$. (Übungsaufgabe: Verifiziere, dass $A_1$, $A_2$ und $A_3$ paarweise unabhängig, aber nicht unabhängig sind.)

(c) Unabhängigkeit ist keine Eigenschaft von *Mengen* von Ereignissen, sondern von *Tupeln* von Ereignissen, die allerdings nicht von der Reihenfolge abhängt. Dies ist wichtig, wenn z. B. eines der Ereignisse in dem betrachteten Tupel mehrmals auftritt. Falls das Paar $(A, A)$ unabhängig ist, so folgt $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A) = \mathbb{P}(A)^2$, also $\mathbb{P}(A) \in \{0,1\}$.

(d) Unabhängigkeit kann sogar trotz Kausalität vorliegen, wie das folgende Beispiel zeigt. Beim Würfeln mit zwei fairen Würfeln (also $\Omega = \{1,\dots,6\}^2$ mit der Gleichverteilung)

betrachten wir die Ereignisse $A = \{\text{Augensumme ist } 7\} = \{(i,j) \in \Omega\colon i+j=7\}$ und $B = \{\text{erster Würfel zeigt } 6\} = \{(i,j) \in \Omega\colon i=6\}$. Dann gilt $\mathbb{P}(A \cap B) = \mathbb{P}(\{(6,1)\}) = \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = \mathbb{P}(A)\mathbb{P}(B)$. Also sind $A$ und $B$ unabhängig, obwohl offensichtlich eine Kausalität zwischen diesen beiden Ereignissen vorliegt.

$\diamond$

Ein hilfreiches Kriterium für Unabhängigkeit von Ereignissen ist das Folgende. Es sagt, dass die Unabhängigkeit von $n$ Ereignissen gleichbedeutend ist mit der Gültigkeit der Produktformel in (2.2.1) für $J = \{1,\dots,n\}$, aber man muss für jedes der Ereignisse auch das Komplement zulassen können. Zur bequemsten Formulierung benötigen wir die Notation $A^1 = A$ für Ereignisse $A$, während $A^{\mathrm{c}}$ wie üblich das Komplement von $A$ ist.

**Lemma 2.2.6.** *Ereignisse $A_1,\dots,A_n$ sind genau dann unabhängig, wenn für alle $k_1,\dots,k_n \in \{1,\mathrm{c}\}$ gilt:*

$$\mathbb{P}\Big(\bigcap_{i=1}^{n} A_i^{k_i}\Big) = \prod_{i=1}^{n} \mathbb{P}(A_i^{k_i}). \tag{2.2.2}$$

**Beweis.**

'$\Longrightarrow$': Wir setzen die Unabhängigkeit der $A_1,\dots,A_n$ voraus und zeigen (2.2.2) mit einer Induktion nach $n$. Für $n=1$ ist nichts zu zeigen. Wir nehmen nun den Schluss von $n$ auf $n+1$ vor und setzen voraus, dass $A_1,\dots,A_{n+1}$ unabhängig sind. Nun zeigen wir (2.2.2) für $n+1$ statt $n$ mit einer weiteren Induktion über die Anzahl $m$ der 'c' in $k_1,\dots,k_{n+1}$.

Für $m=0$ folgt die Behauptung aus der Definition der Unabhängigkeit der $A_1,\dots,A_{n+1}$. Nun machen wir den Induktionsschluss von $m \in \{0,\dots,n-1\}$ auf $m+1$: Es seien $m+1$ Komplementzeichen in $k_1,\dots,k_{n+1}$. Durch eine geeignete Permutation der Ereignisse können wir annehmen, dass $k_{n+1} = \mathrm{c}$ ist. Dann haben wir

$$\mathbb{P}\Big(\bigcap_{i=1}^{n+1} A_i^{k_i}\Big) = \mathbb{P}\Big(\bigcap_{i=1}^{n} A_i^{k_i} \cap A_{n+1}^{\mathrm{c}}\Big) = \mathbb{P}\Big(\bigcap_{i=1}^{n} A_i^{k_i}\Big) - \mathbb{P}\Big(\bigcap_{i=1}^{n} A_i^{k_i} \cap A_{n+1}\Big).$$

Der erste Summand ist nach Induktionsvoraussetzung an $n$ gleich $\prod_{i=1}^{n} \mathbb{P}(A_i^{k_i})$, und der zweite nach Induktionsvoraussetzung an $m$ gleich $[\prod_{i=1}^{n} \mathbb{P}(A_i^{k_i})]\mathbb{P}(A_{n+1})$. Nun setzt man diese zwei Gleichungen ein, klammert geeignet aus und erhält die beabsichtigte Gleichung:

$$\mathbb{P}\Big(\bigcap_{i=1}^{n+1} A_i^{k_i}\Big) = \prod_{i=1}^{n+1} \mathbb{P}(A_i^{k_i}).$$

'$\Longleftarrow$': Wir setzen also die Gültigkeit von (2.2.2) voraus und zeigen nun die Unabhängigkeit von $A_1,\dots,A_n$. Sei $\{i_1,\dots,i_k\} \subset \{1,\dots,n\}$, und sei $\{j_1,\dots,j_m\}$ das Komplement von $\{i_1,\dots,i_k\}$ in $\{1,\dots,n\}$. Dann lässt sich $\bigcap_{l=1}^{k} A_{i_l}$ wie folgt als disjunkte Vereinigung schreiben:

$$\bigcap_{l=1}^{k} A_{i_l} = \bigcup_{k_1,\dots,k_m \in \{1,\mathrm{c}\}} \bigcap_{l=1}^{k} A_{i_l} \cap \bigcap_{s=1}^{m} A_{j_s}^{k_s}.$$

Die Wahrscheinlichkeit von der Menge auf der rechten Seite ist nach unserer Voraussetzung gleich

$$\sum_{k_1,\dots,k_m \in \{1,\mathrm{c}\}} \prod_{l=1}^{k} \mathbb{P}(A_{i_l}) \times \prod_{s=1}^{m} \mathbb{P}(A_{j_s}^{k_s}) = \prod_{l=1}^{k} \mathbb{P}(A_{i_l}),$$

wobei wir auch die Additivität bei paarweiser Disjunktheit benutzten. □

Da Lemma 2.2.6 symmetrisch ist in den Ereignissen und ihren Komplementen, ist die folgende Folgerung klar.

**Korollar 2.2.7.** *Ereignisse $A_1, \dots, A_n$ sind unabhängig genau dann, wenn ihre Komplemente $A_1^{\mathrm{c}}, \dots, A_n^{\mathrm{c}}$ unabhängig sind.*

Eine sehr hübsche Anwendung betrifft die Riemannsche Zetafunktion und die Eulersche Produktformel:

**Beispiel 2.2.8 (Eulersche Primzahlformel).** Die *Riemannsche Zetafunktion* ist definiert durch

$$\zeta(s) = \sum_{k=1}^{\infty} k^{-s}, \qquad s > 1.$$

Die *Eulersche Primzahlformel* oder *Produktdarstellung* dieser Funktion lautet

$$\zeta(s) = \prod_{p \text{ prim}} \left(1 - p^{-s}\right)^{-1}.$$

Diese Formel wollen wir nun mit probabilistischen Mitteln beweisen. Wir skizzieren hier nur den Weg, die Ausformulierung der Details ist eine Übungsaufgabe.

Wir erhalten einen Wahrscheinlichkeitsraum $(\mathbb{N}, q)$, indem wir $q(k) = k^{-s}\zeta(s)^{-1}$ für $k \in \mathbb{N}$ setzen. Man zeigt, dass die Mengen $p\mathbb{N} = \{pk \colon k \in \mathbb{N}\}$ (die Mengen der durch $p$ teilbaren Zahlen) unabhängig sind, wenn $p$ über alle Primzahlen genommen wird. Aufgrund von Folgerung 2.2.7 sind also auch die Komplemente $(p\mathbb{N})^{\mathrm{c}}$ unabhängig. Deren Schnitt ist gleich der Menge $\{1\}$, deren Wahrscheinlichkeit gleich $\zeta(s)^{-1}$ ist. Die Unabhängigkeit liefert, dass der (unendliche) Schnitt der Mengen $(p\mathbb{N})^{\mathrm{c}}$ mit $p$ eine Primzahl die Wahrscheinlichkeit hat, die durch das Produkt der Wahrscheinlichkeiten dieser Mengen gegeben ist. Diese identifiziert man mit $(1 - p^{-s})$, und daraus folgt die Eulersche Produktdarstellung der Riemannschen Zetafunktion. ◇

## 2.3 Produkträume

Der Begriff der Unabhängigkeit ist eng verknüpft mit Produkträumen von mehreren Wahrscheinlichkeitsräumen. Man denke an $n$ nacheinander und unabhängig von einander ausgeführten Zufallsexperimenten, die jeweils durch einen Wahrscheinlichkeitsraum beschrieben werden. Die gesamte Versuchsreihe dieser $n$ Experimente wird in natürlicher Weise durch den Produktraum beschrieben.

**Definition 2.3.1.** *Es seien $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$ diskrete Wahrscheinlichkeitsräume. Auf der Produktmenge*

$$\Omega = \Omega_1 \times \cdots \times \Omega_n = \big\{(\omega_1, \dots, \omega_n) \colon \omega_1 \in \Omega_1, \dots, \omega_n \in \Omega_n\big\}$$

*definieren wir $p \colon \Omega \to [0,1]$ durch*

$$p\big((\omega_1, \dots, \omega_n)\big) = \prod_{i=1}^{n} p_i(\omega_i).$$

*Dann ist $(\Omega, p)$ ein diskreter Wahrscheinlichkeitsraum, der der* Produktwahrscheinlichkeitsraum *der Räume $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$ genannt wird. Er wird auch mit $\otimes_{i=1}^{n}(\Omega_i, p_i)$ bezeichnet. Falls die Räume $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$ identisch sind, schreiben wir auch $(\Omega, p) = (\Omega_1, p_1)^{\otimes n}$.*

Das Beispiel 2.2.2 wird im folgenden Satz verallgemeinert.

**Satz 2.3.2.** *Seien $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$ diskrete Wahrscheinlichkeitsräume, und seien $A_1 \subset \Omega_1, \dots, A_n \subset \Omega_n$ Ereignisse in den jeweiligen Räumen. Dann sind die Ereignisse $A_1^{(1)}, \dots, A_n^{(n)}$ unabhängig im Produktraum $(\Omega, p)$ der Räume $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$, wobei*

$$A_i^{(i)} = \{(\omega_1, \dots, \omega_n) \in \Omega \colon \omega_i \in A_i\}.$$

(Die Ereignisse $A_i^{(i)}$ sind nichts weiter als eine Art Einbettung des Ereignisses '$A_i$ tritt ein' in den Produktraum an die $i$-te Stelle.)

**Beweis.** Mit $\mathbb{P}$ bezeichnen wir das von $p$ auf dem Produktraum $\Omega$ induzierte Wahrscheinlichkeitsmaß.

Sei $J \subset \{1, \dots, n\}$, dann ist die Produktformel in (2.2.1) für die Ereignisse $A_i^{(i)}$ mit $i \in J$ zu zeigen. Die Wahrscheinlichkeit des Schnittes dieser Mengen drücken wir durch Summation über alle Einzelereignisse aus:

$$\begin{aligned}\mathbb{P}\Big(\bigcap_{i \in J} A_i^{(i)}\Big) &= \mathbb{P}\big(\{\omega \colon \omega_i \in A_i \text{ für alle } i \in J\}\big) \\ &= \sum_{\omega \colon \omega_i \in A_i \forall i \in J} p(\omega) = \sum_{\omega_1, \dots, \omega_n \colon \omega_i \in A_i \forall i \in J} p_1(\omega_1) \dots p_n(\omega_n).\end{aligned}$$

Den letzten Ausdruck können wir mit der Notation $B_i = A_i$, falls $i \in J$ und $B_i = \Omega_i$, falls $i \notin J$, zusammenfassen als

$$\sum_{\omega_1 \in B_1} p_1(\omega_1) \cdots \sum_{\omega_n \in B_n} p_n(\omega_n) = \prod_{i=1}^{n} \sum_{\omega_i \in B_i} p_i(\omega_i) = \prod_{i=1}^{n} \mathbb{P}_i(B_i) = \prod_{i \in J} \mathbb{P}_i(A_i),$$

da ja $\mathbb{P}_i(B_i) = \mathbb{P}_i(\Omega_i) = 1$ für $i \notin J$. (Natürlich bezeichnet $\mathbb{P}_i$ das zu $(\Omega_i, p_i)$ gehörige Wahrscheinlichkeitsmaß.) Nun beachte man, dass gilt:

$$A_i^{(i)} = \Omega_1 \times \cdots \times \Omega_{i-1} \times A_i \times \Omega_{i+1} \times \cdots \times \Omega_n,$$

woraus folgt, dass $\mathbb{P}_i(A_i) = \mathbb{P}(A_i^{(i)})$. Dies beendet den Beweis. □

# Kapitel 3

# Zufallsgrößen, Erwartungswerte und Varianzen

Wir wollen nicht nur die Wahrscheinlichkeiten von Ereignissen bestimmen, sondern auch zufällige Größen behandeln. In diesem Abschnitt präzisieren wir, was eine Zufallsgröße ist, und was wir unter ihrem Erwartungswert und ihrer Varianz verstehen wollen. Außerdem erläutern wir, was Unabhängigkeit von Zufallsgrößen ist.

Wir legen wieder dem gesamten Kapitel einen diskreten Wahrscheinlichkeitsraum $(\Omega, p)$ zu Grunde.

## 3.1 Zufallsgrößen

**Definition 3.1.1 (Zufallsgröße).** *Jede Abbildung $X\colon \Omega \to \mathbb{R}$ heißt eine (reellwertige)* Zufallsgröße *oder* Zufallsvariable.

**Beispiel 3.1.2.** Die Augensumme bei zwei Würfen mit einem fairen Würfel ist die auf $\Omega = \{1,\dots,6\}^2$ mit der Gleichverteilung definierte Zufallsvariable $X\colon \Omega \to \mathbb{R}$ mit $X(i,j) = i + j$. $\diamond$

**Beispiel 3.1.3.** Die Anzahl der Erfolge im Bernoulli-Experiment (also $\Omega = \{0,1\}^n$ mit $q(\omega) = p^{\sum_{i=1}^n \omega_i}(1-p)^{n-\sum_{i=1}^n \omega_i}$, wobei $p \in [0,1]$ ein Parameter ist; siehe Beispiel 1.3.3) ist die Zufallsvariable $X\colon \Omega \to \mathbb{R}$, gegeben durch $X(\omega) = \sum_{i=1}^n \omega_i$. $\diamond$

Im Umgang mit Zufallsgrößen haben sich einige sehr handliche Konventionen eingebürgert. Wir schreiben $X(\Omega)$ für die (höchstens abzählbare) Menge $\{X(\omega) \in \mathbb{R}\colon \omega \in \Omega\}$, das Bild von $\Omega$ unter $X$. Für eine Menge $A \subset \mathbb{R}$ ist die Menge $X^{-1}(A) = \{\omega \in \Omega\colon X(\omega) \in A\}$ (das Urbild von $A$ unter $X$) das Ereignis '$X$ nimmt einen Wert in $A$ an' oder '$X$ liegt in $A$'. Wir benutzen die Kurzschreibweisen

$$\{X \in A\} = X^{-1}(A) = \{\omega \in \Omega\colon X(\omega) \in A\},$$
$$\{X = z\} = X^{-1}(\{z\}) = \{\omega \in \Omega\colon X(\omega) = z\},$$
$$\{X \le z\} = X^{-1}((-\infty, z]) = \{\omega \in \Omega\colon X(\omega) \le z\}$$

und so weiter. Statt $\mathbb{P}(\{X \in A\})$ schreiben wir $\mathbb{P}(X \in A)$ etc. Mit $\mathbb{P}(X \in A, Y \in B)$ meinen

wir immer $\mathbb{P}(\{X \in A\} \cap \{Y \in B\})$, d. h., das Komma steht immer für 'und' bzw. für den mengentheoretischen Schnitt.

**Definition 3.1.4 (Verteilung einer Zufallsgröße).** *Sei $X$ eine Zufallsgröße, dann ist das Paar $(X(\Omega), p_X)$ definiert durch $p_X(x) = \mathbb{P}(X = x)$, ein diskreter Wahrscheinlichkeitsraum. Das induzierte Wahrscheinlichkeitsmaß $\mathbb{P} \circ X^{-1}$, definiert durch $\mathbb{P} \circ X^{-1}(A) = \sum_{x \in A} p_X(x)$, erfüllt $\mathbb{P} \circ X^{-1}(A) = \mathbb{P}(X \in A)$ für alle $A \subset X(\Omega)$ und wird die* Verteilung von $X$ *genannt.*

**Bemerkung 3.1.5.** Wenn man $X^{-1}$ als Urbildoperator von der Potenzmenge von $X(\Omega)$ in die Potenzmenge von $\Omega$ auffasst, dann ist $\mathbb{P} \circ X^{-1}$ die Hintereinanderausführung der beiden Abbildungen $X^{-1}\colon \mathcal{P}(X(\Omega)) \to \mathcal{P}(\Omega)$ und $\mathbb{P}\colon \mathcal{P}(\Omega) \to [0,1]$.

Wir können $\mathbb{P} \circ X^{-1}(A) = \mathbb{P}(X \in A)$ für *jede* Teilmenge von $\mathbb{R}$ betrachten und meinen damit $\mathbb{P} \circ X^{-1}(A \cap X(\Omega))$. ◇

Je nachdem, ob das Wahrscheinlichkeitsmaß $\mathbb{P} \circ X^{-1}$ die Binomial-, geometrische oder hypergeometrische Verteilung ist, nennen wir die Zufallsgröße $X$ binomialverteilt, geometrisch verteilt oder hypergeometrisch verteilt, analog für jede andere Verteilung. Insbesondere ist also die Anzahl der Erfolge in einem Bernoulli-Experiment binomialverteilt, und die Wartezeit auf den ersten Erfolg ist geometrisch verteilt. Letztere Verteilung besitzt eine interessante charakteristische Eigenschaft, die (bei oberflächlicher Betrachtung) der Intuition widerspricht: Wenn man eine große Anzahl von erfolglosen Spielen beobachtet hat, erwartet man vielleicht, dass die Erfolgswahrscheinlichkeit im nächsten Spiel größer sein sollte, denn im Durchschnitt sollten sich ja über lange Spielserien die Erfolge und Misserfolge im Verhältnis ihrer Wahrscheinlichkeiten ausgleichen.[1] Der nächste Satz zeigt, dass dies ein Trugschluss ist.

**Lemma 3.1.6 (Gedächtnislosigkeit der geometrischen Verteilung).** *Sei $X$ eine geometrisch auf $\mathbb{N}$ verteilte Zufallsgröße, und sei $n \in \mathbb{N}$. Dann ist für jedes $k \in \mathbb{N}$*

$$\mathbb{P}(X = k) = \mathbb{P}(X = n - 1 + k \mid X \geq n).$$

Insbesondere ist also, wenn man $n-1$ erfolglosen Spielen beigewohnt hat, die Wahrscheinlichkeit für den nächsten Erfolg nach genau $k$ weiteren Spielen unabhängig von $n$, also unabhängig vom bisherigen Spielverlauf.

**Beweis.** Wir rechnen die rechte Seite explizit aus:

$$\begin{aligned}\mathbb{P}(X = n - 1 + k \mid X \geq n) &= \frac{\mathbb{P}(X = n - 1 + k, X \geq n)}{\mathbb{P}(X \geq n)} = \frac{\mathbb{P}(X = n - 1 + k)}{\sum_{l=n}^{\infty} \mathbb{P}(X = l)} \\ &= \frac{p(1-p)^{n+k-2}}{\sum_{l=n}^{\infty} p(1-p)^{l-1}} = \frac{(1-p)^{k-1}}{\sum_{l=1}^{\infty} (1-p)^{l-1}} = p(1-p)^{k-1} \\ &= \mathbb{P}(X = k).\end{aligned}$$

□

[1] Dieser Gedanke wird in Satz 6.1.4 unter dem Namen 'Gesetz der Großen Zahlen' präzisiert werden.

## 3.2 Unabhängigkeit von Zufallsgrößen

Wir beginnen mit einer recht allgemeinen Definition der Unabhängigkeit von diskreten Zufallsgrößen.

**Definition 3.2.1.** *Sei $(X_i)_{i\in I}$ eine Familie von Zufallsgrößen $X_i\colon \Omega \to \mathbb{R}$, wobei $I$ eine beliebige Indexmenge ist. Wir sagen, die Familie $(X_i)_{i\in I}$ ist* unabhängig *(oder auch, die $X_i$ mit $i \in I$ seien* unabhängig*), wenn für jede Familie $(A_i)_{i\in I}$ von reellen Mengen $A_i \subset \mathbb{R}$ die Familie der Ereignisse $(\{X_i \in A_i\})_{i\in I}$ unabhängig ist.*

Tatsächlich lässt das Konzept von diskreten Wahrscheinlichkeitsräumen, das wir in diesem Skript behandeln, es nicht zu, mehr als endlich viele unabhängige Zufallsgrößen zu konstruieren. Die Existenz einer unendlich großen Familie von unabhängigen Zufallsgrößen erfordert die Theorie von Wahrscheinlichkeiten auf überabzählbar großen Mengen, denn alleine wenn man schon abzählbar unendlich viele unabhängige Zufallsgrößen definieren möchte, die jeweils nur zwei verschiedene Werte annnehmen, sagen wir, 0 und 1, muss man das auf einer Menge tun, die die Menge $\{0,1\}^{\mathbb{N}}$ (die Menge aller unendlichen Folgen mit Koeffizienten in $\{0,1\}$) enthält, aber diese Menge ist überabzählbar. Die Wahrscheinlichkeitstheorie auf überabzählbaren Mengen ist Gegenstand der Vorlesung *Stochastik I*. Im Folgenden werden wir also ausschließlich *endlich viele* unabhängige Zufallsgrößen betrachten.

Wir bringen zunächst eine Charakterisierung der Unabhängigkeit von endlich vielen Zufallsgrößen. Ausführlicher könnte man die Unabhängigkeit von Zufallsgrößen $X_1, \dots, X_n$ formulieren, indem man die Gültigkeit die Produktformel (2.2.1) für alle endlichen Teilmengen $J$ von $\{1, \dots, n\}$ und alle Ereignisse der Form $\{X_i \in A_i\}$ fordert. Das folgende Lemma sagt, dass man sich zurück ziehen kann auf einelementige Mengen $A_i$ und auf die Wahl $J = \{1, \dots, n\}$. Der Beweis zeigt, dass man nämlich alle anderen Ereignisse aus dieser speziellen Wahl kombinieren kann durch Vereinigungsbildung.

**Lemma 3.2.2.** *Zufallsgrößen $X_1, \dots, X_n$ sind genau dann unabhängig, wenn für alle $x_1 \in X_1(\Omega), \dots, x_n \in X_n(\Omega)$ gilt:*

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i).$$

**Beweis.**

'$\Longrightarrow$': Dies folgt aus einer Anwendung der Definition 2.2.1 auf die Teilmenge $J = \{1, \dots, n\}$ und die Ereignisse $\{X_i = x_i\}$.

'$\Longleftarrow$' Seien $A_1, \dots, A_n \subset \mathbb{R}$, und sei $J$ eine beliebige nichtleere Teilmenge der Indexmenge $\{1, \dots, n\}$. Wir zeigen die Gültigkeit der Produktformel (2.2.1) für die Ereignisse $\{X_i \in A_i\}$ mit $i \in J$, wobei wir voraus setzen dürfen, dass $A_i \subset X_i(\Omega)$. Hierzu schreiben wir den Schnitt dieser Ereignisse als eine disjunkte Vereinigung von Schnitten von Ereignissen der Form $\{X_i = x_i\}$ mit $i \in \{1, \dots, n\}$. Auf diese Ereignisse können wir die Produktformel laut Voraussetzung anwenden.

Die Einführung der Notation $B_i = A_i$, falls $i \in J$, und $B_i = X_i(\Omega)$ sonst, vereinfacht die

Notation. Also gilt

$$\begin{aligned}\mathbb{P}\Big(\bigcap_{i\in J}\{X_i\in A_i\}\Big) &= \mathbb{P}\Big(\bigcap_{i=1}^{n}\{X_i\in B_i\}\Big) = \mathbb{P}\Big(\bigcup_{x_1\in B_1,\dots,x_n\in B_n}\bigcap_{i=1}^{n}\{X_i=x_i\}\Big)\\ &= \sum_{x_1\in B_1,\dots,x_n\in B_n}\mathbb{P}\Big(\bigcap_{i=1}^{n}\{X_i=x_i\}\Big) = \sum_{x_1\in B_1,\dots,x_n\in B_n}\prod_{i=1}^{n}\mathbb{P}(X_i=x_i)\\ &= \prod_{i=1}^{n}\sum_{x_i\in B_i}\mathbb{P}(X_i=x_i) = \prod_{i=1}^{n}\mathbb{P}(X_i\in B_i) = \prod_{i\in J}\mathbb{P}(X_i\in A_i).\end{aligned}$$

□

**Beispiel 3.2.3 (Indikatorvariable).** Für ein beliebiges Ereignis $A$ bezeichnen wir mit $\mathbb{1}_A$ die durch

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

definierte *Indikatorvariable* auf $A$. Es folgt leicht aus einer Kombination von Lemma 2.2.6 und Lemma 3.2.2, dass Ereignisse $A_1,\dots,A_n$ genau dann unabhängig sind, wenn die Indikatorvariablen $\mathbb{1}_{A_1},\dots,\mathbb{1}_{A_n}$ unabhängig sind. ◇

**Beispiel 3.2.4 (Bernoulli-Zufallsgrößen).** In einem oft auftretenden Spezialfall nehmen die unabhängigen Zufallsgrößen $X_1,\dots,X_n$ die Werte 1 und 0 jeweils mit Wahrscheinlichkeit $p$ bzw. $1-p$ an. Man sagt, $X_1,\dots,X_n$ sind *Bernoulli-Zufallsgrößen*. Man kann diese Größen auf dem in Beispiel 1.3.3 (siehe auch Beispiele 2.2.3 und 3.1.3) eingeführten Wahrscheinlichkeitsraum $(\Omega,q)$ formal definieren, indem man $X_i(\omega)=\omega_i$ setzt, d. h., $X_i$ ist 1, wenn das $i$-te Spiel einen Erfolg hatte. Also kann man Bernoulli-Zufallsgrößen auch auffassen als Indikatorvariable auf unabhängigen Ereignissen, die jeweils mit Wahrscheinlichkeit $p$ eintreten. ◇

Bernoulli-Zufallsgrößen benutzt man oft, um gewisse Modelle der statistischen Physik zu definieren:

**Beispiel 3.2.5 (Perkolation).** Ein Modell für die (zufällige) Durchlässigkeit eines porösen Materials erhält man auf folgende Weise. Wir sagen, zwei Punkte $i$ und $j$ im Gitter $\mathbb{Z}^d$ sind *benachbart*, und wir schreiben $i\sim j$, falls sich $i$ und $j$ nur in genau einer Komponenten unterscheiden, und zwar um 1 oder $-1$. Wir betrachten eine endliche Box $\Lambda\subset\mathbb{Z}^d$, die den Ursprung enthält. Jede Kante zwischen Nachbarn in $\Lambda$ habe den Wert 'offen' mit Wahrscheinlichkeit $p\in(0,1)$ und den Wert 'geschlossen' sonst. Die Offenheit der Kanten sei unabhängig. Also haben wir eine Kollektion von Bernoulli-Zufallsgrößen $(X_{\{i,j\}})_{i,j\in\Lambda,i\sim j}$ mit Werten in einer zweielementigen Menge.

Wir denken uns eine Wasserquelle im Ursprung, und das Wasser kann frei entlang offener Kanten laufen, aber nicht entlang der anderen Kanten. Eine typische Frage, die man sich nun stellt, ist: 'Mit welcher Wahrscheinlichkeit gibt es an der Oberfläche von $\Lambda$ feuchte Punkte?'. Besonders interessant ist diese Frage im Grenzübergang $\Lambda\uparrow\mathbb{Z}^d$, d. h., die Frage nach der Existenz eines offenen Weges nach Unendlich. (Man sagt in diesem Fall, dass die Flüssigkeit durchsickert, d. h. sie *perkoliert*.) Eine andere natürliche Frage ist die nach der erwarteten Anzahl (siehe den nächsten Abschnitt) von durchfeuchteten Zellen in dem Material, und im Fall $\Lambda\uparrow\mathbb{Z}^d$ die Frage, ob der feuchte Teil erwartungsgemäß endlich ist oder nicht.

Es ist klar, dass die Antworten auf diese Fragen nur von $p$ abhängen, und man erwartet einen sogenannten *Phasenübergang* zwischen kleinen und großen Werten von $p$, d. h. einen kritischen Wert von $p$, an dem die Antworten drastisch umschlagen. ◇

Ein weiteres Standardmodell der statistischen Physik wird mit Hilfe von abhängigen Zufallsgrößen konstruiert:

**Beispiel 3.2.6 (Ising-Modell).** Ein Modell für die Magnetisierung eines Stücks Eisen in einem Magnetfeld wird folgendermaßen definiert. Wieder sei $\Lambda \subset \mathbb{Z}^d$ eine endliche Box, und wir definieren die Verteilung von $\{-1,1\}$-wertigen Zufallsgrößen $X_i$ mit $i \in \Lambda$ wie folgt. (Wir interpretieren $X_i$ als die Ladung im Punkt $i$.) Für jede Konfiguration $(x_i)_{i\in\Lambda} \in \{-1,1\}^\Lambda$ sei

$$\mathbb{P}\big((X_i)_{i\in\Lambda} = (x_i)_{i\in\Lambda}\big) = \frac{1}{Z_{\beta,h,\Lambda}} \exp\Big\{\beta \sum_{i,j\in\Lambda, i\sim j} x_i x_j + h \sum_{i\in\Lambda} x_i\Big\},$$

wobei $\beta, h > 0$ Parameter seien und $Z_{\beta,h,\Lambda}$ eine geeignete Normierungskonstante. Das Maß $\mathbb{P}$ favorisiert Konfigurationen $(x_i)_{i\in\Lambda}$ mit vielen Übergängen zwischen gleichen Ladungen und mit vielen positiven Ladungen. Die Parameter $\beta$ bzw. $h$ regeln die Stärke dieser Effekte; insbesondere ist $h$ die Stärke des äußeren Magnetfeldes. ◇

Im folgenden Beispiel wird eine Ahnung davon gegeben, dass man bei geometrisch verteilten Zufallsgrößen sehr viele Wahrscheinlichkeiten explizit ausrechen kann.

**Beispiel 3.2.7 (unabhängige geometrische Zufallsgrößen).** Es seien $X$ und $Y$ zwei unabhängige, zum Parameter $p \in [0,1]$ geometrisch auf $\mathbb{N}_0$ verteilte Zufallsgrößen, d. h., wir haben $\mathbb{P}(X = k) = \mathbb{P}(Y = k) = (1-p)^k p$ für jedes $k \in \mathbb{N}_0$, und für alle $k, n \in \mathbb{N}_0$ gilt $\mathbb{P}(X = k, Y = n) = \mathbb{P}(X = k)\mathbb{P}(Y = n)$. Wir wollen die Wahrscheinlichkeiten gewisser mit $X$ und $Y$ beschriebener Ereignisse errechnen.

Als Beispiel berechnen wir $\mathbb{P}(X = Y)$, also die Wahrscheinlichkeit, dass $X$ und $Y$ den selben Wert annehmen. Da das Ereignis $\{X = Y\}$ die disjunkte Vereinigung der Ereignisse $\{X = k, Y = k\}$ mit $k \in \mathbb{N}_0$ ist, erhalten wir

$$\begin{aligned}\mathbb{P}(X = Y) &= \sum_{k\in\mathbb{N}_0} \mathbb{P}(X = k, Y = k) = \sum_{k\in\mathbb{N}_0} \mathbb{P}(X = k)\mathbb{P}(Y = k) \\ &= p^2 \sum_{k\in\mathbb{N}_0} (1-p)^k (1-p)^k = p^2 \sum_{k\in\mathbb{N}_0} [(1-p)^2]^k = p^2 \frac{1}{1-(1-p)^2} \\ &= \frac{p}{2-p}.\end{aligned}$$

Als Übungsaufgabe berechne man die Werte der Wahrscheinlichkeiten $\mathbb{P}(X \le Y)$, $\mathbb{P}(X < Y)$ und $\mathbb{P}(X = Y = Z)$, wobei $X$, $Y$ und $Z$ drei geometrisch verteilte Zufallsgrößen sind. ◇

Wie bei Unabhängigkeit von Ereignissen gibt es auch einen engen Zusammenhang zwischen Unabhängigkeit von Zufallsgrößen und Produkträumen. Im folgenden Lemma charakterisieren wir die Unabhängigkeit von Zufallsgrößen mit Hilfe der gemeinsamen Verteilung der Größen. Zunächst erklären wir, was man unter der gemeinsamen Verteilung von mehreren Zufallsgrößen versteht.

**Definition 3.2.8 (Gemeinsame Verteilung, Randverteilung).** *Es seien $X_1, \dots, X_n$ Zufallsgrößen. Dann verstehen wir unter der* gemeinsamen Verteilung *der $X_1, \dots, X_n$ die* Verteilung des Zufallsvektors $X = (X_1, \dots, X_n)$. *Dies ist (analog zu Definition 3.1.4) das Wahrscheinlichkeitsmaß $\mathbb{P} \circ X^{-1}$, das durch $p_X$ induziert wird, und $p_X$ wird auf der Bildmenge $X(\Omega) = \{(X_1(\omega), \dots, X_n(\omega)) \colon \omega \in \Omega\}$ definiert durch*

$$p_X(x_1, \dots, x_n) = \mathbb{P}\big(X = (x_1, \dots, x_n)\big) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n).$$

*Die Verteilung der einzelnen Zufallsgrößen $X_i$ erhält man, indem man die $i$-te* Randverteilung *oder* Marginalverteilung *von $p_X$ bildet, die gegeben ist durch*

$$\mathbb{P}(X_i = x_i) = p_{X_i}(x_i) = \sum_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n} p_X(x_1, \dots, x_n), \qquad x_i \in X_i(\Omega).$$

Beispiel 3.2.6 enthielt ein Beispiel für eine gemeinsame Verteilung von nicht unabhängigen Zufallsgrößen. Nun folgt der angekündigte Zusammenhang zwischen Unabhängigkeit und Produktverteilungen.

**Lemma 3.2.9 (Zufallsvektoren mit unabhängigen Komponenten).** *Es seien $X_1, \dots, X_n$ Zufallsgrößen. Dann sind $X_1, \dots, X_n$ genau dann unabhängig, wenn die Verteilung von $X$ gleich der Produktverteilung der Verteilungen der $X_1, \dots, X_n$ ist.*

**Beweis.** Diese Aussage ist nur eine Umformulierung von Lemma 3.2.2, wie man sich leicht klar macht. □

In der Situation von Lemma 3.2.9 sagt man, der Zufallsvektor $X$ habe unabhängige Komponenten $X_1, \dots, X_n$.

Die folgende Bemerkung ist analog zu Lemma 2.3.2.

**Bemerkung 3.2.10.** Falls die Zufallsgrößen $X_1, \dots, X_n$ auf eventuell verschiedenen Wahrscheinlichkeitsräumen $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$ definiert sind, können wir sie auf dem Produktraum $(\Omega, p)$ der $(\Omega_1, p_1), \dots, (\Omega_n, p_n)$ (siehe Definition 2.3.1) in einer solchen Weise definieren, dass sie unabhängig sind. Dazu betrachten wir die Zufallsgrößen $X_1^{(1)}, \dots, X_n^{(n)}$, die durch $X_i^{(i)}((\omega_1, \dots, \omega_n)) = X_i(\omega_i)$ definiert sind. Dann ist die Verteilung von $X_i^{(i)}$ (auf $\Omega$) gleich der von $X_i$ (auf $\Omega_i$). Ferner sind die Variablen $X_1^{(1)}, \dots, X_n^{(n)}$ unabhängig, denn für alle $x_1, \dots, x_n$ gilt

$$\begin{aligned}
\mathbb{P}(X_1^{(1)} = x_1, \dots, X_n^{(n)} = x_n) &= \sum_{\omega \in \Omega \colon X_i^{(i)}(\omega) = x_i \forall i} p(\omega_1, \dots, \omega_n) \\
&= \sum_{\omega_1 \in \Omega_1 \colon X_1(\omega_1) = x_1} \cdots \sum_{\omega_n \in \Omega_n \colon X_n(\omega_n) = x_n} \prod_{i=1}^{n} p_i(\omega_i) \\
&= \prod_{i=1}^{n} \sum_{\omega_i \in \Omega_i \colon X_i(\omega_i) = x_1} p_i(\omega_i) \\
&= \prod_{i=1}^{n} \mathbb{P}_i(X_i = x_i) = \prod_{i=1}^{n} \mathbb{P}(X_i^{(i)} = x_i).
\end{aligned}$$

Hierbei ist $p$ in Definition 2.3.1 eingeführt worden, und $\mathbb{P}$ und $\mathbb{P}_i$ sind die zu $p$ bzw. zu $p_i$ gehörigen Wahrscheinlichkeitsmaße auf $\Omega$ bzw. auf $\Omega_i$. ◇

Aus Familien unabhängiger Zufallsgrößen kann man andere Familien unabhängiger Zufallsgrößen gewinnen, indem man sie auf disjunkte Weise irgendwie mit einander kombiniert:

**Lemma 3.2.11 (Kombinationen unabhängiger Zufallsgrößen).** *Es sei $(X_i)_{i\in I}$ eine Familie unabhängiger Zufallsgrößen, wobei $I$ eine beliebige Indexmenge sei. Ferner seien $I_1, I_2, \ldots$ endliche, paarweise disjunkte Teilmengen von $I$, und für jedes $j \in \mathbb{N}$ sei $Y_j$ eine Zufallsgröße, die aus den Zufallsvariablen $X_i$ mit $i \in I_j$ gebildet sei, also $Y_j = f_j((X_i)_{i\in I_j})$ für eine geeignete Funktion $f_j$. Dann sind die Zufallsvariablen $Y_j$ mit $j \in \mathbb{N}$ unabhängig.*

**Beweis.** (Wegen der notationellen Unhandlichkeit formulieren wir diesen Beweis nicht voll aus.)

Es seien $k \in \mathbb{N}$ und $y_1 \in Y_1(\Omega), , \ldots, y_k \in Y_k(\Omega)$, dann müssen wir die Gültigkeit der Produktformel $\mathbb{P}(Y_1 = y_1, \ldots, Y_k = y_k) = \prod_{j=1}^{k} \mathbb{P}(Y_j = y_j)$ zeigen. Um dies zu tun, zerlegen wir das Ereignis $\{Y_j = y_j\}$ zunächst in eine disjunkte Vereinigung von Ereignissen, die mit Hilfe der $X_i$ mit $i \in I_j$ gebildet werden:

$$\{Y_j = y_j\} = \bigcup_{(x_i)_{i\in I_j} \colon f_j\left((x_i)_{i\in I_j}\right)=y_j} \left\{(X_i)_{i\in I_j} = (x_i)_{i\in I_j}\right\}. \tag{3.2.1}$$

Nun erhält man einen Ausdruck für das Produkt $\prod_{j=1}^{k} \mathbb{P}(Y_j = y_j)$, indem man in (3.2.1) zu den Wahrscheinlichkeiten übergeht, die Summenformel für disjunkte Vereinigungen benutzt, über $j = 1, \ldots, k$ multipliziert und die Wahrscheinlichkeiten der Ereignisse auf der rechten Seite von (3.2.1) mit Hilfe der Unabhängigkeit der $X_i$ in Produkte von einzelnen Wahrscheinlichkeiten zerlegt.

Auf der anderen Seite erhält man analog zu (3.2.1) eine Darstellung des Ereignisses $\{Y_1 = y_1, \ldots, Y_k = y_k\}$ als disjunkte Vereinigung von Ereignissen, die mit Hilfe von $(X_i)_{i\in I_j}$ für $j \in \{1, \ldots, k\}$ gebildet werden. Die Wahrscheinlichkeit dieser Vereinigung errechnet man analog und stellt durch Vergleich mit dem Resultat der obigen Rechnung fest, dass sie identisch ist mit dem Produkt der Wahrscheinlichkeiten der Ereignisse $\{Y_j = y_j\}$. □

Das folgende wichtige Korollar zu Lemma 3.2.11 erhält man durch die Wahl von einelementigen Teilindexmengen:

**Korollar 3.2.12.** *Falls die Zufallsvariablen $X_i$ mit $i \in \mathbb{N}$ unabhängig sind, so auch die Zufallsvariablen $f_i(X_i)$ mit $i \in \mathbb{N}$, wobei $f_i$ beliebige Funktionen sind.*

## 3.3 Erwartungswerte

Wir führen einen zentralen Begriff der Wahrscheinlichkeitstheorie ein: den Begriff des 'erwarteten Wertes' einer Zufallsgröße. Man erhält diesen Wert, indem man über alle Werte, die die Zufallsgröße annehmen kann, mittelt mit den Gewichten, die durch die Einzelwahrscheinlichkeiten gegeben sind.

**Definition 3.3.1 (Erwartungswert).** *Wir sagen, eine Zufallsgröße $X\colon \Omega \to \mathbb{R}$ besitzt einen* Erwartungswert, *falls die Reihe $\sum_{\omega\in\Omega} p(\omega)|X(\omega)|$ konvergiert. In diesem Fall schreiben wir auch $X \in \mathcal{L}^1(\mathbb{P})$ (oder auch einfach $X \in \mathcal{L}^1$) und definieren wir den* Erwartungswert von $X$ *als die Zahl*

$$\mathbb{E}(X) = \sum_{\omega\in\Omega} p(\omega)X(\omega).$$

In Erweiterung von Definition 3.3.1 können wir auch für jede *nicht negative* Zufallsgröße $X$ den Erwartungswert von $X$ als $\mathbb{E}(X) = \sum_{\omega\in\Omega} p(\omega)X(\omega)$ definieren, auch wenn die Reihe divergiert. In letzterem Fall setzen wir $\mathbb{E}(X) = \infty$. In der selben Weise können wir auch Zufallsgrößen behandeln, die nach unten beschränkt sind oder nach oben beschränkt.

Die Voraussetzung der *absoluten* Konvergenz der Reihe $\sum_{\omega\in\Omega} p(\omega)X(\omega)$ sichert, dass der Wert dieser Reihe nicht von der gewählten Aufzählung von $\Omega$ abhängt. Außerdem impliziert sie etliche Rechenregeln, die man vom Begriff des Erwartungswertes erwartet:

**Lemma 3.3.2 (Eigenschaften des Erwartungswertes).** *(a) Eine Zufallsgröße $X$ liegt genau dann in $\mathcal{L}^1$, wenn die Reihe $\sum_{x\in X(\Omega)} |x|\mathbb{P}(X = x)$ konvergiert. In diesem Fall gilt $\mathbb{E}(X) = \sum_{x\in X(\Omega)} x\mathbb{P}(X = x)$.*

*(b) Falls $X, Y \in \mathcal{L}^1$ mit $X \le Y$, so gilt $\mathbb{E}(X) \le \mathbb{E}(Y)$.* (Monotonie des Erwartungswertes)

*(c) Falls $X, Y \in \mathcal{L}^1$ und $c \in \mathbb{R}$, so ist $X + cY \in \mathcal{L}^1$, und es gilt $\mathbb{E}(X + cY) = \mathbb{E}(X) + c\mathbb{E}(Y)$.* (Linearität des Erwartungswertes)

*(d) Falls $X, Y \in \mathcal{L}^1$ unabhängig sind, so ist auch $XY \in \mathcal{L}^1$, und es gilt $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.* (Produktregel bei Unabhängigkeit)

**Beweis.**

($a$) Die erste Aussage wird gezeigt durch die Rechnung

$$\begin{aligned}\sum_{x\in X(\Omega)} |x|\mathbb{P}(X = x) &= \sum_{x\in X(\Omega)} |x| \sum_{\omega\colon X(\omega)=x} p(\omega) = \sum_{x\in X(\Omega)} \sum_{\omega\colon X(\omega)=x} |X(\omega)|p(\omega)\\ &= \sum_{\omega\in\Omega} |X(\omega)|p(\omega),\end{aligned}$$

und die zweite Aussage folgt aus einer Wiederholung dieser Rechnung ohne Betragstriche.

($b$) folgt mit Hilfe von (a) aus den Regeln für absolut konvergente Reihen.

($c$) folgt mit Hilfe von Definition 3.3.1 aus den Rechenregeln für absolut konvergente Reihen.

($d$) Wir spalten auf nach allen Werten von $X$:

$$\begin{aligned}\sum_z |z|\mathbb{P}(XY = z) &= \sum_{z\neq 0}\sum_x |z|\mathbb{P}(XY = z, X = x) = \sum_{x,z\neq 0} |z|\mathbb{P}(X = x, Y = z/x)\\ &= \sum_{x,y\neq 0} |x|\,|y|\mathbb{P}(X = x, Y = y) = \sum_{x,y\neq 0} |x|\,|y|\mathbb{P}(X = x)\mathbb{P}(Y = y)\\ &= \sum_x |x|\mathbb{P}(X = x) \sum_y |y|\mathbb{P}(Y = y),\end{aligned}$$

wobei wir im vorletzten Schritt die Unabhängigkeit von $X$ und $Y$ benutzten. Dies zeigt, dass $XY \in \mathcal{L}^1$. Die behauptete Gleichung folgt durch eine Wiederholung der obigen Rechnung ohne Betragstriche. □

Für die Behandlung des Erwartungswertes von zusammengesetzten Zufallsgrößen ist das folgende Lemma nützlich. Wir erinnern an die Definition 3.2.8.

**Lemma 3.3.3.** *Es seien $X_1, \dots, X_n$ Zufallsgrößen, und es sei $g\colon X_1(\Omega) \times \cdots \times X_n(\Omega) \to \mathbb{R}$ eine Abbildung. Dann existiert der Erwartungswert der Zufallsgröße $Y = g(X_1, \dots, X_n) = g \circ (X_1, \dots, X_n)$ genau dann, wenn die Reihe*

$$\sum_{x_1 \in X_1(\Omega)} \cdots \sum_{x_n \in X_n(\Omega)} g(x_1, \dots, x_k)\mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$$

*absolut konvergiert, und der Wert der Reihe ist dann gleich $\mathbb{E}(Y)$.*

**Beweis.** Wir machen $\Omega' = X_1(\Omega) \times \cdots \times X_n(\Omega)$ zu einem Wahrscheinlichkeitsraum $(\Omega', p')$, indem wir $p'(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$ setzen. Dann ist die Verteilung von $g$ auf $\Omega'$ identisch mit der von $Y$ auf $\Omega$. Also folgt die Aussage aus Lemma 3.3.2(a). □

Nun bestimmen wir die Erwartungswerte einiger wichtiger Verteilungen.

**Beispiel 3.3.4 (Erwartungswert der Binomialverteilung).** Sei $X$ eine binomialverteilte Zufallsgröße, also $\mathbb{P}(X = k) = \mathrm{Bi}_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}$ für $k \in \{0, \dots, n\}$, wobei $n \in \mathbb{N}_0$ und $p \in [0, 1]$. Es ist durchaus möglich, den Erwartungswert von $X$ mit Hilfe von Lemma 3.3.2(a) zu berechnen (Übungsaufgabe: Man führe dies durch), doch wir schlagen einen weniger rechnenaufwändigen Weg ein, indem wir Lemma 3.3.2(c) benutzen. Wir erinnern uns (siehe Beispiele 3.1.3 und 3.2.4), dass wir $X$ auf dem Raum $\Omega = \{0, 1\}^n$ mit $q(\omega) = p^{\sum_{i=1}^n \omega_i}(1-p)^{n-\sum_{i=1}^n \omega_i}$ definieren können, indem wir $X(\omega) = \sum_{i=1}^n \omega_i$ setzen. Mit anderen Worten, wir haben $X = \sum_{i=1}^n X_i$ mit den in Beispiel 3.2.4 definieren Bernoulli-Größen. Jedes $X_i$ nimmt die Werte 1 und 0 mit Wahrscheinlichkeit $p$ bzw. $1-p$ an, also sieht man leicht, dass $\mathbb{E}(X_i) = 1 \cdot p + 0 \cdot (1-p) = p$ ist. Mit Hilfe der Linearität des Erwartungwertes erhält man also leicht, dass $\mathbb{E}(X) = \sum_{i=1}^n \mathbb{E}(X_i) = np$. ◇

**Beispiel 3.3.5 (Erwartungswert der geometrischen Verteilung).** Es sei $X$ eine zum Parameter $p \in (0, 1)$ auf $\mathbb{N}$ geometrisch verteilte Zufallsgröße, also $\mathbb{P}(X = k) = \mathrm{Geo}_p(k) = p(1-p)^{k-1}$ für $k \in \mathbb{N}$. Zur Berechnung des Erwartungswertes von $X$ (und für viele andere Berechnungen für geometrisch verteilte Zufallsgrößen) ist der folgende Trick sehr hilfreich. Eine gliedweise Differenziation der Identität $\sum_{k=1}^\infty x^k = \frac{x}{1-x}$ nach $x$ für $|x| < 1$ liefert die Gleichung $\sum_{k=1}^\infty k x^{k-1} = \frac{1}{(1-x)^2}$. Eine Anwendung auf $x = 1-p$ liefert dann mit Hilfe von Lemma 3.3.2(a):

$$\mathbb{E}(X) = \sum_{k=1}^\infty k\mathbb{P}(X = k) = p\sum_{k=1}^\infty k(1-p)^{k-1} = p\frac{1}{p^2} = \frac{1}{p}.$$

Eine auf $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ geometrisch verteilte Zufallsgröße hat die Verteilung von $X - 1$, also den Erwartungswert $\frac{1}{p} - 1$. ◇

**Beispiel 3.3.6 (Erwartungswert der Poisson-Verteilung).** Sei $X$ eine zum Parameter $\alpha > 0$ Poisson-verteilte Zufallsgröße, also $\mathbb{P}(X = k) = \frac{\alpha^k}{k!} e^{-\alpha}$ für $k \in \mathbb{N}_0$. Der Erwartungswert

von $X$ kann leicht mit Lemma 3.3.2(a) berechnet werden:

$$\mathbb{E}(X) = \sum_{k\in\mathbb{N}_0} k\mathbb{P}(X=k) = e^{-\alpha}\sum_{k=1}^{\infty}\frac{\alpha^k}{(k-1)!} = \alpha e^{-\alpha}\sum_{k=0}^{\infty}\frac{\alpha^k}{k!} = \alpha,$$

wobei wir im vorletzten Schritt eine Laufindexverschiebung durchführten. ◇

**Beispiel 3.3.7.** Wir erinnern an die in Beispiel 2.2.8 eingeführte Verteilung auf $\mathbb{N}$, die durch $q(k) = k^{-s}\zeta(s)^{-1}$ gegeben ist, wobei $s > 1$. Da die Reihe $\sum_{k\in\mathbb{N}} kq(k)$ genau dann konvergiert, wenn $s > 2$, existiert der Erwartungswert dieser Verteilung also nur im Fall $s > 2$. Ihr Wert ist dann $\zeta(s-1)/\zeta(s)$. ◇

Die folgende Formel ist manchmal hilfreich, wenn eine $\mathbb{N}_0$-wertige Zufallsgröße $X$ nur durch Angabe ihrer 'Schwänze' $\mathbb{P}(X > k)$ gegeben ist. (Letzteres ist manchmal einfacher, weil man bei dieser Festlegung der Verteilung von $X$ nicht auf Normierung achten muss, sondern nur auf Monotonie.)

**Lemma 3.3.8.** *Der Erwartungswert einer $\mathbb{N}_0$-wertigen Zufallsgröße $X$ (egal, ob er endlich ist oder nicht) ist gegeben durch*

$$\mathbb{E}(X) = \sum_{k=0}^{\infty}\mathbb{P}(X > k).$$

**Beweis.** Übungsaufgabe. □

## 3.4 Varianzen

Die Kenngröße 'Erwartungswert' gibt natürlich bei weitem keine erschöpfende Information über die Zufallsgröße. Eine nützliche zusätzliche Information ist zum Beispiel die Antwort auf die Frage, wie stark die Zufallsgröße im Durchschnitt von ihrem Erwartungswert abweicht. Eine Kenngröße, die dies angibt, ist die Varianz.

**Definition 3.4.1.** *Es sei $X$ eine Zufallsgröße mit existierendem Erwartungswert $\mathbb{E}(X)$. Die* Varianz *von $X$ ist der Ausdruck*

$$\mathbb{V}(X) = \sum_x (x - \mathbb{E}(X))^2\mathbb{P}(X=x) \in [0,\infty].$$

*Wir sagen, die* Varianz existiert, *falls $\mathbb{V}(X) < \infty$. In diesem Fall definieren wir die* Standardabweichung *von $X$ als $S(X) = \sqrt{\mathbb{V}(X)}$.*

Man sieht leicht mit Hilfe von Lemma 3.3.2(a) ein, dass $\mathbb{V}(X)$ im Falle der Existenz der Erwartungswert der Zufallsgröße $\omega \mapsto (X(\omega) - \mathbb{E}(X))^2$ ist, also

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

**Beispiel 3.4.2 (Varianz einer Gleichverteilung).** Wenn eine Zufallsgröße $X$ auf einer $n$-elementigen Menge $\{x_1, \ldots, x_n\}$ gleichverteilt ist, so sind

$$\mathbb{E}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{und} \qquad \mathbb{V}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbb{E}(X))^2.$$

Also ist $\mathbb{E}(X)$ der Mittelwert der $x_1, \ldots, x_n$, und $\mathbb{V}(X)$ ist die mittlere quadratische Abweichung davon. ◇

**Beispiel 3.4.3 (Varianz der Bernoulli-Verteilung).** Eine Bernoulli-verteilte Zufallsgröße $X$ nimmt die Werte 1 und 0 mit Wahrscheinlichkeit $p \in [0,1]$ bzw. $(1-p)$ an. Also ist $\mathbb{E}(X) = p$, und die Varianz berechnet sich nach Definition 3.4.1 zu $\mathbb{V}(X) = (0 - \mathbb{E}(X))^2\mathbb{P}(X = 0) + (1 - \mathbb{E}(X))^2\mathbb{P}(X = 1) = p^2(1-p) + (1-p)^2 p = p(1-p)$. ◇

Die Varianz einer binomialverteilten Zufallsgröße kann man durchaus unter Verwendung der Definition 3.4.1 direkt berechnen, aber in Beispiel 3.5.4 werden wir einen eleganten Weg präsentieren.

Einige einfache Eigenschaften der Varianz sind im folgenden Lemma aufgelistet.

**Lemma 3.4.4.** *Es seien $X, Y \in \mathcal{L}^1$.*

*(a) Die Varianz von $X$ existiert genau dann, wenn $\mathbb{E}(X^2) < \infty$. In diesem Fall schreiben wir $X \in \mathcal{L}^2$, und es gilt die Formel*

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

*(b) Seien $a, b \in \mathbb{R}$. Falls die Varianz von $X$ existiert, dann auch die von $a + bX$, und es gilt $\mathbb{V}(a + bX) = b^2\mathbb{V}(X)$.*

*(c) Falls $X$ und $Y$ unabhängig sind mit existierenden Varianzen, dann existiert auch die Varianz von $X + Y$, und es gilt $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$.* (Satz von Bienaymé)

*(d) Falls $\mathbb{V}(X)$ existiert und $\mathbb{V}(X) = 0$ ist, so existiert ein $x \in \mathbb{R}$ mit $\mathbb{P}(X = x) = 1$.*

**Beweis.**

($a$): Wegen

$$(x - \mathbb{E}(X))^2\mathbb{P}(X = x) = x^2\mathbb{P}(X = x) - 2x\mathbb{E}(X)\mathbb{P}(X = x) + \mathbb{E}(X)^2\mathbb{P}(X = x)$$

ist die erste Behauptung klar, denn die absolute Konvergenz von $\sum_x x\mathbb{P}(X = x)$ ist vorausgesetzt und die von $\sum_x \mathbb{P}(X = x)$ ist klar. Die behauptete Gleichung erechnet man leicht, indem man die obige Beziehung über $x$ summiert und zusammenfasst.

($b$): Dies errechnet sich leicht mit Hilfe von (a) und der Linearität des Erwartungswertes.

($c$): Wir benutzen (a) für $X + Y$ und multiplizieren aus und erhalten

$$\mathbb{V}(X + Y) = \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) - \mathbb{E}(X)^2 - 2\mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(Y)^2.$$

Nun sehen wir mit Hilfe von Lemma 3.3.2(d), dass die rechte Seite gleich $\mathbb{E}(X^2) - \mathbb{E}(X)^2 + \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$ ist, nach (a) also gleich $\mathbb{V}(X) + \mathbb{V}(Y)$.

($d$): Dies sieht man leicht aus der Definition 3.4.1: Falls $\mathbb{V}(X) = 0$, so muss für jedes $x \in \mathbb{R}$ entweder $x = \mathbb{E}(X)$ oder $\mathbb{P}(X = x) = 0$ sein.

□

**Beispiel 3.4.5 (Varianz der Poisson-Verteilung).** Die Varianz einer zum Parameter $\alpha > 0$ Poisson-verteilten Zufallsgröße $X$ (siehe Beispiel 3.3.6) ist $\mathbb{V}(X) = \alpha$. (Übungsaufgabe) ◇

**Beispiel 3.4.6 (Varianz der geometrischen Verteilung).** Als Übungsaufgabe berechne man die Varianz einer geometrisch verteilten Zufallsgröße. (Man verwende den in Beispiel 3.3.5 erläuterten Trick ein zweites Mal.) ◇

Die Varianz bzw. der Erwartungswert besitzt die folgende Minimaleigenschaft:

**Lemma 3.4.7 (Minimale quadratische Abweichung).** *Für jede Zufallsgröße $X \in \mathcal{L}^2$ gilt die Abschätzung*

$$\mathbb{E}\big((X-a)^2\big) \geq \mathbb{V}(X), \qquad a \in \mathbb{R},$$

*mit Gleichheit genau dann, wenn $a = \mathbb{E}(X)$.*

**Beweis.** Mit Hilfe der Linearität des Erwartungswerts errechnet man leicht, dass $\mathbb{E}((X-a)^2) = \mathbb{V}(X) + (a - \mathbb{E}(X))^2$ für jedes $a \in \mathbb{R}$. Also ist die Aussage evident. □

## 3.5 Kovarianzen

In diesem Abschnitt stellen wir eine Kenngröße vor, die über die Abhängigkeiten zweier gegebener Zufallsgrößen eine gewisse Aussage macht.

**Definition 3.5.1.** *Es seien $X$ und $Y$ zwei Zufallsgrößen mit existierenden Varianzen. Die* Kovarianz *von $X$ und $Y$ ist die Zahl* $\operatorname{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$. *Wir nennen $X$ und $Y$* unkorreliert, *falls* $\operatorname{cov}(X,Y) = 0$.

Die Kovarianz ist wohldefiniert, denn der Erwartungswert von $XY$ existiert auf Grund der Abschätzung $2|XY| \leq X^2 + Y^2$ und Lemma 3.4.4(a).

Einige evidente Eigenschaften der Kovarianz werden nun gesammelt:

**Lemma 3.5.2.** *(a) Für je zwei Zufallsgrößen $X$ und $Y$ mit existierenden Varianzen gelten die Beziehungen*

$$\begin{aligned}
\operatorname{cov}(X,Y) &= \mathbb{E}\big((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big),\\
\operatorname{cov}(X,X) &= \mathbb{V}(X),\\
\operatorname{cov}(X,Y) &= \operatorname{cov}(Y,X),\\
\operatorname{cov}(aX,bY) &= ab\operatorname{cov}(X,Y), \qquad \text{für alle } a,b \in \mathbb{R}.
\end{aligned}$$

*(b) Für je $n$ Zufallsgrößen $X_1, \ldots, X_n$ gilt*

$$\mathbb{V}\Big(\sum_{i=1}^n X_i\Big) = \sum_{i=1}^n \mathbb{V}(X_i) + \sum_{\substack{i,j=1\\ i\neq j}}^n \operatorname{cov}(X_i, X_j).$$

*(c) Falls $X$ und $Y$ unabhängige Zufallsgrößen mit existierenden Varianzen sind, so sind $X$ und $Y$ unkorreliert.*

**Beweis.** Übungsaufgabe. □

Aus der ersten Beziehung in Lemma 3.5.2(a) folgt, dass sich die Kovarianz von $X$ und $Y$ nicht ändert, wenn zu $X$ oder $Y$ Konstanten addiert werden.

Die Aussage in Lemma 3.5.2(c) kann nicht ohne Weiteres umgekehrt werden, wie das folgende Beispiel zeigt. Wir betrachten auf $\Omega = \{-1, 0, 1\}$ mit der Gleichverteilung die Zufallsgröße $X$, gegeben durch $X(\omega) = \omega$. Als Übungsaufgabe vergewissere man sich, dass die beiden Zufallsgrößen $X$ und $|X|$ zwar unkorreliert, aber nicht unabhängig sind.

Eine direkte Folgerung aus Lemma 3.5.2(b) ist die folgende Aussage.

**Korollar 3.5.3 (Satz von Bienaymé).** *Für paarweise unkorrelierte Zufallsgrößen $X_1, \dots, X_n$ mit existierenden Varianzen gilt $\mathbb{V}(X_1 + \cdots + X_n) = \mathbb{V}(X_1) + \cdots + \mathbb{V}(X_n)$.*

**Beispiel 3.5.4 (Varianz der Binomialverteilung).** Im Kontext von Beispiel 3.3.4 wissen wir schon aus Beispiel 3.4.3, dass $\mathbb{V}(X_i) = p(1-p)$. Da die Zufallsgrößen $X_1, \dots, X_n$ unabhängig sind, erlaubt der Satz von Bienaymé, die Varianz von $X$ wie folgt zu berechen: $\mathbb{V}(X) = \mathbb{V}(X_1 + \cdots + X_n) = \mathbb{V}(X_1) + \cdots + \mathbb{V}(X_n) = np(1-p)$. ◇

Die folgende Minimaleigenschaft der Kovarianz ist manchmal hilfreich, wenn man eine (eventuell schwer zugängliche) Zufallsvariable $Y$ mit Hilfe einer linearen Funktion einer (leichter zu erhaltenden) Zufallsvariablen $X$ annähern möchte.

**Lemma 3.5.5 (Beste lineare Vorhersage).** *Es seien $X, Y \in \mathcal{L}^2$ mit $\mathbb{V}(X) = 1$. Dann wird die quadratische Abweichung*

$$\mathbb{E}\big((Y - a - bX)^2\big)$$

*zwischen $Y$ und der linearen Funktion $a + bX$ von $X$ minimiert genau für $b = \operatorname{cov}(X, Y)$ und $a = \mathbb{E}(Y - bX)$. Falls insbesondere $X$ und $Y$ unkorreliert sind, so hängt die Lösung $b = 0$ und $a = \mathbb{E}(Y)$ nicht von $X$ ab.*

**Beweis.** Eine Ausnutzung der Linearität des Erwartungswertes zeigt, dass der betrachtete Ausdruck ein Polynom zweiter Ordnung in $a$ und $b$ ist, das im Unendlichen gegen Unendlich geht und daher sein Minimum an der Nullstelle des Gradienten annimmt. Diese Nullstelle ist durch die behaupteten Gleichungen bestimmt. □

Eine der wichtigsten Ungleichungen ist die folgende.

**Satz 3.5.6 (Cauchy-Schwarz-Ungleichung).** *Für je zwei Zufallsgrößen $X$ und $Y$ gilt*

$$\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(X^2)}\,\sqrt{\mathbb{E}(Y^2)}.$$

*Gleichheit gilt genau dann, wenn es reelle Zahlen $a$, $b$ gibt, die nicht beide Null sind, sodass $\mathbb{P}(aX + bY = 0) = 1$, d. h. wenn $X$ und $Y$ konstante Vielfache von einander sind.*

**Beweis.** Wir setzen $\alpha = \mathbb{E}(Y^2)$ und $\beta = \mathbb{E}(XY)$. Wir dürfen annehmen, dass $\alpha > 0$, denn sonst wäre $\mathbb{P}(Y = 0) = 1$, also auch $\mathbb{E}(XY) = 0$, und die behauptete Ungleichung stimmt

trivialerweise. Nun errechnen wir durch Ausmultiplizieren und mit Hilfe der Linearität des Erwartungswertes:

$$\begin{aligned} 0 \leq \mathbb{E}((\alpha X - \beta Y)^2) &= \alpha^2 \mathbb{E}(X^2) - 2\alpha\beta\, \mathbb{E}(XY) + \beta^2 \mathbb{E}(Y^2) \\ &= \alpha\big(\mathbb{E}(X^2)\mathbb{E}(Y^2) - \mathbb{E}(XY)^2\big). \end{aligned}$$

Da $\alpha > 0$, folgt die behauptete Ungleichung.

Falls Gleichheit gilt, so ist der Erwartungswert von $(\alpha X - \beta Y)^2$ gleich Null, also folgt $\mathbb{P}(\alpha X - \beta Y = 0) = 1$. Falls $\alpha > 0$, so können wir $\alpha = a$ und $\beta = b$ wählen. Falls $\alpha = 0$, so können wir $a = 0$ und $b = 1$ wählen. □

Aus einer Anwendung der Cauchy-Schwarz-Ungleichung auf die Zufallsgrößen $X - \mathbb{E}(X)$ und $Y - \mathbb{E}(Y)$ folgt insbesondere die Ungleichung

$$-1 \leq \frac{\operatorname{cov}(X, Y)}{S(X)S(Y)} \leq 1$$

für alle Zufallsgrößen $X$ und $Y$, deren Varianzen existieren, wobei wir daran erinnern, dass $S(X) = \sqrt{\mathbb{V}(X)}$ die Standardabweichung von $X$ ist.

Die Bedingung $\operatorname{cov}(X, Y) > 0$ bedeutet, dass eine Tendenz vorliegt, nach der das Ereignis $\{X > 0\}$ öfter mit dem Ereignis $\{Y > 0\}$ zusammen auftritt als mit $\{Y \leq 0\}$. Man sagt, $X$ und $Y$ seien *positiv korreliert*. Das impliziert allerdings noch lange nicht, dass $X$ eine (Mit-)Ursache für $Y$ ist oder umgekehrt, auch wenn dieser Fehlschluss oft und gerne gemacht wird.

# Kapitel 4

# Summen unabhängiger Zufallsgrößen

In diesem Kapitel behandeln wir Summen unabhängiger Zufallsvariabler, ein Thema, das immer wieder wichtige Rollen spielte und spielen wird. Wir identifizieren die Verteilung dieser Summe auf zweifache Weise, indem wir zunächst den Zusammenhang mit der Faltung beleuchten und danach das kraftvolle Hilfsmittel der erzeugenden Funktionen einsetzen. Außerdem stellen wir eines der grundlegenden stochastischen Modelle vor, die sogenannte eindimensionale Irrfahrt.

## 4.1 Faltungen

Wenn $X$ und $Y$ zwei unabhängige Zufallsgrößen sind, was ist dann die Verteilung der Summe $X + Y$? Wir geben eine Antwort durch Summation über alle Werte, die $X$ annehmen kann. Die Ausnutzung der Unabhängigkeit führt uns auf natürliche Weise zum Begriff der Faltung. In diesem Abschnitt werden wir nur $\mathbb{Z}$-wertige Zufallsgrößen betrachten. Daher ist die Menge $\mathbb{Z}$ der natürliche Indexbereich der von uns betrachteten Folgen.

**Definition 4.1.1.** *Die* Faltung *zweier absolut summierbarer Folgen $a = (a_x)_{x\in\mathbb{Z}}$ und $b = (b_y)_{y\in\mathbb{Z}}$ ist die Folge $c = (c_z)_{z\in\mathbb{Z}}$, die gegeben ist durch $c_z = \sum_{x\in\mathbb{Z}} a_x b_{z-x}$. Wir schreiben $c = a \star b$.*

Man sieht leicht, dass $a \star b = b \star a$ und dass $a \star b$ eine absolut summierbare Folge ist, wenn $a$ und $b$ dies sind.

**Satz 4.1.2 (Faltungssatz).** *Seien $X$ und $Y$ zwei unabhängige $\mathbb{Z}$-wertige Zufallsgrößen mit Verteilungen $p_X$ und $p_Y$, d. h. $p_X(x) = \mathbb{P}(X = x)$ und $p_Y(y) = \mathbb{P}(Y = y)$ für alle $x, y \in \mathbb{Z}$. Dann ist die Verteilung von $X+Y$ gleich der Faltung $p_X \star p_Y$, d. h. $\mathbb{P}(X+Y = z) = (p_X \star p_Y)(z)$ für alle $z \in \mathbb{Z}$.*

**Beweis.** Wir summieren über alle Werte, die $X$ annehmen kann, und erhalten für alle $z \in \mathbb{Z}$:

$$\begin{aligned}\mathbb{P}(X+Y=z) &= \sum_{x\in\mathbb{Z}} \mathbb{P}(X=x, Y=z-x) = \sum_{x\in\mathbb{Z}} \mathbb{P}(X=x)\mathbb{P}(Y=z-x)\\ &= \sum_{x\in\mathbb{Z}} p_X(x)p_Y(z-x) = (p_X \star p_Y)(z).\end{aligned}$$

□

**Beispiel 4.1.3 (Binomialverteilung).** Die Summe zweier unabhängiger binomialverteiler Zufallsgrößen zu den Parametern $n_1$ und $p$ bzw. $n_2$ und $p$ ist eine zu den Parametern $n_1+n_2$ und $p$ binomialverteilte Zufallsgröße. Dies sieht man am einfachsten ein, indem man die beiden Zufallsgrößen jeweils als Summe von $n_1$ bzw. $n_2$ unabhängigen Bernoulli-Zufallsgrößen darstellt. Man sagt, die Familie der zum Parameter $n$ binomialverteilten Zufallsgrößen (mit festem zweiten Parameter $p$) bildet eine *Faltungshalbgruppe*.

Als eine Anwendung von Satz 4.1.2 beweisen wir die eingangs erwähnte Aussage noch einmal, indem wir zeigen, dass $\mathrm{Bi}_{n_1,p} \star \mathrm{Bi}_{n_2,p} = \mathrm{Bi}_{n_1+n_2,p}$ gilt, wobei wir die Folge $\mathrm{Bi}_{n,p}$ (die ja nur auf $\{0,\dots,n\}$ definiert ist) trivial mit Null zu einer Folge mit Indexmenge $\mathbb{Z}$ fortsetzen.

Es sei also $k \in \{0,\dots,n_1+n_2\}$, dann gilt

$$\begin{aligned}\mathrm{Bi}_{n_1,p} \star \mathrm{Bi}_{n_2,p}(k) &= \sum_{l\in\mathbb{Z}} \mathrm{Bi}_{n_1,p}(l)\mathrm{Bi}_{n_2,p}(k-l)\\ &= \sum_{l=\max\{0,k-n_2\}}^{\min\{n_1,k\}} \binom{n_1}{l} p^l (1-p)^{n_1-l} \binom{n_2}{k-l} p^{k-l}(1-p)^{n_2-k+l}\\ &= \mathrm{Bi}_{n_1+n_2,p}(k) \sum_{l=\max\{0,k-n_2\}}^{\min\{n_1,k\}} \frac{\binom{n_1}{l}\binom{n_2}{k-l}}{\binom{n_1+n_2}{k}},\end{aligned}$$

wie eine direkte Rechnung ergibt. Nun beachte man, dass der Quotient hinter dem Summenzeichen die hypergeometrische Wahrscheinlichkeit mit Parametern $n_1$, $n_2$ und $k$ ist, ausgewertet in $l$. Da der Summationsbereich genau derjenige ist, auf der diese Verteilung definiert ist, ist also die Summme gleich Eins. Dies beweist die Behauptung $\mathrm{Bi}_{n_1,p} \star \mathrm{Bi}_{n_2,p} = \mathrm{Bi}_{n_1+n_2,p}$ auf $\{0,\dots,n_1+n_2\}$, und man sieht leicht, dass sie trivialerweise auch in $\mathbb{Z}\setminus\{0,\dots,n_1+n_2\}$ erfüllt ist. ◇

**Beispiel 4.1.4 (Negative Binomialverteilung).** Es seien $X_1,\dots,X_n$ unabhängige, zum Parameter $p \in (0,1)$ auf $\mathbb{N}_0$ geometrisch verteilte Zufallsgrößen (siehe Beispiel 1.3.5), also $\mathbb{P}(X_i=k) = p(1-p)^k$ für $k \in \mathbb{N}_0$. Wir setzen $X = X_1+\cdots+X_n$, also hat $X$ die Verteilung $\widetilde{\mathrm{Geo}}_p^{\star n}$, womit wir die $n$-fache Faltung der geometrischen Verteilung auf $\mathbb{N}_0$ bezeichnen.

Wir behaupten, dass die Verteilung von $X$ identifiziert wird als

$$\mathbb{P}(X=k) = \binom{n+k-1}{k} p^n(1-p)^k = \binom{-n}{k} p^n(p-1)^k = \mathrm{Neg}_{n,p}(k), \qquad k\in\mathbb{N}_0, \tag{4.1.1}$$

wobei

$$\binom{-n}{k} = \frac{(-n)(-n-1)(-n-2)\dots(-n-k+1)}{k!}, \qquad n \in (0,\infty),$$

der allgemeine Binomialkoeffizient ist. Die Verteilung $\mathrm{Neg}_{n,p}$ ist unter dem Namen *negative Binomialverteilung* zu den Parametern $p$ und $n$ bekannt. Sie kann ohne Probleme auch für beliebiges $n \in (0,\infty)$ definiert werden, besitzt aber die Interpretation als Summe von geometrisch verteilten Zufallsgrößen nur für natürliche Zahlen $n$. Insbesondere ist $\mathrm{Neg}_{1,p} = \widetilde{\mathrm{Geo}}_p$.

Wir bieten nun zwei Wege an, die Faltungsformel

$$\widetilde{\mathrm{Geo}}_p^{\star n} = \mathrm{Neg}_{n,p}$$

für $n \in \mathbb{N}$ zu zeigen, ein dritter Weg (der sogar für alle $n \in (0,\infty)$ funktioniert), folgt in Beispiel 4.2.11. Der erste Weg macht Gebrauch von der oben erwähnten Interpretation von $X_i + 1$ als die Wartezeit nach dem $i$-ten bis zum $(i+1)$-ten Erfolg in einem unendlich langen Bernoulli-Experiment. Dann ist also $X+n$ die Wartezeit auf den $n$-ten Erfolg, gerechnet ab dem Beginn der Serie. Also ist das Ereignis $\{X = k\} = \{X + n = k + n\}$ das Ereignis, dass unter den ersten $k+n-1$ Spielen genau $n-1$ Erfolge und $k$ Misserfolge sind und dass das $(n+k)$-te Spiel erfolgreich ist. Jede einzelne dieser Serien der Länge $n+k$ hat die Wahrscheinlichkeit $p^n(1-p)^k$, und es gibt genau $\binom{n-1+k}{k}$ solche Serien. Dies zeigt, dass die Verteilung von $X$ tatsächlich durch die Formel in (4.1.1) gegeben ist.

Der zweite Weg benutzt den Faltungssatz und eine kombinatorische Überlegung. Man errechnet leicht, dass für $n_1, n_2 \in \mathbb{N}_0$ gilt:

$$\begin{aligned}
\mathrm{Neg}_{n_1,p} \star \mathrm{Neg}_{n_2,p}(k) &= \sum_{l\in\mathbb{Z}} \mathrm{Neg}_{n_1,p}(l)\mathrm{Neg}_{n_2,p}(k-l) \\
&= \sum_{l=0}^{k} \binom{n_1 - 1 + l}{l} p^{n_1}(1-p)^l \binom{n_2 - 1 + k - l}{k-l} p^{n_2}(1-p)^{k-l} \\
&= \mathrm{Neg}_{n_1+n_2,p}(k) \sum_{l=0}^{k} \frac{\binom{n_1-1+l}{l}\binom{n_2-1+k-l}{k-l}}{\binom{n_1+n_2-1+k}{k}}.
\end{aligned}$$

Dass die Summe über $l$ den Wert Eins hat, sieht man folgendermaßen ein. Die Zahl $\binom{n_1-1+l}{l}$ ist die Anzahl der Möglichkeiten, $n_1$ Einsen und $l$ Nullen hinter einander in eine Reihe zu legen, so dass die Reihe mit einer Eins endet. Analog ist $\binom{n_2-1+k-l}{k-l}$ die Anzahl der Möglichkeiten, $n_2$ Einsen und $k-l$ Nullen in eine Reihe zu legen, so dass sie mit einer Eins endet. Wenn man je eine solche Reihe hintereinander legt für irgendein $l \in \{0,\dots,k\}$, so erhält man eine Reihe mit $n_1+n_2$ Einsen und $k$ Nullen, die mit einer Eins endet. Anders herum kann man jede solche Reihe eindeutig aufspalten in zwei Reihen mit $n_1$ bzw. $n_2$ Einsen und $l$ bzw. $k-l$ Nullen für ein geeignetes $l \in \{0,\dots,k\}$, so dass diese beiden Teilreihen jeweils mit einer Eins enden. Dies zeigt auf kombinatorische Weise, dass $\sum_{l=0}^{k} \binom{n_1-1+l}{l}\binom{n_2-1+k-l}{k-l} = \binom{n_1+n_2-1+k}{k}$ für $n_1, n_2 \in \mathbb{N}_0$. ◇

**Beispiel 4.1.5 (Poisson-Verteilung).** Eine weitere Faltungshalbgruppe ist die der Poisson-Verteilungen: Die Summe zweier unabhängiger zum Parameter $\alpha > 0$ bzw. $\beta > 0$ Poisson-verteilter Zufallsgrößen ist eine zum Parameter $\alpha+\beta$ Poisson-verteilte Zufallsgröße. Den Beweis führt man wiederum mit Hilfe von Satz 4.1.2 (Übungsaufgabe). Ein eleganterer Beweis folgt in Beispiel 4.2.12. ◇

## 4.2 Erzeugende Funktionen

In diesem Abschnitt betrachten wir ausschließlich Verteilungen auf $\mathbb{N}_0 = \{0,1,2,\dots,\}$ und kombinieren ihre Betrachtung mit der gewisser Potenzreihen. Auf diesem Wege setzen wir einige

bekannte Ergebnisse aus der Analysis für die Beschreibung von Zufallsgrößen ein.

**Definition 4.2.1 (erzeugende Funktion).** *Die* erzeugende Funktion *einer Wahrscheinlichkeitsverteilung $(p_k)_{k\in\mathbb{N}_0}$ auf $\mathbb{N}_0$ ist die Funktion $\varphi$, die gegeben ist durch*

$$\varphi(s) = \sum_{k\in\mathbb{N}_0} p_k s^k, \qquad |s| < 1.$$

**Bemerkung 4.2.2.** Da die Reihe der $p_k$ absolut konvergiert, hat die zugehörige erzeugende Funktion $\varphi$ mindestens den Konvergenzradius Eins, d. h., sie konvergiert mindestens im Innern des (komplexen) Einheitskreises. Insbesondere kann man die Potenzreihe beliebig oft im Intervall $(-1, 1)$ gliedweise differenzieren. Auf Grund des Satzes von Taylor kann man aus der erzeugenden Funktion mit Hilfe der Formel

$$p_k = \frac{\varphi^{(k)}(0)}{k!}, \qquad k \in \mathbb{N}_0,$$

eindeutig die Koeffizienten $p_k$ erhalten, wobei $\varphi^{(k)}$ die $k$-te Ableitung bedeutet. Es besteht also ein eineindeutiger Zusammenhang zwischen der Verteilung und ihrer erzeugenden Funktion. ◇

**Beispiel 4.2.3 (Binomialverteilung).** Für $n \in \mathbb{N}$ und $p \in [0, 1]$ ist die erzeugende Funktion der Binomialverteilung zu den Parametern $n$ und $p$ (siehe Beispiel 1.3.4) gegeben durch

$$\varphi(s) = \sum_{k=0}^{n} \mathrm{Bi}_{n,p}(k) s^k = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} s^k = (1 - p + sp)^n,$$

wobei wir den binomischen Lehrsatz benutzten. ◇

**Beispiel 4.2.4 (Negative Binomialverteilung).** Die erzeugende Funktion einer negativ zu den Parametern $n \in (0, \infty)$ und $p \in [0, 1]$ binomialverteilten Zufallsgröße $X$ (siehe Beispiel 4.1.4) errechnet man als

$$\varphi(s) = p^n \sum_{k\in\mathbb{N}_0} \binom{-n}{k} (p-1)^k s^k = \Big(\frac{p}{1+sp-s}\Big)^n \sum_{k\in\mathbb{N}_0} \mathrm{Neg}_{n,1+sp-s}(k) = \Big(\frac{p}{1+sp-s}\Big)^n,$$

zunächst nur für $s \in \mathbb{R}$ mit $1 + sp - s \in (0, 1]$. Doch da beide Seiten der Gleichung analytisch im Einheitskreis sind, gilt sie auch dort. ◇

**Beispiel 4.2.5 (Poisson-Verteilung).** Die erzeugende Funktion der Poisson-Verteilung zum Parameter $\alpha \in (0, \infty)$ (siehe Beispiel 1.3.6) ist gegeben durch

$$\varphi(s) = \sum_{k\in\mathbb{N}_0} \mathrm{Po}_\alpha(k) s^k = \sum_{k\in\mathbb{N}_0} e^{-\alpha} \frac{\alpha^k}{k!} s^k = e^{-\alpha(1-s)}.$$

◇

Da die erzeugende Funktion die Verteilung eindeutig fest legt, ist es klar, dass auch der Erwartungswert und die Varianz der Verteilung mit Hilfe der erzeugenden Funktion ausgedrückt werden können. Mit $\mathbb{E}(P) = \sum_{k\in\mathbb{N}_0} k p_k$ und $\mathbb{V}(P) = \sum_{k\in\mathbb{N}_0} (k - \mathbb{E}(P))^2 p_k$ bezeichnen wir den Erwartungswert und die Varianz einer Verteilung $P = (p_k)_{k\in\mathbb{N}_0}$ auf $\mathbb{N}_0$.

**Satz 4.2.6.** *Es sei $P = (p_k)_{k\in\mathbb{N}_0}$ eine Verteilung auf $\mathbb{N}_0$ mit erzeugender Funktion $\varphi$.*

*(i) $\mathbb{E}(P)$ existiert genau dann, wenn $\varphi'(1-) = \lim_{s\uparrow 1}\varphi'(s)$ existiert, und dann gilt $\mathbb{E}(P) = \varphi'(1-)$.*

*(ii) $\mathbb{V}(P)$ existiert genau dann, wenn $\varphi''(1-) = \lim_{s\uparrow 1}\varphi''(s)$ existiert, und dann gilt $\mathbb{V}(P) = \varphi''(1-) - \mathbb{E}(P)^2 + \mathbb{E}(P)$.*

**Beweis.**

(i) Für $|s| < 1$ existiert $\varphi'(s)$, und für $s \uparrow 1$ gilt

$$\varphi'(s) = \sum_{k\in\mathbb{N}_0} p_k k s^{k-1} \uparrow \sum_{k\in\mathbb{N}_0} p_k k = \mathbb{E}(P).$$

(ii) Für $|s| < 1$ existiert $\varphi''(s)$, und für $s \uparrow 1$ gilt

$$\varphi''(s) = \sum_{k\in\mathbb{N}_0} p_k k(k-1) s^{k-2} \uparrow \sum_{k\in\mathbb{N}_0} p_k (k^2 - k).$$

Es ist leicht zu sehen, dass die Reihe $\sum_{k\in\mathbb{N}_0} p_k(k^2-k)$ genau dann konvergiert, wenn $\sum_{k\in\mathbb{N}_0} p_k k^2$ konvergiert, d. h., wenn $\mathbb{V}(P)$ existiert. In diesem Fall errechnet man leicht, dass

$$\varphi''(1-) - \mathbb{E}(P)^2 + \mathbb{E}(P) = \sum_{k\in\mathbb{N}_0} p_k k^2 - \mathbb{E}(P)^2 = \mathbb{V}(P).$$

□

**Beispiel 4.2.7.** Indem man die in den Beispielen 4.2.3 bzw. 4.2.5 errechneten erzeugenden Funktionen der Binomial- bzw. Poisson-Verteilung bei Eins ein bzw. zwei Mal differenziert, erhält man als eine Übungsaufgabe, dass sie die Erwartungswerte $np$ bzw. $\alpha$ und die Varianzen $np(1-p)$ bzw. $\alpha$ besitzen. (Natürlich stehen diese Ergebnisse in Übereinstimmung mit Beispielen 3.3.4 und 3.5.4 bzw. 3.3.6 und 3.4.5.) ◇

Den Zusammenhang zwischen Verteilungen auf $\mathbb{N}_0$ und den zugehörigen erzeugenden Funktionen kann man auch ausnützen bei der Behandlung von Summen unabhängiger Zufallsgrößen. Wir bezeichnen nun die erzeugende Funktion einer Verteilung $P = (p_k)_{k\in\mathbb{N}_0}$ auf $\mathbb{N}_0$ mit $\varphi_P$. Die erzeugende Funktion $\varphi_X$ einer $\mathbb{N}_0$-wertigen Zufallsvariablen $X$ ist definiert als die erzeugende Funktion der Verteilung von $X$, also

$$\varphi_X(s) = \varphi_{\mathbb{P}\circ X^{-1}}(s) = \sum_{k\in\mathbb{N}_0} \mathbb{P}(X = k) s^k = \mathbb{E}(s^X), \qquad |s| < 1.$$

Wir erinnern daran (siehe Abschnitt 4.1), dass die Faltung $P_1 \star P_2$ zweier Verteilungen $P_1$ und $P_2$ auf $\mathbb{N}_0$ die Verteilung der Summe zweier unabhängiger Zufallsgrößen mit Verteilung $P_1$ bzw. $P_2$ ist. Es stellt sich heraus, dass die erzeugende Funktion dieser Summe gleich dem punktweisen Produkt der beiden erzeugenden Funktionen ist:

**Satz 4.2.8 (Faltung und Unabhängigkeit).** *(i) Für je zwei unabhängige $\mathbb{N}_0$-wertige Zufallsgrößen $X_1$ und $X_2$ gilt*

$$\varphi_{X_1+X_2}(s) = \varphi_{X_1}(s)\varphi_{X_2}(s), \qquad |s| < 1.$$

*(ii) Für je zwei Verteilungen $P_1$ und $P_2$ auf $\mathbb{N}_0$ gilt*

$$\varphi_{P_1 \star P_2}(s) = \varphi_{P_1}(s)\varphi_{P_2}(s), \qquad |s| < 1.$$

**Beweis.** (i) Nach Korollar 3.2.12 sind auch die beiden Zufallsvariablen $s^{X_1}$ und $s^{X_2}$ unabhängig. Nach dem Produktsatz für Erwartungswerte bei Unabhängigkeit (siehe Satz 3.3.2(d)) haben wir

$$\varphi_{X_1+X_2}(s) = \mathbb{E}(s^{X_1+X_2}) = \mathbb{E}(s^{X_1}s^{X_2}) = \mathbb{E}(s^{X_1})\mathbb{E}(s^{X_2}) = \varphi_{X_1}(s)\varphi_{X_2}(s).$$

Die Aussage in (ii) ist der Spezialfall von (i) für identische Zufallsgrößen $X_1$ und $X_2$. □

**Bemerkung 4.2.9.** Den Satz 4.2.8 kann man statt auf probabilistischem Wege auch mit Hilfe aus der Analysis beweisen, wenn man benutzt, dass das punktweise Produkt $\varphi_{X_1}(s)\varphi_{X_2}(s)$ wieder eine Potenzreihe ist, deren Koeffizientenfolge die Faltung der Koeffizientenfolgen von $\varphi_{X_1}(s)$ und $\varphi_{X_2}(s)$ sind, also

$$\begin{aligned}\varphi_{X_1}(s)\varphi_{X_2}(s) &= \Big(\sum_{k=0}^{\infty} p_{X_1}(k)s^k\Big)\Big(\sum_{k=0}^{\infty} p_{X_2}(k)s^k\Big) = \sum_{k=0}^{\infty}(p_{X_1} \star p_{X_2})(k)s^k \\ &= \sum_{k=0}^{\infty} p_{X_1+X_2}(k)s^k = \varphi_{X_1+X_2}(s),\end{aligned}$$

wobei im vorletzten Schritt der Faltungssatz 4.1.2 benutzt wurde. ◇

Die bemerkenswerte Aussage von Satz 4.2.8 gibt uns ein weiteres Mittel in die Hand, Verteilungen von Summen unabhängiger Zufallsgrößen zu identifizieren:

**Beispiel 4.2.10 (Binomialverteilung).** Die erzeugende Funktion der Binomialverteilung mit Parametern $n$ und $p$ (siehe Beispiel 4.2.3) ist offensichtlich das $n$-fache Produkt der erzeugenden Funktion einer Binomialverteilung mit Parametern 1 und $p$. Dies reflektiert die bekannte Tatsache (siehe Beispiel 4.1.3) dass eine binomialverteilte Zufallsgröße eine Summe unabhängiger Bernoulli-Zufallsgrößen ist, oder auch die Summe unabhängiger binomialverteilter Zufallsgrößen mit gewissen Parametern. ◇

**Beispiel 4.2.11 (Negative Binomialverteilung).** Wie man im Beispiel 4.2.4 gesehen hat, ist auch die erzeugende Funktion der Negativen Binomialverteilung mit Parametern $n \in (0,\infty)$ und $p \in [0,1]$ die $n$-te Potenz derselben Verteilung mit Parametern 1 und $p$. Daher ist die Summe unabhängiger negativ zu den Parametern $n_1 \in (0,\infty)$ und $p$ bzw. $n_2 \in (0,\infty)$ und $p$ binomialverteilter Zufallsgrößen negativ binomialverteilt zu den Parametern $n_1+n_2$ und $p$. Also gilt insbesondere die Faltungsformel

$$\mathrm{Neg}_{n_1,p} \star \mathrm{Neg}_{n_2,p} = \mathrm{Neg}_{n_1+n_2,p},$$

die wir in Beispiel 4.1.4 nur für natürliche Zahlen $n_1$ und $n_2$ bewiesen. ◇

**Beispiel 4.2.12 (Poisson-Verteilung).** Ein sehr eleganter und kurzer Beweis für die Tatsache (siehe Beispiel 4.1.5), dass die Summe zweier unabhängiger Poisson-verteilter Zufallsgrößen $X$ und $Y$ mit Parametern $\alpha$ und $\beta$ Poisson-verteilt ist mit Parameter $\alpha+\beta$, ist nun mit Satz 4.2.8 und Beispiel 4.2.5 möglich: Die erzeugende Funktion der Summe ist gegeben durch

$$\varphi_{X+Y}(s) = \varphi_X(s)\varphi_Y(s) = e^{-\alpha(s-1)}e^{-\beta(s-1)} = e^{-(\alpha+\beta)(s-1)}.$$

Also wird die erzeugende Funktion von $X+Y$ identifiziert mit der einer Poisson-Verteilung mit Parameter $\alpha+\beta$. Wegen der Eindeutigkeit der erzeugenden Funktion ist $X+Y$ also Poisson-verteilt mit diesem Parameter. ◇

## 4.3 Die eindimensionale Irrfahrt

In diesem Abschnitt führen wir ein grundlegendes stochastisches Modell ein, das sehr einfach definiert wird, aber eine Fülle von interessanten Fragestellungen aufwirft und eine sehr große Zahl von Anwendungen in verschiedenen Bereichen hat. Wir betrachten einen Punkt auf dem eindimensionalen Gitter $\mathbb{Z}$, der zum Zeitpunkt Null im Ursprung startet und danach zu den diskreten Zeitpunkten $1, 2, 3, \dots$ jeweils einen Sprung zu einem der beiden Nachbarn seines aktuellen Aufenthaltortes ausführt. Die Sprungentscheidungen werden zufällig und unabhängig getroffen, und die Wahrscheinlichkeiten der beiden möglichen Sprungentscheidungen sind jeweils gleich $\frac{1}{2}$. Dieses Modell nennt man die *eindimensionale symmetrische Irrfahrt*. Es gibt Interpretationen und Anwendungen dieses Modells als Auszählungen der Stimmen für zwei konkurrierende Kandidaten, als den zufälligen Weg eines kleinen Teilchens durch eine eindimensionale Umgebung, als einen Aktienkurs in diskreter Zeit und vieles mehr. Für die meisten dieser Interpretationen ist dieses Modell natürlich viel zu simpel, und man betrachtet dann oft kompliziertere Versionen, aber dies soll uns hier nicht beschäftigen.

Mathematisch kann man das Modell einführen, indem man eine Folge $X_1, X_2, \dots$ von unabhängigen symmetrisch verteilten $\{-1, 1\}$-wertigen Zufallsgrößen betrachtet (das sind die Größen der Sprünge zu den Zeitpunkten $1, 2, \dots$) und dann die Position des Teilchens zum Zeitpunkt $n \in \mathbb{N}_0$ mit $S_n = X_1 + X_2 + \cdots + X_n$ angibt.[1] Der Pfad $(S_0, S_1, \dots, S_n)$, den das Teilchen bis zum Zeitpunkt $n$ zurück gelegt hat, ist also ein Element des Raumes

$$\Omega_n = \{(s_0, \dots, s_n) \in \mathbb{Z}^{n+1} \colon s_0 = 0 \text{ und } |s_i - s_{i-1}| = 1 \text{ für alle } i = 1, \dots, n\},$$

und jedes Element in $\Omega_n$ hat die selbe Wahrscheinlichkeit $2^{-n}$. Das zugehörige Wahrscheinlichkeitsmaß bezeichen wir mit $\mathbb{P}_n$. Wir werden manchmal anschaulich argumentieren und einen Sprung der Größe 1 bzw. $-1$ als einen Sprung aufwärts bzw. abwärts interpretieren.

Uns interessieren die folgenden Fragen:

(i) Mit welcher Wahrscheinlichkeit ist das Teilchen zum Zeitpunkt $n$ im Ursprung?

(ii) Mit welcher Wahrscheinlichkeit erreicht das Teilchen bis zum Zeitpunkt $n$ einen gegebenen Tiefpunkt?

(iii) Mit welcher Wahrscheinlichkeit war das Teilchen nie in $-\mathbb{N}$ bis zum Zeitpunkt $n$?

[1] Wie schon oft bemerkt, können wir nicht alle Zufallsvariablen $X_1, X_2, \dots$ gleichzeitig auf einem Wahrscheinlichkeitsraum definieren, aber für die korrekte Behandlung von $S_n$ reicht unser mathematischer Apparat vollauf aus.

(iv) Wieviel Zeit verbringt das Teilchen mit welcher Wahrscheinlichkeit in $\mathbb{N}$?

Da wir auf $\Omega_n$ die Gleichverteilung betrachten, müssen wir also effektive Zählmethoden entwickeln, und dies soll in diesem Kapitel getan werden.

Ferner sind wir an den Asymptoten gewisser Wahrscheinlichkeiten für große $n$ interessiert. Das wichtigste analytische Hilfsmittel hierbei wird die *Stirlingsche Formel* sein, die besagt:

$$n! \sim \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n, \qquad n \to \infty, \tag{4.3.1}$$

wobei wir wie immer $\sim$ für asymptotische Äquivalenz benutzen.

Wenden wir uns zunächst der Frage zu, mit welcher Wahrscheinlichkeit das Teilchen zum Zeitpunkt $n$ sich in einem gegebenen Punkt befindet, d. h., die Wahrscheinlichkeit des Ereignisses $\{S_n = i\}$ für $i \in \mathbb{Z}$. Das folgende Lemma ist nahezu offensichtlich.

**Lemma 4.3.1.** *Für alle $n \in \mathbb{N}$ und $i \in \mathbb{Z}$ gilt*

$$\mathbb{P}_n(S_n = i) = \begin{cases} 0 & \text{falls } n+i \text{ ungerade oder } |i| > n, \\ 2^{-n}\binom{n}{(n+i)/2} & \text{sonst.} \end{cases}$$

**Beweis.** Es ist klar, dass das Ereignis $\{S_n = i\}$ nicht eintreten kann, wenn $|i| > n$.

Da das Teilchen zu jedem Zeitpunkt um eine Einheit springen muss, kann es zu einem geraden Zeitpunkt nicht in einem ungeraden Raumpunkt sein und umgekehrt. Falls $n+i$ gerade ist, dann muss das Teilchen, um zum Zeitpunkt $n$ in $i$ zu sein, genau $(n+i)/2$ Mal aufwärts springen und genau $(n-i)/2$ Mal abwärts. Es gibt genau $\binom{n}{(n+i)/2}$ Pfade, die dies tun. □

Also können wir schon die Frage nach der Aufenthaltswahrscheinlichkeit in Null sowie die Asymptotik dieser Wahrscheinlichkeit wie folgt beantworten. Wir definieren

$$u_{2n} = \mathbb{P}_{2n}(S_{2n} = 0), \qquad n \in \mathbb{N}.$$

Das folgende Korollar beweist man als Übungsaufgabe mit Hilfe von Lemma 4.3.1 und der Stirlingschen Formel in (4.3.1).

**Korollar 4.3.2.** *Es gilt*

$$u_{2n} = 2^{-2n}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}, \qquad n \to \infty.$$

Nun wenden wir uns der Verteilung des Minimums des Pfades zu, also der Zufallsgröße

$$M_n = \min\{S_0, \ldots, S_n\}.$$

Eines der wichtigsten Hilfsmittel hierfür ist das *Spiegelungsprinzip*, das durch geschicktes Spiegeln eines Teils des Pfades einen Vergleich zwischen gewissen Pfadklassen herstellt. Im Folgenden bestimmen wir die Wahrscheinlichkeit der Menge der Pfade, die den Punkt $j \in -\mathbb{N}_0$ erreichen und nach insgesamt $n$ Schritten in $i \geq j$ enden. Zur Veranschaulichung des folgenden Prinzips ist eine Skizze hilfreich.

**Lemma 4.3.3 (Spiegelungsprinzip).** *Für alle $n \in \mathbb{N}$ und alle $i, j \in \mathbb{Z}$ mit $j \leq 0$ und $i \geq j$ gilt $\mathbb{P}(M_n \leq j, S_n = i) = \mathbb{P}(S_n = i - 2j)$.*

**Beweis.** Wir brauchen nur den Fall zu betrachten, dass $n + i$ gerade ist, sonst sind beide betrachteten Ereignisse leer.

Für einen Pfad $s$ in $\{M_n \leq j, S_n = i\}$ betrachten wir das kleinste $k \in \{1, \ldots, n\}$ mit $s_k = j$, also den ersten Zeitpunkt, an dem das Teilchen den Wert $j$ erreicht. Nun spiegeln wir das Pfadstück $(s_k, \ldots, s_n)$ an $s_k = j$ und erhalten einen Pfad $\tilde{s} = (\tilde{s}_0, \ldots, \tilde{s}_n) \in \Omega_n$ mit $\tilde{s}_n = 2j - i$. Dieser Pfad liegt also in dem Ereignis $\{S_n = 2j - i\}$.

Nun überlegt man sich leicht, dass die oben durchgeführte Abbildung (Spiegeln ab dem ersten Erreichenszeitpunkt des Niveaus $j$) sogar eine bijektive Abbildung zwischen den Ereignissen $\{M_n \leq j, S_n = i\}$ und $\{S_n = 2j - i\}$ ist. Die Umkehrabbildung erhält man, indem man einen Pfad aus $\{S_n = 2j - i\}$ ab dem ersten Zeitpunkt, an dem er das Niveau $j$ erreicht, spiegelt. (Er muss dieses Niveau spätestens zum Zeitpunkt $n$ erreichen, da $2j - i \leq j \leq 0$.) Also enthalten die Mengen $\{M_n \leq j, S_n = i\}$ und $\{S_n = 2j - i\}$ die selbe Anzahl von Pfaden. Daraus folgt die Behauptung, denn $\mathbb{P}(M_n \leq j, S_n = i) = \mathbb{P}(S_n = 2j - i) = \mathbb{P}(S_n = i - 2j)$. □

Mit Hilfe des Spiegelungsprinzips können wir nun die gemeinsame Verteilung des Endpunkts $S_n$ und des Minimums $M_n$ bestimmen:

**Satz 4.3.4 (Verteilung des Minimums des Pfades).** *Für alle $n \in \mathbb{N}$ und alle $i, j \in \mathbb{Z}$ mit $j \leq 0$ und $i \geq j$ gelten:*

$$\mathbb{P}_n(M_n = j, S_n = i) = \mathbb{P}_n(S_n = i - 2j) - \mathbb{P}_n(S_n = i - 2j + 2),$$
$$\mathbb{P}_n(M_n = j) = \mathbb{P}_n(S_n \in \{j, j - 1\}).$$

**Beweis.** Übungsaufgabe.

□

Aus Symmetriegründen erhält man aus Satz 4.3.4 natürlich auch die gemeinsame Verteilung des Endpunkts und des Maximums des Pfades.

Nun betrachten wir die Ereignisse, dass das Teilchen erst nach $2n$ Schritten zum Ursprung zurück kehrt bzw. nicht mehr bzw. nie den negativen Bereich betritt:

$$A_{2n} = \{S_1 \neq 0, \ldots, S_{2n-1} \neq 0, S_{2n} = 0\},$$
$$B_{2n} = \{S_i \neq 0 \text{ für alle } i \in \{1, \ldots, 2n\}\},$$
$$C_{2n} = \{S_i \geq 0 \text{ für alle } i \in \{1, \ldots, 2n\}\}.$$

Wir erinnern daran, dass $u_{2n}$ die Wahrscheinlichkeit dafür ist, dass das Teilchen zum Zeitpunkt $2n$ sich in Null befindet. Offensichtlich ist also $\mathbb{P}_{2n}(A_{2n}) \leq u_{2n}$. In den beiden Ereignissen $B_{2n}$ und $A_{2n}$ kann sich das Teilchen im Zeitintervall $\{1, \ldots, 2n - 1\}$ entweder in $\mathbb{N}$ oder in $-\mathbb{N}$ aufhalten.

**Lemma 4.3.5.** *Für jedes $n \in \mathbb{N}$ gelten die Beziehungen*

$$\mathbb{P}_{2n}(A_{2n}) = \frac{1}{2n} u_{2n-2} = u_{2n-2} - u_{2n}, \tag{4.3.2}$$
$$\mathbb{P}_{2n}(B_{2n}) = u_{2n}, \tag{4.3.3}$$
$$\mathbb{P}_{2n}(C_{2n}) = u_{2n}. \tag{4.3.4}$$

**Beweis.** (4.3.2): Wir zählen die Pfade in $A_{2n}$, die im positiven Bereich verlaufen, und multiplizieren deren Anzahl mit 2. Ein solcher Pfad ist zu den Zeitpunkten 1 und $2n-1$ in 1. Die Zahl der $(2n-2)$-schrittigen in 1 startenden und endenden Pfade, die nie die Null betreten, ist gleich der Zahl der $(2n-2)$-schrittigen in 1 startenden und endenden Pfade minus die Zahl solcher Pfade, die zwischendurch die Null betreten. Diese beiden Anzahlen sind gleich der Anzahl der Pfade in dem Ereignis $\{S_{2n-2}=0\}$ bzw. in $\{M_{2n-2}\le -1, S_{2n-2}=0\}$. Also haben wir

$$\begin{aligned}\mathbb{P}_{2n}(A_{2n}) &= 2\mathbb{P}_{2n}(S_1>0,\dots,S_{2n-1}>0,S_{2n}=0)\\ &= 2^{-2n+1}\big(|\{S_{2n-2}=0\}| - |\{M_{2n-2}\le -1, S_{2n-2}=0\}|\big).\end{aligned}$$

Nach dem Spiegelungsprinzip ist der letzte Term gleich $|\{S_{2n-2}=2\}|$. Mit Hilfe von Lemma 4.3.1 können wir nun ausrechnen:

$$\mathbb{P}_{2n}(A_{2n}) = 2^{-2(n-1)}\frac{1}{2}\Big(\binom{2n-2}{n-1} - \binom{2n-2}{n}\Big) = \frac{1}{2n}u_{2n-2}.$$

Dies beweist die erste Gleichung in (4.3.2); die zweite rechnet man leicht nach.

(4.3.3): Das Gegenereignis von $B_{2n}$ ist das Ereignis, dass das Teilchen zu einem der Zeitpunkte $2j$ mit $j\in\{1,\dots,n\}$ zum ersten Mal zurück zur Null kehrt, also die disjunkte Vereinigung der Ereignisse $\{S_1\neq 0,\dots,S_{2j-1}\neq 0, S_{2j}=0\}$. Man überlegt sich leicht, dass dieses Ereignis die Wahrscheinlichkeit $\mathbb{P}_{2j}(A_{2j})$ hat. Wenn wir (4.3.2) für $j$ an Stelle von $n$ anwenden, erhalten wir also, dass

$$\mathbb{P}_{2n}(B_{2n}) = 1-\sum_{j=1}^{n}\mathbb{P}_{2j}(A_{2j}) = 1-\sum_{j=1}^{n}(u_{2(j-1)}-u_{2j}) = u_{2n}.$$

(4.3.4): Übungsaufgabe. □

**Bemerkung 4.3.6 (Rekurrenz und Nullrekurrenz).** Eine interessante Folgerung ergibt sich aus (4.3.3) in Kombination mit der Tatsache $\lim_{n\to\infty}u_{2n}=0$ (siehe Korollar 4.3.2). Wir betrachten den ersten Zeitpunkt einer Rückkehr zum Startpunkt:

$$T = \inf\{k\in\mathbb{N}\colon S_k=0\}\in\mathbb{N}_0\cup\{\infty\}.$$

(Da die Betrachtung dieser Zufallsgröße die Existenz einer unendlich langen Irrfahrt voraus setzt, können wir sie mit unserem mathematischen Apparat nicht korrekt behandeln, siehe die Vorlesung *Stochastik I.*) Die Wahrscheinlichkeit $\mathbb{P}(T>2n)$, bis zum Zeitpunkt $2n$ nicht mehr zum Startpunkt zurück zu kehren, ist die Wahrscheinlichkeit des Ereignisses $B_{2n}$, tendiert also nach (4.3.3) gegen Null für $n\to\infty$. Man sollte daraus schließen (und es ist auch völlig richtig), dass $\mathbb{P}(T<\infty)=1$. Die Interpretation ist, dass das Teilchen mit Sicherheit irgendwann einmal wieder zum Ursprung zurück kehren wird. Diese Aussage nennt man die *Rekurrenz* der eindimensionalen symmetrischen Irrfahrt.

Eine weitere heuristische Folgerung aus Lemma 4.3.5 betrifft die erwartete Rückkehrzeit zum Ursprung, d. h. der Erwartungswert $\mathbb{E}(T)$ von $T$. Wir gehen davon aus (und das ist auch völlig in Ordnung), dass $\mathbb{E}(T)$ gegeben ist durch die Reihe $\sum_{k=1}^{\infty}k\mathbb{P}(T=k)$. Das Ereignis $\{T=2n\}$ ist identisch mit dem Ereignis $A_{2n}$. Also liefert die erste Gleichung in (4.3.3), dass diese Reihe divergiert:

$$\mathbb{E}(T) = \sum_{n=1}^{\infty}2n\mathbb{P}(A_{2n}) = \sum_{n=1}^{\infty}2n\frac{1}{2n}u_{2n-2} = \sum_{n=1}^{\infty}u_{2n-2},$$

und wegen Korollar 4.3.2 divergiert diese Reihe. Die Interpretation ist also, dass das Teilchen zwar mit Sicherheit zum Ursprung zurück kehren wird, dafür aber erwartungsgemäß unendlich viel Zeit benötigen wird. Diese Eigenschaft nennt man die *Nullrekurrenz* der eindimensionalen symmetrischen Irrfahrt. ◇

Nun bearbeiten wir die Frage, mit welcher Wahrscheinlichkeit das Teilchen sich eine gegebene Anzahl von Zeitpunkten im positiven Bereich aufhält. Wir betrachten also die Zufallsgröße, die die Zeit angibt, die der zufällige $2n$-schrittige Pfad im positiven Bereich verbringt:

$$X_{2n} = 2|\{i \in \{1, \dots, n\} \colon S_{2i-1} > 0\}|.$$

Die Zufallsgröße $X_{2n}$ nimmt nur Werte in $\{0, 2, 4, \dots, 2n\}$ an. Das Ereignis $\{X_{2n} = 0\}$ ist das Ereignis $\{S_i \leq 0$ für alle $i \in \{1, \dots, 2n\}\}$, und das Ereignis $\{X_{2n} = 2n\}$ ist identisch mit $\{S_i \geq 0$ für alle $i \in \{1, \dots, 2n\}\}$, also mit $C_{2n}$. Nach (4.3.4) ist also $\mathbb{P}_{2n}(X_{2n} = 0) = \mathbb{P}_{2n}(X_{2n} = 2n\} = u_{2n}$. Außerdem ist die Abbildung $j \mapsto \mathbb{P}_{2n}(X_{2n} = 2j)$ symmetrisch in dem Sinne, dass $\mathbb{P}_{2n}(X_{2n} = 2j) = \mathbb{P}_{2n}(X_{2n} = 2(n-j))$. Wir bestimmen nun die Verteilung von $X_{2n}$:

**Lemma 4.3.7.** *Für jedes $n \in \mathbb{N}$ und alle $j \in \{0, \dots, n\}$ gilt $\mathbb{P}_{2n}(X_{2n} = 2j) = u_{2j}u_{2(n-j)}$.*

**Beweis.** Wir führen eine Induktion nach $n$. Der Fall $n = 1$ ist klar.

Wir nehmen nun an, die Aussage trifft zu für alle $k \leq n-1$, und wir beweisen sie für $n$. Oben hatten wir schon darauf hin gewiesen, dass die Aussage für $j = 0$ und für $j = n$ zutrifft, also behandeln wir nur noch den Fall $1 \leq j \leq n-1$. Im Ereignis $\{X_{2n} = 2j\}$ muss das Teilchen zu einem der Zeitpunkte $2, 4, 6, \dots, 2n-2$ in Null sein, und wir spalten auf nach dem ersten solchen Zeitpunkt und danach, ob der Pfad zuvor im positiven oder im negativen Bereich verläuft. Diesen Zeitpunkt nennen wir $2l$. Im ersten Fall (d. h., wenn der Pfad bis $2l$ in $\mathbb{N}$ verläuft) muss $l \leq j$ gelten, und nach dem Zeitpunkt $2l$ bleibt er genau $2(j-l)$ Zeitpunkte im positiven Bereich. Im zweiten Fall muss $l \leq n-j$ gelten, denn nach dem Zeitpunkt $2l$ muss der Pfad ja noch genau $2j$ Zeitpunkte im Positiven bleiben. Diese Überlegungen führen zu der Formel

$$\begin{aligned}\mathbb{P}_{2n}(X_{2n} = 2j) = &\sum_{l=1}^{j} \mathbb{P}_{2l}(S_1 > 0, \dots, S_{2l-1} > 0, S_{2l} = 0)\mathbb{P}_{2(n-l)}(X_{2(n-l)} = 2(j-l)) \\ &+ \sum_{l=1}^{n-j} \mathbb{P}_{2l}(S_1 < 0, \dots, S_{2l-1} < 0, S_{2l} = 0)\mathbb{P}_{2(n-l)}(X_{2(n-l)} = 2j).\end{aligned}$$

Nun können wir die Induktionsvoraussetzung einsetzen und erhalten

$$\begin{aligned}\mathbb{P}_{2n}(X_{2n} = j) = &\frac{1}{2}u_{2(n-j)} \sum_{l=1}^{j} \mathbb{P}_{2l}(S_1 \neq 0, \dots, S_{2l-1} \neq 0, S_{2l} = 0)u_{2(j-l)} \\ &+ \frac{1}{2}u_{2j} \sum_{l=1}^{n-j} \mathbb{P}_{2l}(S_1 \neq 0, \dots, S_{2l-1} \neq 0, S_{2l} = 0)u_{2(n-j-l)}.\end{aligned}$$

Nun beachte man, dass die erste Summe die Wahrscheinlichkeit der disjunkten Vereinigung über $l \in \{1, \dots, j\}$ der Ereignisse ist, dass der Pfad zum Zeitpunkt $2l$ der erste Mal in Null ist und dann zum Zeitpunkt $2j$ ebenfalls. Also ist die erste Summe gleich der Wahrscheinlichkeit des

Ereignisses $\{S_{2j} = 0\}$, d. h. gleich $u_{2j}$. Analog ist die zweite Summe gleich $u_{2(n-j)}$. Damit ist der Beweis beendet. □

Aus Lemma 4.3.7 und Korollar 4.3.2 erhält man mit ein wenig Rechnung, dass

$$\frac{\mathbb{P}_{2n}(X_{2n} = 2j)}{\mathbb{P}_{2n}(X_{2n} = 2(j+1))} = 1 + \frac{n-1-2j}{(n-j)(2j+1)},$$

und dieser Quotient ist positiv für $j \in [0, \frac{1}{2}(n-1)]$ und negativ für $j \in [\frac{1}{2}(n-1), n]$. Die Wahrscheinlichkeiten $\mathbb{P}_{2n}(X_{2n} = 2j)$ fallen also in der linken Hälfte des Definitionsbereiches $\{0, \dots, n\}$ und steigen in der rechten. Sie sind also für $j = 0$ und $j = n$ am größten. Wenn also zwei gleich starke Tennisspieler eine Serie von Matches gegeneinander spielen, so ist es viel wahrscheinlicher, dass einer von ihnen die gesamte Zeit über führt, als dass die Dauer der Führung sich ausgleicht. Dies sieht man auch an den Asymptoten

$$\mathbb{P}_{2n}(X_{2n} = 0) = \mathbb{P}_{2n}(X_{2n} = 2n) \sim \frac{1}{\sqrt{\pi n}} \qquad \text{und} \qquad \mathbb{P}_{2n}(X_{2n} = 2\lfloor n/2 \rfloor) \sim \frac{2}{\pi n},$$

das heißt, die Wahrscheinlichkeit für einen Ausgleich der Führungsdauern geht sogar doppelt so schnell gegen Null wie die Wahrscheinlichkeit für ständige Führung eines der Spieler.

Mit Hilfe der Asymptotik in Korollar 4.3.2 erhalten wir sogar einen sehr schönen Grenzwertsatz:

**Satz 4.3.8 (Arcussinus-Gesetz).** *Für alle $0 < a < b < 1$ gilt*

$$\lim_{n\to\infty} \mathbb{P}_{2n}\Big(a \le \frac{X_{2n}}{2n} \le b\Big) = \frac{1}{\pi} \int_a^b \frac{1}{\sqrt{x(1-x)}}\, \mathrm{d}x = \frac{2}{\pi}\big(\arcsin\sqrt{b} - \arcsin\sqrt{a}\big).$$

**Beweisskizze.** Es gilt für $n \to \infty$:

$$\begin{aligned}\mathbb{P}_{2n}\Big(a \le \frac{X_{2n}}{2n} \le b\Big) &\sim \sum_{j\approx an}^{\approx bn} \mathbb{P}_{2n}(X_{2n} = 2j) = \sum_{j\approx an}^{\approx bn} u_{2j} u_{2(n-j)} \sim \sum_{j\approx an}^{\approx bn} \frac{1}{\sqrt{\pi j \pi (n-j)}} \\ &= \frac{1}{\pi}\frac{1}{n} \sum_{j\approx an}^{\approx bn} \frac{1}{\sqrt{\frac{j}{n}(1-\frac{j}{n})}},\end{aligned}$$

wobei wir Randeffekte bei $j \approx an, bn$ vernachlässigten. Der letzte Ausdruck ist offensichtlich eine Riemannsumme für das Integral $\frac{1}{\pi}\int_a^b (x(1-x))^{-1/2}\,\mathrm{d}x$ für äquidistante Unterteilung in Intervalle der Länge $\frac{1}{n}$. Also konvergiert dieser Ausdruck gegen das Integral. □

Man nennt eine Zufallsgröße $X$ *Arcussinus-verteilt*, wenn für alle $0 < a < b < 1$ gilt:

$$\mathbb{P}(a \le X \le b) = \frac{1}{\pi} \int_a^b \frac{1}{\sqrt{x(1-x)}}\, \mathrm{d}x.$$

Man formuliert das Arcussinus-Gesetz auch, indem man sagt, die (Verteilung der) Zufallsgröße $X_{2n}/2n$ *konvergiert schwach* gegen (die Verteilung) eine(r) Arcussinus-verteilte(n) Zufallsgröße. Die schwache Konvergenz wird in Abschnitt 6.2 ausführlicher behandelt.

# Kapitel 5

# Wahrscheinlichkeit mit Dichten

In diesem Kapitel erweitern bzw. übertragen wir die bisher behandelte Theorie auf Wahrscheinlichkeitsmaße bzw. Zufallsgrößen, die mit Hilfe von Integralen über Dichten beschrieben werden, wie z. B. die Gleichverteilung auf einem beschränkten Intervall, die Normal- und die Exponentialverteilungen. Alle in diesem Kapitel auftretenden Integrale werden als Riemann-Integrale aufgefasst, sodass Vorkenntnisse in Analysis I ausreichen und z. B. eine Kenntnis des Lebesgue-Integrals nicht benötigt wird. Wir werden auch darauf verzichten, Wahrscheinlichkeitsräume anzugeben, denn dies wird nicht nötig sein für eine Behandlung der Theorie auf dem Niveau, auf dem wir sie betreiben. Mit anderen Worten, wir verzichten in diesem Kapitel auf eine mathematische Fundierung. Eine solche ist Gegenstand der Vorlesung *Stochastik I*, in der die nötigen maßtheoretischen Konzepte eingeführt werden.

Der Verzicht auf Maßtheorie und Lebesgue-Integral im vorliegenden Skript hat den Vorteil, dass auch Studenten ab dem dritten Semester folgen können, denn wir werden schnell zu interessanten Beispielen kommen, und der theoretische Aufwand wird gering bleiben. Der Nachteil ist allerdings die Unmöglichkeit, die diskrete Wahrscheinlichkeitstheorie und die Theorie der Wahrscheinlichkeit mit Dichten unter das Dach eines übergeordneten Konzepts zu stellen; sie werden scheinbar verbindungslos neben einander stehen bleiben, mit allerdings nicht zu übersehenden Verwandtschaften. Tatsächlich werden wir nicht einmal im Stande sein, eine diskrete Zufallsgröße und eine über eine Dichte definierte Zufallsgröße mit einander zu addieren. Dieser Defekt gibt sicherlich eine weitere Motivation, die Vorlesung *Stochastik I* zu besuchen.

## 5.1 Grundbegriffe

**Definition 5.1.1 (Dichte, Verteilungsfunktion).** *(a) Eine Abbildung* $f\colon \mathbb{R} \to [0,\infty)$*, so dass* $\int_{\mathbb{R}} f(x)\,\mathrm{d}x$ *existiert und den Wert Eins hat, heißt eine* Wahrscheinlichkeitsdichte *oder kurz eine* Dichte*.*

*(b) Eine Abbildung* $F\colon \mathbb{R} \to [0,1]$ *heißt eine* Verteilungsfunktion*, falls gelten:*

*(i)* $F$ *ist monoton steigend,*

*(ii)* $\lim_{t\to\infty} F(t) = 1$ *und* $\lim_{t\to-\infty} F(t) = 0$*,*

*(iii)* $F$ *ist rechtsseitig stetig (d. h.* $\lim_{s\downarrow t} F(s) = F(t)$ *für jedes* $t \in \mathbb{R}$*).*

**Bemerkung 5.1.2.** (a) Falls $f$ eine Dichte ist, so definiert $F(t) = \int_{-\infty}^{t} f(x)\,\mathrm{d}x$ eine Verteilungsfunktion, und zwar sogar eine stetige. Man nennt dann $f$ eine Dichte von $F$. Nicht jede stetige Verteilungsfunktion besitzt eine Dichte.

(b) Falls eine Dichte $f$ in endlich vielen Punkten abgeändert wird, erhält man eine neue Dichte $\tilde{f}$. Für jedes Intervall $I$ gilt dann $\int_I f(x)\,\mathrm{d}x = \int_I \tilde{f}(x)\,\mathrm{d}x$.

(c) Falls eine Dichte $f$ einer Verteilungsfunktion $F$ stetig in einem Punkte $a$ ist, so gilt $F'(a) = f(a)$ nach dem Hauptsatz der Differential- und Integralrechnung.

◇

**Definition 5.1.3 (Verteilungsfunktion und Dichte einer Zufallsgröße).** *Für eine reellwertige Zufallsgröße $X$ heißt die Abbildung $F_X\colon \mathbb{R} \to [0,1]$, definiert durch $F_X(t) = \mathbb{P}(X \le t)$, die* Verteilungsfunktion von $X$. *Wir sagen, $X$* hat eine Dichte, *wenn ihre Verteilungsfunktion $F_X$ eine hat.*

**Bemerkung 5.1.4.** (a) Falls $X$ eine diskrete Zufallsgröße ist, so ist $F_X$ die rechtsstetige Treppenfunktion, die in den Punkten $x$ mit $\mathbb{P}(X = x) > 0$ einen Sprung der Größe $\mathbb{P}(X = x)$ macht. Insbesondere hat $X$ keine Dichte.

(b) Wenn eine Zufallsgröße $X$ eine Dichte $f$ hat, dann gilt

$$\mathbb{P}\big(X \in A\big) = \mathbb{P} \circ X^{-1}(A) = \int_A f(x)\,\mathrm{d}x, \tag{5.1.1}$$

für alle Mengen $A \subset \mathbb{R}$, für die die Abbildung $f\mathbb{1}_A$ Riemann-integrierbar ist, also mindestens für alle endlichen Vereinigungen $A$ von Intervallen. Insbesondere ist $\mathbb{P}(X = x) = 0$ für alle $x \in \mathbb{R}$, denn

$$0 \le \mathbb{P}(X = x) \le \mathbb{P}\Big(x \le X \le x + \frac{1}{n}\Big) = \int_x^{x+\frac{1}{n}} f(y)\,\mathrm{d}y \longrightarrow 0 \qquad \text{für } n \to \infty.$$

Da $X$ einzelne Werte mit Wahrscheinlichkeit Null annimmt, gilt insbesondere $\mathbb{P}(X \in [a,b]) = \mathbb{P}(X \in (a,b]) = \mathbb{P}(X \in (a,b))$ etc. Wir werden allerdings keinen Wahrscheinlichkeitsraum angeben, auf dem $X$ definiert wäre.

(c) Wir sagen, ein Wahrscheinlichkeitsmaß $\mathbb{P}$ auf $\Omega = \mathbb{R}$ besitzt eine Dichte $f$, wenn die identische Zufallsgröße $X(\omega) = \omega$ eine hat. In diesem Fall gilt also $\mathbb{P}(A) = \int_A f(x)\,\mathrm{d}x$ für alle Mengen $A \subset \mathbb{R}$, für die $f\mathbb{1}_A$ Riemann-integrierbar ist, mindestens aber für alle endlichen Vereinigungen von Intervallen.

◇

## 5.2 Übertragung der bisherigen Ergebnisse

Die meisten allgemeinen Aussagen der voran gegangenen Kapitel über Wahrscheinlichkeiten, Erwartungswerte, Varianzen, Kovarianzen und Unabhängigkeit gelten analog auch für Wahrscheinlichkeitsmaße bzw. Zufallsgrößen mit Riemann-integrierbaren Dichten, und die Beweise laufen analog. In den meisten Fällen genügt es, die auftretenden Summen über $x \in X(\Omega)$ durch

die entsprechenden Integrale $\int_{\mathbb{R}} \dots \mathrm{d}x$ und den Term $\mathbb{P}(X = x)$ durch die Dichte $f(x)$ zu ersetzen und die Sprechweise anzupassen. Ausdrücke, die explizit den Wahrscheinlichkeitsraum $\Omega$ involvieren, müssen wir außen vor lassen.

Da das Riemann-Integral allerdings (anders als das Lebesgue-Integral) keine abzählbaren Additivitätseigenschaften aufweist[1], können also diejenigen Eigenschaften, die auf der zweiten Aussage in den Kolmogorovschen Axiomen in Bemerkung 1.1.3 beruhen, nicht gefolgert werden und müssen durch die entsprechende *endliche* Additivität ersetzt werden.

Im Einzelnen gilt Folgendes.

### Grundbegriffe

Die Rechenregeln für Wahrscheinlichkeiten in Lemma 1.1.4(a)-(c) gelten wörtlich auch für Wahrscheinlichkeitsmaße $\mathbb{P}$ mit Dichten, aber die Aussagen in (d) und (e) müssen auf endliche Familien von Ereignissen eingeschränkt werden. Die bedingte Wahrscheinlichkeit wird wie in Definition 2.1.2 definiert, und ihre Eigenschaften in Lemma 2.1.4 (allerdings nur für *endliche* Familien von Ereignissen) und die Multiplikationsformel in Lemma 2.1.6 gelten ebenso für Wahrscheinlichkeiten mit Dichten.

### Gemeinsame Verteilungen und Randdichten

**Definition 5.2.1 (gemeinsame Verteilungsfunktion und Dichte).** *Die* gemeinsame Verteilungsfunktion *von Zufallsgrößen $X_1, \dots, X_n$, also die Verteilungsfunktion des Zufallsvektors $X = (X_1, \dots, X_n)$, ist definiert durch*

$$\begin{aligned} F_X(t_1, \dots, t_n) &= \mathbb{P}\big(X_1 \le t_1, \dots, X_n \le t_n\big) \\ &= \mathbb{P} \circ X^{-1}\big((-\infty, t_1] \times \dots \times (-\infty, t_n]\big), \qquad t_1, \dots, t_n \in \mathbb{R}. \end{aligned}$$

*Wir sagen, $X_1, \dots, X_n$ haben eine* gemeinsame Dichte $f\colon \mathbb{R}^n \to [0, \infty)$ *(oder der Zufallsvektor $X = (X_1, \dots, X_n)$ habe eine Dichte $f$), falls gilt*

$$P(X \in A) = P \circ X^{-1}(A) = \int_A f(x_1, \dots, x_n)\, \mathrm{d}x_1 \dots \mathrm{d}x_n$$

*für alle $A \subset \mathbb{R}^n$, für die die Abbildung $f \mathbb{1}_A$ Riemann-integrierbar ist.*

**Bemerkung 5.2.2.** (a) Wenn $X_1, \dots, X_n$ die Verteilungsfunktion $F_X$ hat, so hat jedes $X_i$ die Verteilungsfunktion

$$\begin{aligned} F_{X_i}(t_i) &= \mathbb{P}(X_1 < \infty, \dots, X_{i-1} < \infty, X_i \le t_i, X_{i+1} < \infty, \dots, X_n < \infty) \\ &= \lim_{t_1, \dots, t_{i-1}, t_{i+1}, \dots, t_n \to \infty} F_X(t_1, \dots, t_n). \end{aligned}$$

[1] Zum Beispiel ist für jedes $x \in \mathbb{Q} \cap [0, 1]$ die Abbildung $\mathbb{1}_{\{x\}}$ Riemann-integrierbar, nicht aber ihre abzählbare Summe $\mathbb{1}_{\mathbb{Q} \cap [0,1]}$.)

(b) Wenn $X_1, \dots, X_n$ eine gemeinsame Dichte $f$ haben, dann gilt insbesondere

$$F_X(t_1 \dots, t_n) = \mathbb{P} \circ X^{-1}\Big(\prod_{i=1}^{n}(-\infty, t_i]\Big) = \int_{\prod_{i=1}^{n}(-\infty,t_i]} f(x_1, \dots, x_n)\,\mathrm{d}x_1 \dots \mathrm{d}x_n$$
$$= \int_{-\infty}^{t_n} \cdots \int_{-\infty}^{t_1} f(x_1, \dots, x_n)\,\mathrm{d}x_1 \dots \mathrm{d}x_n, \qquad t_1, \dots, t_n \in \mathbb{R}.$$

(Der *Satz von Fubini* garantiert, dass der Wert dieses $n$-fachen Integrals nicht von der Reihenfolge der Integration abhängt.) Insbesondere besitzen auch die einzelnen Zufallsgrößen $X_1, \dots, X_n$ jeweils eine Dichte, und zwar erhält man eine Dichte von $X_i$, indem man $f$ über alle Werte, die die anderen Zufallsgrößen annehmen können, integriert:

$$\mathbb{P}(X_i \le t_i) = \int_{\mathbb{R}^{i-1}\times(-\infty,t_i]\times\mathbb{R}^{n-i}} f(x_1, \dots, x_n)\,\mathrm{d}x_1 \dots \mathrm{d}x_n$$
$$= \int_{(-\infty,t_i]} \Big( \int_{\mathbb{R}^{n-1}} f(x_1, \dots, x_n)\,\mathrm{d}x_1 \dots \mathrm{d}x_{i-1}\mathrm{d}x_{i+1} \dots \mathrm{d}x_n \Big)\mathrm{d}x_i.$$

Der Ausdruck in den Klammern als Funktion von $x_i$ nennt man die $i$-te *Randdichte* von $f$; diese Funktion ist eine (eindimensionale) Dichte von $X_i$. In analoger Weise kann man für jeden Teilvektor von $X = (X_1, \dots, X_n)$ eine Dichte erhalten, indem man über alle Indices, die nicht in dem Vektor verwendet werden, ausintegriert.

◇

### Unabhängigkeit

Für die Definition der Unabhängigkeit von Zufallsgrößen adaptieren wir die Aussage von Lemma 3.2.2:

**Definition 5.2.3 (Unabhängigkeit von Zufallsgrößen).** *Es seien $X_1, \dots, X_n$ beliebige Zufallsgrößen. Wir nennen $X_1, \dots, X_n$* unabhängig, *falls für alle $t_1, \dots, t_n \in \mathbb{R}$ gilt:*

$$\mathbb{P}(X_1 \le t_1, \dots, X_n \le t_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \le t_i).$$

Wie in Lemma 3.2.9 sind die Zufallsgrößen $X_1, \dots, X_n$ mit Dichten genau dann unabhängig, wenn die Verteilung des Vektors $X = (X_1, \dots, X_n)$ gleich dem Produkt der Verteilungen der $X_1, \dots, X_n$ ist. Unabhängigkeit schlägt sich auch in der Produktstruktur der Dichten nieder:

**Lemma 5.2.4.** *Es seien Zufallsgrößen $X_1, \dots, X_n$ mit Dichten $f_1, \dots, f_n \colon \mathbb{R} \to [0, \infty)$ gegeben (wir setzen nicht voraus, dass eine gemeinsame Dichte existiert). Dann sind $X_1, \dots, X_n$ genau dann unabhängig, wenn eine gemeinsame Dichte gegeben ist durch die Abbildung*

$$(x_1, \dots, x_n) \mapsto \prod_{i=1}^{n} f_i(x_i), \qquad x_1, \dots, x_n \in \mathbb{R}.$$

**Beweis.** Übungsaufgabe. □

### Erwartungswerte, Varianzen und Kovarianzen

Erwartungswerte, Varianzen und Kovarianzen von Zufallsgrößen mit Dichten werden analog zu den jeweils entsprechenden Begriffen für diskrete Zufallsgrößen definiert:

**Definition 5.2.5 (Erwartungswert,Varianz).** *Es sei $X$ eine Zufallsgröße mit Dichte $f$.*

*(a) Der* Erwartungswert von $X$ *existiert genau dann (und wir schreiben dann $X \in \mathcal{L}^1$), wenn das Integral $\int_{\mathbb{R}} |x| f(x)\,\mathrm{d}x$ konvergiert, und der Erwartungswert ist dann gegeben als*

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

*(b) Wenn $X \in \mathcal{L}^1$, so heißt*

$$\mathbb{V}(X) = \int_{\mathbb{R}} (x - \mathbb{E}(X))^2 f(x)\,\mathrm{d}x = \mathbb{E}\big((X - \mathbb{E}(X))^2\big)$$

*die* Varianz von $X$ *und $S(X) = \sqrt{\mathbb{V}(X)}$ die* Standardabweichung von $X$.

Die Eigenschaften des Erwartungswertes in Lemma 3.3.2 und die Formel für Erwartungswerte von zusammengesetzen Zufallsgrößen in Lemma 3.3.3 gelten wörtlich bzw. analog; in Lemma 3.3.2(a) und Lemma 3.3.3 müssen die Summen durch Integrale ersetzt werden. Die Eigenschaften von Varianzen und Kovarianzen in Lemmas 3.4.4 und 3.5.2 sowie die Cauchy-Schwarz-Ungleichung in Satz 3.5.6 gelten wörtlich auch für Zufallsvariable mit Dichten.

### Faltung

Den Begriff der *Faltung* in Definition 4.1.1 überträgt man wie folgt auf Integrale: Die Faltung zweier integrierbarer Funktionen $f, g\colon \mathbb{R} \to \mathbb{R}$ ist definiert als die Funktion $f \star g\colon \mathbb{R} \to \mathbb{R}$, gegeben durch

$$f \star g(y) = \int_{\mathbb{R}} f(x) g(y - x)\,\mathrm{d}x, \qquad y \in \mathbb{R}.$$

Man kann leicht zeigen, dass $f \star g = g \star f$, und dass $f \star g$ absolut integrierbar ist, falls $f$ und $g$ es sind. Es gilt das Analogon zum Faltungssatz 4.1.2:

**Satz 5.2.6 (Faltungssatz).** *Für je zwei unabhängige Zufallsgrößen $X$ und $Y$ mit Dichten $f$ bzw. $g$ hat die Zufallsgröße $X + Y$ die Dichte $f \star g$.*

**Beweis.** Nach Lemma 5.2.4 hat $(X, Y)$ die Dichte $(x, y) \mapsto f(x)g(y)$. Es sei $z \in \mathbb{R}$ und $A_z =$

$\{(x,y) \in \mathbb{R}^2 \colon x + y \le z\}$. Dann ist

$$\begin{aligned}\mathbb{P}(X+Y \le z) &= \mathbb{P}\big((X,Y) \in A_z\big) = \int_{A_z} f(x)g(y)\,\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{\infty} \mathrm{d}x\, f(x) \int_{-\infty}^{z-x} \mathrm{d}y\, g(y) \\ &= \int_{-\infty}^{\infty} \mathrm{d}x\, f(x) \int_{-\infty}^{z} \mathrm{d}y\, g(y-x) = \int_{-\infty}^{z} \mathrm{d}y \left( \int_{\mathbb{R}} \mathrm{d}x\, g(y-x) f(x) \right) \\ &= \int_{-\infty}^{z} \mathrm{d}y\, g \star f(y).\end{aligned}$$

Also ist $g \star f = f \star g$ eine Dichte von $X + Y$. □

## 5.3 Beispiele

Es folgen die wichtigsten Beispiele. Wir benutzen die *Indikatorfunktion* $\mathbb{1}_A \colon \mathbb{R} \to [0,1]$ auf einer Menge $A$, die gegeben ist durch

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

**Beispiel 5.3.1 (Gleichförmige Verteilung).** Seien $a, b \in \mathbb{R}$ mit $a < b$, dann ist durch

$$f(t) = \frac{1}{b-a}\, \mathbb{1}_{[a,b]}(t), \qquad \text{für } t \in \mathbb{R},$$

eine Dichte auf $\mathbb{R}$ gegeben, die Dichte der *gleichförmigen Verteilung auf* $[a,b]$. Die zugehörige Verteilungsfunktion $F$ hat ein lineares Stück vom Wert Null bei $a$ zum Wert Eins bei $b$. Eine Zufallsgröße $X$ mit Dichte $f$ besitzt den Erwartungswert

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f(x)\,\mathrm{d}x = \frac{1}{b-a} \int_a^b x\,\mathrm{d}x = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$$

und die Varianz

$$\begin{aligned}\mathbb{V}(X) &= \int_{\mathbb{R}} (x - \mathbb{E}(X))^2 f(x)\,\mathrm{d}x = \frac{1}{b-a} \int_a^b (x - (a+b)/2)^2\,\mathrm{d}x \\ &= \frac{1}{3(b-a)} \Big( \big(\frac{b-a}{2}\big)^3 - \big(\frac{a-b}{2}\big)^3 \Big) = \frac{1}{12}(b-a)^2.\end{aligned}$$

Analog werden gleichförmige Verteilungen auf beliebigen Teilmengen des $\mathbb{R}^d$ definiert, deren Indikatorfunktion Riemann-integrierbar ist. ◇

**Beispiel 5.3.2 (Exponentialverteilung).** Wir definieren mit einem Parameter $\alpha \in (0,\infty)$ die Dichte

$$f(t) = \alpha e^{-\alpha t} \mathbb{1}_{[0,\infty)}(t), \qquad \text{für } t \in \mathbb{R},$$

die die Dichte der *Exponentialverteilung* zum Parameter $\alpha$ genannt wird. Die zugehörige Verteilungsfunktion ist gegeben durch $F(t) = (1 - e^{-\alpha t}) \mathbb{1}_{[0,\infty)}(t)$, und den Erwartungswert errechnet man mit Hilfe einer partiellen Integration als

$$\mathbb{E}(X) = \int_{\mathbb{R}} t f(t)\,\mathrm{d}t = \int_0^\infty t\alpha e^{-\alpha t}\,\mathrm{d}t = -te^{-\alpha t}\Big|_0^\infty + \int_0^\infty e^{-\alpha t}\,\mathrm{d}t = \frac{1}{\alpha}.$$

Mit einer weiteren partiellen Integration errechnet man leicht, dass die Varianz gegeben ist als $\mathbb{V}(X) = 1/\alpha^2$ (Übungsaufgabe).

Die Exponentialverteilung ist das kontinuierliche Gegenstück zur geometrischen Verteilung. Insbesondere besitzt sie ebenfalls die Eigenschaft der Gedächtnislosigkeit (siehe Lemma 3.1.6):

**Lemma 5.3.3 (Gedächtnislosigkeit der Exponentialverteilung).** *Sei $X$ eine exponentiell verteilte Zufallsvariable. Dann gilt für jede $s, t > 0$*

$$\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t).$$

**Beweis.** Übungsaufgabe. □

Die Interpretation ist analog zu der von Lemma 3.1.6: Wenn man auf das Eintreten einer exponentiell verteilten Zufallszeit wartet und sie bis zum Zeitpunkt $s$ noch nicht eingetreten ist, so ist die Wahrscheinlichkeit, dass sie nach weiteren $t$ Zeiteinheiten eintritt, die gleiche, als wenn man das Nichteintreten in den letzten $s$ Zeiteinheiten nicht kennen würde. Weitere interessante Eigenschaften dieser Wartezeitverteilung treten bei der Behandlung des Poisson-Prozesses in Abschnitt 5.4 auf.

Das Maximum unabhängiger gleichförmig verteiler Zufallsgrößen steht in einer interessanten Beziehung zur Exponentialverteilung:

**Lemma 5.3.4.** *Es seien $X_1, \dots, X_n$ unabhängige, auf dem Intervall $[0, \alpha]$ gleichförmig verteilte Zufallsgrößen. Dann ist eine Dichte der Zufallsgröße $M_n = \max\{X_1, \dots, X_n\}$ gegeben durch $x \mapsto \mathbb{1}_{[0,\alpha]}(x) n\alpha^{-n}x^{n-1}$, und ihr Erwartungswert ist $\mathbb{E}(M_n) = \frac{n}{n+1}\alpha$. Für $n \to \infty$ konvergiert die Verteilungsfunktion der Zufallsgröße $Y_n = n(\alpha - M_n)$ gegen die Verteilungsfunktion der Exponentialverteilung mit Parameter $\frac{1}{\alpha}$.*

**Beweis.** Übungsaufgabe. □

◇

**Beispiel 5.3.5 (Gamma-Verteilung).** Mit zwei Parametern $\alpha > 0$ und $r > 0$ definieren wir die Dichte $\gamma_{\alpha,r} \colon (0, \infty) \to [0, \infty)$ durch

$$\gamma_{\alpha,r}(t) = \frac{\alpha^r}{\Gamma(r)} t^{r-1} e^{-\alpha t} \mathbb{1}_{[0,\infty)}(t),$$

wobei $\Gamma \colon (0, \infty) \to (0, \infty)$ die bekannte Gamma-Funktion[2] ist:

$$\Gamma(r) = \int_0^\infty y^{r-1} e^{-y} \, \mathrm{d}y, \qquad r > 0.$$

Mit Hilfe der Substitution $\alpha t = y$ sieht man leicht, dass $\gamma_{\alpha,r}$ eine Dichte ist, die Dichte der *Gamma-Verteilung* mit Parametern $\alpha$ und $r$. Der Spezialfall $r = 1$ ist die Dichte der Exponentialverteilung, siehe Beispiel 5.3.2. Die Gamma-Verteilung ist das kontinuierliche Analogon zur

[2] Die Gamma-Funktion ist die einzige logarithmisch konvexe Funktion $f \colon (0, \infty) \to (0, \infty)$ mit $f(1) = 1$, die die Funktionalgleichung $\Gamma(r + 1) = r\Gamma(r)$ für jedes $r > 0$ erfüllt. Sie interpoliert also auf $\mathbb{N}$ die Fakultät, d. h. $f(k) = (k - 1)!$ für $k \in \mathbb{N}$.

negativen Binomial-Verteilung (siehe Beispiel 4.1.4). Sie besitzt ein paar bemerkenswerte Beziehungen zur Exponential- und zur Poisson-Verteilung. Insbesondere stellt sich heraus, dass die Faltungsgleichung $\gamma_{\alpha,r_1} \star \gamma_{\alpha,r_2} = \gamma_{\alpha,r_1+r_2}$ gilt, d. h., dass die Familie der Gamma-Verteilungen eine Faltungshalbgruppe bildet (siehe auch Abschnitt 4.1):

**Lemma 5.3.6 (Gamma-, Exponential- und Poisson-Verteilung).** *Es sei $\alpha > 0$.*

*(i) Die Summe zweier unabhängiger Gamma-verteilter Zufallsgrößen mit Parametern $\alpha$ und $r_1$ bzw. $\alpha$ und $r_2$ ist Gamma-verteilt mit Parametern $\alpha$ und $r_1 + r_2$. Insbesondere ist für $k \in \mathbb{N}$ die Gamma-Verteilung mit Parameter $\alpha$ und $k$ identisch mit der Verteilung der Summe von $k$ unabhängigen, zum Parameter $\alpha$ exponentiell verteilten Zufallsgrößen.*

*(ii) Für ein $t > 0$ sei $N_{\alpha t}$ eine zum Parameter $\alpha t$ Poisson-verteilte Zufallsgröße und $X_{\alpha,k}$ eine zum Parameter $\alpha$ und $k \in \mathbb{N}$ Gamma-verteilte Zufallsgröße. Dann gilt $\mathbb{P}(N_{\alpha t} \geq k) = \mathbb{P}(X_{\alpha,k} \leq t)$.*

**Beweis.** (i) Es genügt, die Gleichung $\gamma_{\alpha,r_1} \star \gamma_{\alpha,r_2} = \gamma_{\alpha,r_1+r_2}$ zu beweisen: Für $s > 0$ gilt

$$\begin{aligned}\gamma_{\alpha,r_1} \star \gamma_{\alpha,r_2}(s) &= \int_0^s \gamma_{\alpha,r_1}(t)\gamma_{\alpha,r_2}(s-t)\,\mathrm{d}t \\ &= \frac{\alpha^{r_1}}{\Gamma(r_1)}\frac{\alpha^{r_2}}{\Gamma(r_2)} e^{-\alpha s}\int_0^s t^{r_1-1}(s-t)^{r_2-1}\,\mathrm{d}s \\ &= \gamma_{\alpha,r_1+r_2}(s)\frac{\Gamma(r_1+r_2)}{\Gamma(r_1)\Gamma(r_2)}\int_0^1 u^{r_1-1}(1-u)^{r_2-1}\,\mathrm{d}u,\end{aligned}$$

wobei wir die Substitution $t/s = u$ benutzten. Das Integral auf der rechten Seite ist bekannt als das Eulersche Beta-Integral mit Parametern $r_1$ und $r_2$, und es ist bekannt, dass sein Wert gleich dem Kehrwert des Bruches davor ist. Also ist die gesamte rechte Seite gleich $\gamma_{\alpha,r_1+r_2}(s)$, und die Aussage ist bewiesen.

(ii) Wir errechnen

$$\begin{aligned}\mathbb{P}(N_{\alpha t} \geq k) &= 1 - \mathbb{P}(N_{\alpha t} \leq k-1) = 1 - e^{-\alpha t}\sum_{n=0}^{k-1}\frac{(\alpha t)^n}{n!} \\ &= \int_0^t \frac{\alpha^k}{(k-1)!}x^{k-1}e^{-\alpha x}\,\mathrm{d}x = \int_0^t \gamma_{\alpha,k}(x)\,\mathrm{d}x = \mathbb{P}(X_{\alpha,k} \leq t),\end{aligned}$$

wobei der dritte Schritt mit Hilfe einer Differenziation nach $t$ eingesehen wird. □ ◇

**Beispiel 5.3.7 (Normal- oder Gaußverteilung).** Mit zwei Parametern $\mu \in \mathbb{R}$ und $\sigma \in (0,\infty)$ definieren wir $\varphi_{\mu,\sigma}\colon \mathbb{R} \to (0,\infty)$ durch

$$\varphi_{\mu,\sigma}(t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{(t-\mu)^2}{2\sigma^2}\Big), \qquad \text{für } t \in \mathbb{R}.$$

Wir benutzen nun einen kleinen Trick, um zu zeigen, dass $\varphi_{\mu,\sigma}$ tatsächlich eine Wahrscheinlichkeitsdichte ist, d. h. dass $\int_{\mathbb{R}} \varphi_{\mu,\sigma}(t)\,\mathrm{d}t = 1$. Zunächst bemerken wir, dass es genügt, dies nur für $\mu = 0$ und $\sigma = 1$ zu tun, denn das Integral erstreckt sich über die gesamte reelle Achse und kann um $\mu$ verschoben werden, und eine Substitution $t = \tilde{t}\sigma$ führt die Frage auf den Fall $\sigma = 1$ zurück. Wir werden zeigen, dass

$$\Big(\int_{\mathbb{R}} e^{-t^2/2}\,\mathrm{d}t\Big)^2 = 2\pi,$$

woraus die Behauptung folgt. Wir schreiben das Quadrat der Integrale als ein zweidimensionales Integral über $x = (x_1, x_2) \in \mathbb{R}^2$ (man beachte, dass $x_1^2 + x_2^2 = \|x\|_2^2$) und gehen über zu Polarkoordinaten:

$$\Big(\int_{\mathbb{R}} e^{-t^2/2}\,\mathrm{d}t\Big)^2 = \int_{\mathbb{R}} e^{-x_1^2/2}\,\mathrm{d}x_1 \int_{\mathbb{R}} e^{-x_2^2/2}\,\mathrm{d}x_2 = \int_{\mathbb{R}^2} e^{-\|x\|_2^2/2}\,\mathrm{d}x = \int_0^{2\pi}\int_0^{\infty} re^{-r^2/2}\,\mathrm{d}r\mathrm{d}t.$$

Der Lohn dieses Ansatzes ist, dass wir eine explizite Stammfunktion für den Integranden $r \mapsto re^{-r^2/2}$ haben, und zwar $r \mapsto -e^{-r^2/2}$. Daher ist das innere Integral offensichtlich gleich Eins, und der gesamte Ausdruck gleich $2\pi$, wie behauptet. Die Funktion $\varphi_{\mu,\sigma}$ ist also tatsächlich eine Wahrscheinlichkeitsdichte, und zwar die Dichte der *Normal-* oder *Gaußverteilung*. Auf den 10-DM-Scheinen, die bis Ende 2001 im Umlauf waren, war der Graf von $\varphi_{\mu,\sigma}$ abgebildet, man nennt ihn die *Gaußsche Glockenkurve*.

Für die zugehörige Verteilungsfunktion

$$\Phi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\Big(-\frac{(t-\mu)^2}{2\sigma^2}\Big)\,\mathrm{d}t$$

gibt es keinen geschlossenen Ausdruck, aber Tabellen für viele ihrer Werte. Die Rolle der Parameter $\mu$ und $\sigma$ wird klar, wenn man den Erwartungswert und die Varianz ausrechnet. Es sei $X$ eine Zufallsgröße mit Dichte $\varphi_{\mu,\sigma}$, dann gilt:

$$\begin{aligned}\mathbb{E}(X) &= \int_{\mathbb{R}} t\varphi_{\mu,\sigma}(t)\,\mathrm{d}t = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} (t+\mu)\exp\Big(-\frac{t^2}{2\sigma^2}\Big)\,\mathrm{d}t \\ &= \mu + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} t\exp\Big(-\frac{t^2}{2\sigma^2}\Big)\,\mathrm{d}t \\ &= \mu,\end{aligned}$$

denn die Funktion $t \mapsto te^{-t^2/(2\sigma^2)}$ ist antisymmetrisch auf $\mathbb{R}$, und da das Integral existiert (wie man leicht etwa mit Vergleichskriterien sieht), ist sein Wert gleich Null.

Außerdem errechnet man die Varianz mit Hilfe einer Substitution und einer partiellen Integration zu

$$\begin{aligned}\mathbb{V}(X) &= \int_{\mathbb{R}} (t-\mathbb{E}(X))^2\varphi_{\mu,\sigma}(t)\,\mathrm{d}t = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} (t-\mu)^2\exp\Big(-\frac{(t-\mu)^2}{2\sigma^2}\Big)\,\mathrm{d}t \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{\mathbb{R}} s^2 e^{-s^2/2}\,\mathrm{d}s = \frac{\sigma^2}{\sqrt{2\pi}}\Big[-se^{-s^2/2}\Big|_{-\infty}^{\infty} + \int_{\mathbb{R}} e^{-s^2/2}\,\mathrm{d}s\Big] \\ &= \sigma^2.\end{aligned}$$

Also ist $\mu$ der Erwartungswert und $\sigma^2$ die Varianz der Normalverteilung mit Parametern $\mu$ und $\sigma$. Man bezeichnet diese Verteilung auch oft mit $\mathcal{N}(\mu, \sigma^2)$. Im Fall $\mu = 0$ und $\sigma^2 = 1$ sprechen wir von der *Standardnormalverteilung* $\mathcal{N}(0, 1)$.

Die Normalverteilung besitzt mehrere spezielle Eigenschaften und tritt in universeller Weise auf als Grenzverteilung im sehr wichtigen Zentralen Grenzwertsatz, siehe Satz 6.2.2. Wir wollen ihre Faltungseigenschaft beweisen: Die Summe zweier unabhängiger normalverteilter Zufallsgrößen ist wiederum normalverteilt, und ihre ersten Parameter addieren sich und die Quadrate der zweiten ebenfalls:

**Lemma 5.3.8 (Faltungseigenschaft der Normalverteilung).** *Für alle $\mu_1, \mu_2 \in \mathbb{R}$ und alle $\sigma_1, \sigma_2 \in (0, \infty)$ gilt*

$$\varphi_{\mu_1,\sigma_1} \star \varphi_{\mu_2,\sigma_2} = \varphi_{\mu_1+\mu_2,\sigma}, \qquad \text{wobei } \sigma^2 = \sigma_1^2 + \sigma_2^2.$$

**Beweis.** Wir dürfen $\mu_1 = \mu_2 = 0$ annehmen. Sei $t \in \mathbb{R}$. Offensichtlich gilt

$$\frac{\varphi_{0,\sigma_1} \star \varphi_{0,\sigma_2}(t)}{\varphi_{0,\sigma}(t)} = \frac{\sigma}{\sigma_1\sigma_2}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \exp\Big\{\frac{t^2}{2\sigma^2} - \frac{s^2}{2\sigma_1^2} - \frac{(t-s)^2}{2\sigma_2^2}\Big\}\,\mathrm{d}s. \tag{5.3.1}$$

Eine langweilige, aber unkomplizierte Rechnung identifiziert den Term im Exponenten:

$$\frac{t^2}{2\sigma^2} - \frac{s^2}{2\sigma_1^2} - \frac{(t-s)^2}{2\sigma_2^2} = -\frac{1}{2}\Big(s\frac{\sigma}{\sigma_1\sigma_2} - \frac{t}{\sigma}\frac{\sigma_1}{\sigma_2}\Big)^2.$$

Nun benutzt man dies im Integral in (5.3.1) und substituiert den Term zwischen den Klammern im Integral. Also sieht man, dass die rechte Seite gleich Eins ist. □ ◇

**Beispiel 5.3.9 (Mehrdimensionale Normalverteilung).** Es seien $X_1, \dots, X_n$ unabhängige, standardnormalverteilte Zufallsgrößen. Dann hat der Vektor $X = (X_1, \dots, X_n)^{\mathrm{T}}$ die Dichte

$$f(x) = f(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} \exp\Big\{-\frac{1}{2}\sum_{i=1}^n x_i^2\Big\}, \qquad x = (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Wir nennen den Vektor $X$ *$n$-dimensional standardnormalverteilt*. Wir gehen im Folgenden davon aus, dass $X$ ein *Spaltenvektor* ist; mit $X^{\mathrm{T}}$ bezeichnen wir den Zeilenvektor $(X_1, \dots, X_n)$.

Sei nun $A$ eine reguläre $n \times n$-Matrix und $\mu \in \mathbb{R}^n$ ein Vektor, sowie $\theta(x) = Ax + \mu$ für $x \in \mathbb{R}^n$. Wir sagen, dass der Vektor $Y = \theta(X) = (Y_1, \dots, Y_n)^{\mathrm{T}}$ eine *(allgemeine) Normalverteilung* besitzt. Dann besitzt $Y$ die Dichte

$$g(y) = g(y_1, \dots, y_n) = \frac{1}{|\det(C)|^{1/2}(2\pi)^{n/2}} \exp\Big\{-\frac{1}{2}(y-\mu)^{\mathrm{T}}C^{-1}(y-\mu)\Big\}, \qquad y \in \mathbb{R}^n,$$

wobei $C = AA^{\mathrm{T}}$, und $A^{\mathrm{T}}$ ist die Transponierte der Matrix $A$. Dass die Dichte von $Y = AX + \mu$ diese Form haben muss, sieht man ein, indem man die lineare Substitution $y = Ax + \mu$, also $x = A^{-1}(y - \mu)$, durchführt und beachtet, dass gilt:

$$(y-\mu)^{\mathrm{T}}C^{-1}(y-\mu) = (y-\mu)^{\mathrm{T}}(A^{\mathrm{T}})^{-1}A^{-1}(y-\mu) = x^{\mathrm{T}}x = \sum_{i=1}^n x_i^2.$$

Aus der Theorie der mehrdimensionalen Integration ist bekannt, dass die Integrationsvariablen der Regel $\mathrm{d}y = |\det A|\,\mathrm{d}x = \sqrt{|\det C|}\,\mathrm{d}x$ gehorchen.

Wir definieren Erwartungswerte von Zufallsvektoren komponentenweise, also ist $\mathbb{E}(Y) = (\mathbb{E}(Y_1), \dots, \mathbb{E}(Y_n))^{\mathrm{T}}$, und es ergibt sich, dass $\mathbb{E}(Y) = A\mathbb{E}(X) + \mu = \mu$ ist. Ferner ist die *Kovarianzmatrix* $\mathrm{cov}(Y, Y) = (\mathrm{cov}(Y_i, Y_j))_{i,j=1,\dots,n}$ gegeben durch

$$\mathrm{cov}(Y_i, Y_j) = \mathbb{E}\big((Y_i - \mathbb{E}(Y_i))(Y_j - \mathbb{E}(Y_j))\big), \qquad i, j = 1, \dots, n.$$

Sie erfüllt

$$\begin{aligned}\operatorname{cov}(Y,Y) &= \mathbb{E}\big((Y-\mathbb{E}(Y))(Y-\mathbb{E}(Y))^{\mathrm{T}}\big) = \mathbb{E}\big((AX)(AX)^{\mathrm{T}}\big)\\ &= \mathbb{E}(AXX^{\mathrm{T}}A^{\mathrm{T}}) = A\mathbb{E}(XX^{\mathrm{T}})A^{\mathrm{T}} = AA^{\mathrm{T}}\\ &= C.\end{aligned}$$

Also ist (wie im eindimensionalen Fall) die Verteilung des normalverteilten Vektors $Y$ festgelegt durch den Erwartungswert und die Kovarianzmatrix. Man nennt $Y$ die ($n$-dimensionale) Normalverteilung mit Kovarianzmatrix $C$ und Erwartungswertvektor $\mu$ und schreibt auch oft $Y \sim \mathcal{N}(\mu, C)$. ◇

**Beispiel 5.3.10 (Cauchy-Verteilung).** Die Dichte der *Cauchy-Verteilung* mit Parameter $c \in (0,\infty)$ ist gegeben durch

$$f(t) = \frac{c}{\pi}\,\frac{1}{t^2+c^2}, \qquad \text{für } t \in \mathbb{R}.$$

Die Verteilungsfunktion ist gegeben durch $F(t) = \frac{1}{\pi}\arctan(\frac{t}{c})$. Der Erwartungswert der Cauchy-Verteilung existiert nicht, da die Funktion $t \mapsto \frac{|t|}{t^2+1}$ nicht über $\mathbb{R}$ integrierbar ist. ◇

## 5.4 Der Poisson-Prozess

In diesem Abschnitt diskutieren wir ein wichtiges mathematisches Modell für das Eintreten einer Folge von zufälligen Zeitpunkten. Dieses Modell wurde schon in Beispiel 1.3.6 angedeutet, doch für eine befriedigende Behandlung ist eine Kenntnis der Exponential- und der Gamma-Verteilungen notwendig (siehe Beispiele 5.3.2 und 5.3.5). Wir werden zunächst den Poisson-Prozess axiomatisch einführen und den Zusammenhang mit Poisson-verteilten Zufallsvariablen diskutieren. Danach geben wir eine Charakterisierung in Termen von exponentiellen Zufallsvariablen. Wegen unseres Verzichts auf Maßtheorie wird uns es nicht möglich sein, die Existenz des Poisson-Prozesses streng zu beweisen, und wir werden auch aus Zeit- und Platzgründen gewisse Teile der Beweise nur andeuten, aber nicht ausformulieren.

Gegeben seien zufällige Zeitpunkte auf der positiven Zeitachse $(0,\infty)$. Dies können Zeitpunkte von Abfahrten von Bussen von einer Haltestelle sein oder von eingehenden Telefonanrufen oder vieles Andere mehr. Für jedes endliche halboffene Zeitintervall $I = (a,b]$ mit $0 \le a < b < \infty$ (für ein solches Intervall werden wir im Folgenden nur kurz 'Intervall' sagen) sei $N_I$ die Anzahl derjenigen zufälligen Zeitpunkte, die in das Intervall $I$ fallen. Für die Kollektion der $\mathbb{N}_0$-wertigen Zufallsgrößen $N_I$ machen wir folgende Annahmen:

(P1) Die Verteilung von $N_I$ hängt nur von der Länge des Intervalls $I$ ab.

(P2) Wenn $I_1,\dots,I_k$ paarweise disjunkte Intervalle sind, dann sind $N_{I_1},\dots,N_{I_k}$ unabhängig.

(P3) Für jedes Intervall $I$ existiert $\mathbb{E}(N_I)$.

(P4) Es gibt ein Intervall $I$ mit $\mathbb{P}(N_I > 0) > 0$.

(P5) Es gilt $\lim_{\varepsilon \downarrow 0} \varepsilon^{-1}\mathbb{P}(N_{(0,\varepsilon]} \ge 2) = 0$.

Eine Kollektion von Zufallsgrößen $N_I$, die die Eigenschaften (P1)-(P5) erfüllen, nennen wir einen *Poissonschen Punktprozess* oder kurz einen *Poisson-Prozess*. Oft nennt man auch den Prozess $(N_{(0,t]})_{t\in[0,\infty)}$ einen Poisson-Prozess. Die Bedingungen (P1) und (P2) sind starke

Strukturannahmen, die die mathematische Behandlung vereinfachen bzw. ermöglichen sollen. (P3) und (P4) schließen unerwünschte pathologische Fälle aus, und (P5) verhindert, dass sich die Zeitpunkte zu stark häufen können. Wir werden nicht beweisen, dass ein Poisson-Prozess existiert, sondern wir werden dies annehmen und diesen Prozess genauer untersuchen. Es stellt sich heraus, dass alle Zählvariablen $N_I$ Poisson-verteilt sind, eine Tatsache, die den Namen erklärt:

**Lemma 5.4.1.** *Wenn (P1)-(P5) erfüllt sind, so existiert ein $\alpha > 0$, sodass für alle $t, s > 0$ die Zufallsvariable $N_{(t,t+s]}$ Poisson-verteilt ist mit Parameter $\alpha s$. Insbesondere ist $\mathbb{E}(N_I) = \alpha|I|$ für jedes Intervall $I$.*

**Beweis.** Zunächst identifizieren wir den Erwartungswert jedes $N_I$, der ja nach (P3) endlich ist. Die Funktion $\alpha(t) = \mathbb{E}(N_{(0,t]})$ erfüllt $\alpha(0) = 0$ sowie

$$\alpha(t+s) = \mathbb{E}(N_{(0,t]} + N_{(t,t+s]}) = \mathbb{E}(N_{(0,t]}) + \mathbb{E}(N_{(0,s]}) = \alpha(t) + \alpha(s),$$

wobei wir im zweiten Schritt (P1) benutzten. Mit Hilfe von ein wenig Maßtheorie sieht man, dass $\lim_{t\downarrow 0} \alpha(t) = 0$, d. h. $\alpha(\cdot)$ ist stetig in 0. Eine beliebte Übungsaufgabe aus der Analysis zeigt, dass es ein $\alpha \geq 0$ gibt mit $\alpha(t) = \alpha t$ für jedes $t \geq 0$. Wegen (P4) muss $\alpha > 0$ sein.

Nun beweisen wir die erste Aussage des Lemmas. Wegen (P1) reicht es, $N_{(0,s]}$ zu betrachten. Wir zerlegen $(0, s]$ in die Intervalle $I_j^{(k)} = (\frac{s}{k}(j-1), \frac{s}{k}j]$ mit $j \in \{1, \dots, k\}$ und betrachten die Zufallsvariable $X_j^{(k)} = N_{I_j^{(k)}}$. Dann gilt offensichtlich $N_{(0,s]} = \sum_{j=1}^k X_j^{(k)}$. Ferner approximieren wir $X_j^{(k)}$ mit der Hilfsvariable

$$\overline{X}_j^{(k)} = \begin{cases} 1, & \text{falls } X_j^{(k)} > 0, \\ 0 & \text{sonst.} \end{cases}$$

Mit anderen Worten, $\overline{X}_j^{(k)}$ registriert, ob das $j$-te Intervall $I_j^{(k)}$ mindestens einen der zufälligen Punkte erhält oder nicht. Wegen (P2) in Kombination mit Korollar 3.2.12 sind die Variablen $\overline{X}_1^{(k)}, \dots, \overline{X}_k^{(k)}$ unabhängig, und wegen (P1) sind sie identisch verteilt. Mit anderen Worten, sie sind Bernoulli-Variable mit Parameter $p_k = \mathbb{P}(N_{(0,s/k]} > 0)$.

Nun definieren wir $\overline{N}_{(0,s]}^{(k)} = \sum_{j=1}^k \overline{X}_j^{(k)}$, dann ist $\overline{N}_{(0,s]}^{(k)}$ eine binomialverteilte Zufallsvariable mit Parametern $k$ und $p_k$. Es gilt offensichtlich $\overline{N}_{(0,s]}^{(k)} \leq N_{(0,s]}$.

Wir benutzen nun (P5), um zu zeigen, dass für große $k$ die Variablen $\overline{N}_{(0,s]}^{(k)}$ und $N_{(0,s]}$ sehr nahe bei einander liegen, d. h. dass gilt:

$$\lim_{k\to\infty} \mathbb{P}(\overline{N}_{(0,s]}^{(k)} = m) = \mathbb{P}(N_{(0,s]} = m), \qquad m \in \mathbb{N}_0. \tag{5.4.1}$$

Dies sieht man ein, indem man abschätzt:

$$\mathbb{P}(\overline{N}_{(0,s]}^{(k)} \neq N_{(0,s]}) = \mathbb{P}(\overline{N}_{(0,s]}^{(k)} > N_{(0,s]}) \leq \sum_{j=1}^k \mathbb{P}(\overline{X}_j^{(k)} \geq 2) = k\mathbb{P}(N_{(0,s/k]} \geq 2),$$

und dies konvergiert gegen Null für $k \to \infty$ wegen (P5). Da $|\mathbb{P}(\overline{N}_{(0,s]}^{(k)} = m) - \mathbb{P}(N_{(0,s]} = m)| \leq 2\mathbb{P}(\overline{N}_{(0,s]}^{(k)} \neq N_{(0,s]})$, folgt (5.4.1).

Mit Hilfe von (5.4.1) zeigt man nun, dass $\lim_{k\to\infty} kp_k = \alpha s$ ist: Wir haben

$$\lim_{k\to\infty} kp_k = \lim_{k\to\infty} \mathbb{E}(\overline{N}^{(k)}_{(0,s]}) = \lim_{k\to\infty} \sum_{l=1}^{\infty} \mathbb{P}(\overline{N}^{(k)}_{(0,s]} \geq l) = \sum_{l=1}^{\infty} \mathbb{P}(N_{(0,s]} \geq l) = \mathbb{E}(N_{(0,s]}) = \alpha s,$$

wobei wir im zweiten und im vorletzten Schritt Lemma 3.3.8 benutzten und im dritten eine allgemeine Tatsache über Reihen.

Also kann man den Poissonschen Grenzwertsatz (Lemma 1.3.7) anwenden und erhält

$$\lim_{k\to\infty} \mathbb{P}(\overline{N}^{(k)}_{(0,s]} = m) = \lim_{k\to\infty} \mathrm{Bi}_{k,p_k}(m) = \mathrm{Po}_{\alpha s}(m), \qquad m \in \mathbb{N}_0.$$

D. h., die Zufallsvariable $\overline{N}^{(k)}_{(0,s]}$ ist asymptotisch Poisson-verteilt mit Parameter $\alpha s$. Da diese Grenzverteilung mit der Verteilung von $N_{(0,s]}$ übereinstimmt, ist der Beweis beendet. □

Wir nähern uns nun einem anderen Zugang zum Poisson-Prozess. Ausgangspunkt der Überlegung ist folgende Kombination der beiden Beobachtungen von Lemma 5.3.6: Die Wahrscheinlichkeit, dass im Intervall $(0,t]$ mindestens $k$ Punkte liegen (dies ist gleich $\mathbb{P}(N_{(0,\alpha t]} \geq k)$, und $N_{(0,\alpha t]}$ ist nach Lemma 5.4.1 Poisson-verteilt mit Parameter $\alpha t$), ist für jedes $k \in \mathbb{N}_0$ gegeben durch $\mathbb{P}(\sum_{i=1}^k \tau_i \leq t)$, wobei $\tau_1, \tau_2, \ldots$ eine Folge unabhängiger, zum Parameter $\alpha$ exponentiell verteilten Zufallsgrößen ist (denn $\sum_{i=1}^k \tau_i$ ist Gamma-verteilt mit Parametern $\alpha$ und $k$). Dies legt den Schluss nahe, dass die Zufallszeiten $\tau_i$ eine Interpretation als die Wartezeiten zwischen dem Eintreffen des $(i-1)$-ten und $i$-ten zufälligen Zeitpunktes zulassen, und genau das ist der Fall:

**Satz 5.4.2.** *Es sei $(\tau_i)_{i\in\mathbb{N}}$ eine Folge unabhängiger, zum Parameter $\alpha$ exponentiell verteilten Zufallsgrößen. Wir definieren $T_k = \sum_{i=1}^k \tau_i$ für $k \in \mathbb{N}$. Für jedes endliche halboffene Intervall $I$ definieren wir $N_I = |\{k \in \mathbb{N} \colon T_k \in I\}|$ als die Zahl der $T_k$, die in $I$ fallen. Dann erfüllt die Kollektion der Zufallsgrößen $N_I$ die Bedingungen (P1)-(P5).*

**Beweisskizze.** Wir werden nicht die volle Stärke der Aussage beweisen, sondern nur die folgende Aussage: Für jedes $0 < s < t$ sind $N_{(0,s]}$ und $N_{(s,t]}$ unabhängige, zu den Parametern $\alpha s$ bzw. $\alpha(t-s)$ Poisson-verteilte Zufallsgrößen. (Der vollständige Beweis des Satzes ist komplizierter, aber analog.) Ausführlicher gesagt, wir werden für jedes $k, l \in \mathbb{N}_0$ zeigen:

$$\mathbb{P}(N_{(0,s]} = k, N_{(s,t]} = l) = \mathrm{Po}_{\alpha s}(k)\mathrm{Po}_{\alpha(t-s)}(l) = e^{-\alpha t}\alpha^{k+l}\frac{s^k}{k!}\frac{(t-s)^l}{l!}. \tag{5.4.2}$$

Zunächst sieht man, dass $\{N_{(0,s]} = k, N_{(s,t]} = l\} = \{T_k \leq s < T_{k+1}, T_{k+l} \leq t < T_{k+l+1}\}$. Das betrachtete Ereignis lässt sich also mit Hilfe der $\tau_i$ ausdrücken als das Ereignis, dass der Zufallsvektor $(\tau_1, \ldots, \tau_{k+l+1})$ in der Menge

$$A = \big\{x \in [0,\infty)^{k+l+1} \colon S_k(x) \leq s < S_{k+1}(x), S_{k+l}(x) \leq t < S_{k+l+1}(x)\big\}$$

liegt (wobei wir $S_n(x) = x_1 + \cdots + x_n$ setzten):

$$\{N_{(0,s]} = k, N_{(s,t]} = l\} = \big\{(\tau_1, \ldots, \tau_{k+l+1}) \in A\big\}.$$

Nach Lemma 5.2.4 ist eine Dichte des Zufallsvektors $(\tau_1, \ldots, \tau_{k+l+1})$ gegeben durch

$$x = (x_1, \ldots, x_{k+l+1}) \mapsto \mathbb{1}_{[0,\infty)^{k+l+1}}(x)\alpha^{k+l+1}e^{-\alpha S_{k+l+1}(x)}.$$

Wir zeigen nun die Gleichung (5.4.2) für $l \geq 1$ (der Fall $l = 0$ ist analog). Es gilt

$$\begin{aligned}\mathbb{P}(N_{(0,s]} = k, N_{(s,t]} = l) = \mathbb{P}\big((\tau_1, \dots, \tau_{k+l+1}) \in A\big) &= \int_A \mathrm{d}x\, \alpha^{k+l+1} e^{-\alpha S_{k+l+1}(x)} \\ &= \alpha^{k+l} \int_0^\infty \cdots \int_0^\infty \mathrm{d}x_1 \dots \mathrm{d}x_{k+l+1}\, \alpha e^{-\alpha S_{k+l+1}(x)} \\ &\qquad \times \mathbb{1}\{S_k(x) \leq s < S_{k+1}(x), S_{k+l}(x) \leq t < S_{k+l+1}(x)\}.\end{aligned}$$

Wir integrieren nun schrittweise von innen nach außen. Zuerst halten wir $x_1, \dots, x_{k+l}$ fest und substituieren $z = S_{k+l+1}(x)$:

$$\int_0^\infty \mathrm{d}x_{k+l+1}\, \alpha e^{-\alpha S_{k+l+1}(x)} \mathbb{1}\{t < S_{k+l+1}(x)\} = \int_t^\infty \mathrm{d}z\, \alpha e^{-\alpha z} = e^{-\alpha t}.$$

Nun halten wir $x_1, \dots, x_k$ fest und substituieren $y_1 = S_{k+1}(x) - s, y_2 = x_{k+2}, \dots, y_l = x_{k+l}$:

$$\begin{aligned}&\int_0^\infty \cdots \int_0^\infty \mathrm{d}x_{k+1} \dots \mathrm{d}x_{k+l}\, \mathbb{1}\{s < S_{k+1}(x), S_{k+l}(x) \leq t\} \\ &\quad = \int_0^\infty \cdots \int_0^\infty \mathrm{d}y_1 \dots \mathrm{d}y_l\, \mathbb{1}\{y_1 + \cdots + y_l \leq t - s\} = \frac{(t-s)^l}{l!},\end{aligned}$$

wobei wir eine Induktion über $l$ benutzten. Die restlichen $k$ Integrale behandeln wir genauso und erhalten

$$\int_0^\infty \cdots \int_0^\infty \mathrm{d}x_1 \dots \mathrm{d}x_k\, \mathbb{1}\{S_k(x) \leq s\} = \frac{s^k}{k!}.$$

Wenn man dies alles zusammensetzt, ergibt sich die Behauptung. □

Die Konstruktion des Poisson-Prozesses, die durch Lemma 5.4.2 gegeben wird, hängt für uns in der Luft, da wir nicht unendlich viele unabhängige Zufallsgrößen mathematisch korrekt konstruieren können. Abgesehen von dieser Tatsache haben wir aber den Poisson-Prozess sehr befriedigend charakterisiert: Wir können ihn uns vorstellen als der Prozess von Zeitpunkten, zwischen denen unabhängige exponentielle Wartezeiten vergehen. Die erwartete Wartezeit zwischen aufeinander folgenden dieser Zeitpunkte beträgt $1/\alpha$, wobei $\alpha$ der Parameter der benutzten Exponentialverteilung ist.

Wenn man einen Poisson-Prozess als Modell für Zeitpunkte, an denen Busse von einer Haltestelle abfahren, verwendet, stellt man sich vielleicht folgende Frage: Wenn ich zum (festen) Zeitpunkt $t > 0$ an der Haltestelle ankomme und nicht weiß, wann der letzte Bus abgefahren ist, was ist die Verteilung der Wartezeit auf den nächsten Bus? Ist die erwartete Wartezeit vielleicht kürzer als $1/\alpha$? Und wie steht es um die Zeit, die seit der Abfahrt des letzten Busses bis jetzt (also den Zeitpunkt $t$) vergangen ist? Hat die Summe dieser zwei Zeiten den Erwartungswert $1/\alpha$? Die Antwort ist ein wenig überraschend und folgt im nächsten Lemma. Wir nehmen an, dass $N_t = N_{(0,t]}$, wobei $(N_I)_I$ ein wie im Lemma 5.4.2 konstruierter Poisson-Prozess ist. Ferner definieren wir für $t > 0$

$$W_t = -t + \min\{T_k \colon k \in \mathbb{N}, T_k > t\} \qquad V_t = t - \max\{T_k \colon k \in \mathbb{N}, T_k \leq t\}.$$

In Worten: $W_t$ ist die Wartezeit ab $t$ bis zur Abfahrt des nächsten Busses, und $V_t$ ist die Zeitdifferenz zwischen der letzten Abfahrt vor dem Zeitpunkt $t$ und $t$.

**Lemma 5.4.3 (Wartezeitparadox).** *Für jedes $t > 0$ ist $W_t$ zum Parameter $\alpha$ exponentialverteilt, und $V_t$ hat die Verteilung von $\min\{W_t, t\}$. Insbesondere gelten $\mathbb{E}(W_t) = 1/\alpha$ und $\mathbb{E}(V_t) = \frac{1}{\alpha}(1 - e^{-\alpha t})$. Die Zufallszeiten $W_t$ und $V_t$ sind unabhängig.*

**Beweis.** Das Ereignis $\{W_t > s\}$ ist identisch mit dem Ereignis, dass zwischen den Zeitpunkten $t$ und $t + s$ keines der $T_k$ eintrifft. Also gilt $\mathbb{P}(W_t > s) = \mathbb{P}(N_{(t,t+s]} = 0) = \mathrm{Po}_{\alpha s}(0) = e^{-\alpha s}$. Dies zeigt die erste Aussage.

Das Ereignis $\{V_t > s\}$ ist für $s < t$ identisch mit dem Ereignis, dass zwischen den Zeitpunkten $t-s$ und $t$ keines der $T_k$ eintrifft. Analog zu dem Obigen erhält man, dass $\mathbb{P}(V_t > s) = \mathbb{P}(N_{(t-s,t]} = 0) = e^{-\alpha s}$ für $s < t$. Für $s \geq t$ ist $\{V_t > s\} = \{N_{(0,t]} = 0\}$, also $\mathbb{P}(V_t > s) = e^{-\alpha t}$. Setzt man diese zwei Teilaussagen zusammen, erhält man die zweite Aussage des Lemmas.

Die Berechnung des Erwartungswertes von $\min\{W_t, t\}$ ist eine Übungsaufgabe. Die Unabhängigkeit der Zufallszeiten $W_t$ und $V_t$ ist eine einfache Konsequenz von Satz 5.4.2, denn $W_t$ und $V_t$ hängen nur von der Zahl der $T_k$ in $(t, \infty)$ bzw. in $(0, t]$ ab, und diese Anzahlen sind nach (P2) unabhängig (siehe Korollar 3.2.12). □

Die Wartezeit zwischen der Abfahrt des letzten Busses vor dem Zeitpunkt $t$ und des nächsten nach $t$ hat also *nicht* die Verteilung irgendeiner der Wartezeiten $\tau_i$. Tatsächlich ist ihr Erwartungswert größer und tendiert für große $t$ gegen das Zweifache des Erwartungswertes von $\tau_i$. Dieses Phänomen kann man damit interpretieren, dass eine Zwischenwartezeit 'größer' wird durch Beobachtung. Anders formuliert: Die Zwischenwartezeit $V_t + W_t$ hat die besondere Eigenschaft, dass der Zeitpunkt $t$ in dem betrachteten Zeitintervall liegt, und diese Eigenschaft vergrößert ihre Länge.

# Kapitel 6

# Grenzwertsätze

In diesem Kapitel behandeln wir die zwei wichtigsten Grenzwertsätze für Wahrscheinlichkeitsverteilungen: das *Gesetz der Großen Zahlen* und den *Zentralen Grenzwertsatz*, zumindest für wichtige Spezialfälle. Beide Sätze machen asymptotische Aussagen über sehr oft wiederholte unabhängige identisch verteilte Zufallsexperimente: Das Gesetz der Großen Zahlen formuliert, dass der Durchschnitt der dabei auftretenden Ergebnisse sich in gewissem Sinne dem erwarteten Wert eines dieser Experimente annähert, und der Zentrale Grenzwertsatz macht eine feinere Aussage über die Fluktuationen um diesen asymptotischen Wert.

Unter einer Zufallsgröße verstehen wir in diesem Kapitel eine diskrete im Sinne der Definition 3.1.1 oder eine stetige im Sinne von Abschnitt 5.1.

## 6.1 Das Gesetz der Großen Zahlen

Wir beginnen mit zwei einfachen, aber wichtigen allgemeinen Ungleichungen für die Wahrscheinlichkeit von Abweichungen einer Zufallsgröße von Null bzw. von ihrem Erwartungswert. Wir erinnern daran, dass der Erwartungswert einer nicht negativen Zufallsgröße immer definiert ist, aber eventuell gleich $\infty$ ist.

**Satz 6.1.1 (Markov-Ungleichung).** *Es sei $X$ eine Zufallsgröße und $\varphi\colon (0,\infty) \to (0,\infty)$ eine monoton wachsende Funktion. Dann gilt für jedes $\varepsilon > 0$*

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}(\varphi \circ |X|)}{\varphi(\varepsilon)}.$$

**Beweis.** Auf der Menge $\{|X| \geq \varepsilon\} = \{\omega \in \Omega\colon |X(\omega)| \geq \varepsilon\}$ gilt wegen der Monotonie von $\varphi$, dass $\varphi(\varepsilon) \leq \varphi(|X(\omega)|)$. Also gilt die Abschätzung

$$\mathbb{1}_{\{|X|\geq\varepsilon\}} \leq \frac{\varphi \circ |X|}{\varphi(\varepsilon)}.$$

Nun bildet man auf beiden Seiten den Erwartungswert und erhält die behauptete Ungleichung. □

Indem man die Markov-Ungleichung auf die Zufallsvariable $X - \mathbb{E}(X)$ statt $X$ und $\varphi(x) = x^2$ anwendet, erhält man die sehr nützliche folgende Ungleichung:

**Korollar 6.1.2 (Tschebyscheff-Ungleichung).** *Für jede Zufallsgröße $X \in \mathcal{L}^2$ und jedes $\varepsilon > 0$ gilt*

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

Die Tschebyscheff-Ungleichung kann im Allgemeinen nicht verbessert werden, wie das folgende Beispiel zeigt. Man betrachte eine Zufallsvariable $X$, die die Werte $\varepsilon$, 0 und $-\varepsilon$ annimmt mit den Wahrscheinlichkeiten $(2\varepsilon^2)^{-1}$, $1-\varepsilon^{-2}$ und $(2\varepsilon^2)^{-1}$. Dann sind $\mathbb{E}(X) = 0$ und $\mathbb{V}(X) = 1$, und es gilt Gleichheit in der Tschebyscheffschen Ungleichung.

Der Wert der Tschebyscheff-Ungleichung liegt in ihrer einfachen Handhabbarkeit und universellen Anwendbarkeit, sie gibt einen recht guten Eindruck von der Größenordnung der betrachteten Wahrscheinlichkeit. Sie ist immerhin gut genug, um einen kurzen Beweis des zentralen Ergebnisses dieses Abschnittes zu ermöglichen (siehe Satz 6.1.4).

Wir kommen nun zum Gesetz der Großen Zahlen. Wie so oft betrachten wir eine oftmalige Wiederholung eines zufälligen Experiments, bei dem jeweils ein zufälliges Ergebnis erzielt wird, das man mit einer Zufallsgröße angeben kann. Sei also $X_i \in \mathbb{R}$ das Ergebnis der $i$-ten Ausführung. Wir nehmen an, dass jede Zufallsgröße $X_i$ den gleichen Erwartungswert $E = \mathbb{E}(X_1) = \mathbb{E}(X_2) = \dots$ besitzt. Die Intuition sagt, dass die Folge der Mittelwerte $\frac{1}{n}S_n = \frac{1}{n}(X_1 + \cdots + X_n)$ sich für große $n$ der Zahl $E$ annähern sollte. Doch in welchen Sinn sollte das passieren? Eine Aussage wie '$\lim_{n\to\infty} \frac{1}{n}S_n = E$' können wir nicht in dieser Vorlesung behandeln, denn dazu müssten alle (unendlich vielen) Zufallsgrößen $X_1, X_2, \dots$ gleichzeitig definiert sein.[1] Wir werden die Annäherung von $\frac{1}{n}S_n$ an $E$ in der Form formulieren, dass die Wahrscheinlichkeit, dass $\frac{1}{n}S_n$ von $E$ einen gewissen positiven Abstand hat, mit $n \to \infty$ gegen Null geht. Dies gibt Anlass zu einer Definition:

**Definition 6.1.3 (Konvergenz in Wahrscheinlichkeit).** *Wir sagen, eine Folge $(Y_n)_{n\in\mathbb{N}}$ von Zufallsgrößen* konvergiert in Wahrscheinlichkeit *oder* konvergiert stochastisch *gegen eine Zufallsgröße $Y$, falls für jedes $\varepsilon > 0$ gilt:*

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - Y| > \varepsilon) = 0.$$

*In diesem Fall schreiben wir auch $Y_n \xrightarrow{\mathbb{P}} Y$.*

Es ist klar, dass $Y_n \xrightarrow{\mathbb{P}} Y$ genau dann gilt, wenn $Y_n - Y \xrightarrow{\mathbb{P}} 0$. Man beachte, dass die Konvergenz in Wahrscheinlichkeit nicht von den Zufallsgrößen abhängt, sondern nur von ihrer Verteilung. Insbesondere müssen für diesen Konvergenzbegriff nicht unendlich viele Zufallsgrößen auf einem Wahrscheinlichkeitsraum definiert werden, sondern jeweils nur eine einzige, nämlich $Y_n - Y$, dies allerdings für jedes $n \in \mathbb{N}$.

[1] Tatsächlich kann man diese Aussage (unter Zuhilfenahme von Maßtheorie) präzisieren und unter geeigneten Voraussetzungen beweisen; siehe die Vorlesung *Stochastik I*. Eine solche Aussage nennt man das *Starke Gesetz der Großen Zahlen*.

Der Begriff der Konvergenz in Wahrscheinlichkeit ist tatsächlich ein sehr geeigneter für die oben gestellte Frage nach einer sinnvollen Formulierung der Annäherung von $\frac{1}{n}S_n$ an $E$. Wir erinnern daran (siehe Abschnitt 3.5), dass zwei Zufallsvariable $X$ und $Y$ unkorreliert heißen, wenn $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

**Satz 6.1.4 (Schwaches Gesetz der Großen Zahlen).** *Für jedes $n \in \mathbb{N}$ seien paarweise unkorrelierte Zufallsgrößen $X_1, \dots, X_n$ gegeben, die alle den gleichen Erwartungswert $E \in \mathbb{R}$ und die gleiche Varianz $V < \infty$ besitzen. Sei $\frac{1}{n}S_n = \frac{1}{n}(X_1 + \cdots + X_n)$ der Mittelwert. Dann gilt für jedes $\varepsilon > 0$:*

$$\lim_{n\to\infty} \mathbb{P}(|\tfrac{1}{n}S_n - E| > \varepsilon) = 0,$$

*d. h., $\frac{1}{n}S_n$ konvergiert in Wahrscheinlichkeit gegen $E$.*

**Beweis.** Auf Grund der Linearität des Erwartungswerts ist $\mathbb{E}(\frac{1}{n}S_n) = E$, und auf Grund der paarweisen Unkorreliertheit ist

$$\mathbb{V}(\tfrac{1}{n}S_n) = \tfrac{1}{n^2}\mathbb{V}(X_1 + \cdots + X_n) = \tfrac{1}{n^2}n\mathbb{V}(X_1) = \tfrac{1}{n}V; \tag{6.1.1}$$

siehe den Satz von Bienaymé, Korollar 3.5.3. Also liefert eine Anwendung der Tschebyscheff-Ungleichung:

$$\mathbb{P}(|\tfrac{1}{n}S_n - E| > \varepsilon) = \mathbb{P}(|\tfrac{1}{n}S_n - \mathbb{E}(\tfrac{1}{n}S_n)| > \varepsilon) \le \frac{\mathbb{V}(\frac{1}{n}S_n)}{\varepsilon^2} = \frac{1}{n}\frac{V}{\varepsilon^2},$$

und dies konvergiert gegen Null für $n \to \infty$. □

Insbesondere gilt das Gesetz der Großen Zahlen also auch für unabhängige identisch verteilte Zufallsgrößen mit existierenden Varianzen. Natürlich gibt es weit stärkere Versionen des Gesetzes der Großen Zahlen in der Literatur, aber darum kümmern wir uns hier nicht. Der Mittelwert $\frac{1}{n}S_n$ der $n$ Ausführungen der Experimente liegt also in der Nähe des Erwartungswertes einer einzelnen Ausführung in dem Sinne, dass Abweichungen einer gegebenen positiven Größe eine verschwindende Wahrscheinlichkeit haben. Man beachte aber, dass im Allgemeinen *nicht* $\lim_{n\to\infty} \mathbb{P}(\frac{1}{n}S_n \neq E) = 0$ gilt. Zum Beispiel gilt für die eindimensionale Irrfahrt aus Abschnitt 4.3 (hier sind $X_1, X_2, \dots$ unabhängig und nehmen die Werte $-1$ und 1 mit Wahrscheinlichkeit $\frac{1}{2}$ an), dass $\mathbb{P}(\frac{1}{2n}S_{2n} \neq E) = 1 - \mathbb{P}(S_{2n} = 0) \to 1$, wie man an Korollar 4.3.2 sieht.

Unter gewissen zusätzlichen Integrierbarkeitsannahmen kann man sogar erhalten, dass die Geschwindigkeit der Konvergenz im Gesetz der Großen Zahlen sogar *exponentiell* ist:

**Lemma 6.1.5 (Große Abweichungen).** *Für jedes $n \in \mathbb{N}$ seien $X_1, \dots, X_n$ unabhängige, identisch verteilte Zufallsvariablen. Es existiere ein $\alpha > 0$, so dass $\mathbb{E}(e^{\alpha|X_1|}) < \infty$. Wir definieren wieder $E = \mathbb{E}(X_1)$ und $S_n = X_1 + \cdots + X_n$. Dann existiert für jedes $\varepsilon > 0$ ein $C > 0$ (das von $\varepsilon$ und von $\alpha$ abhängt), so dass*

$$\mathbb{P}(|\tfrac{1}{n}S_n - E| > \varepsilon) \le e^{-Cn}, \qquad n \in \mathbb{N}.$$

**Beweis.** Wir dürfen voraus setzen, dass $E = \mathbb{E}(X_1) = 0$. Wir werden nur zeigen, dass $\mathbb{P}(\frac{1}{n}S_n > \varepsilon) \le e^{-Cn}$ für alle $n \in \mathbb{N}$ und ein geeignetes $C$. Der Beweis der anderen Hälfte der Aussage, $\mathbb{P}(\frac{1}{n}S_n < -\varepsilon) \le e^{-Cn}$, läuft analog, und dann folgt die Aussage des Lemmas mit einem eventuell anderen Wert von $C > 0$.

Wir fixieren ein $\beta \in (0, \alpha/2)$ und benutzen die Markov-Ungleichung (siehe Satz 6.1.1) für die Abbildung $\varphi(x) = e^{\beta x}$ wie folgt:

$$\mathbb{P}(\tfrac{1}{n}S_n > \varepsilon) = \mathbb{P}(X_1 + \cdots + X_n > \varepsilon n) = \mathbb{P}\big(e^{\beta(X_1+\cdots+X_n)} > e^{\beta\varepsilon n}\big) \leq \mathbb{E}\big(e^{\beta(X_1+\cdots+X_n)}\big)e^{-\beta\varepsilon n}.$$

Den auftretenden Erwartungswert kann man mit Hilfe der Unabhängigkeit von $e^{\beta X_1}, \ldots, e^{\beta X_n}$ (siehe Korollar 3.2.12) und Lemma 3.3.2(d) berechnen zu

$$\mathbb{E}\big(e^{\beta(X_1+\cdots+X_n)}\big) = \mathbb{E}(e^{\beta X_1})^n.$$

Auf Grund unserer Integrierbarkeitsvoraussetzung und wegen $\beta < \alpha$ ist diese obere Schranke endlich. Wenn wir sie oben einsetzen, erhalten wir die Abschätzung

$$\mathbb{P}(\tfrac{1}{n}S_n > \varepsilon) \leq \exp\Big\{-n\big[\beta\varepsilon - \log \mathbb{E}(e^{\beta X_1})\big]\Big\}. \tag{6.1.2}$$

Nun zeigen wir, dass für genügend kleines $\beta > 0$ der Term in $[\ldots]$ in (6.1.2) positiv ist. Zunächst schätzen wir mit einem großen $R > 0$ für $\beta \in (0, \varepsilon/2)$ ab:

$$\begin{aligned}
\mathbb{E}(e^{\beta X_1}) &\leq \mathbb{E}(e^{\beta X_1}\mathbb{1}\{X_1 \leq R\}) + \mathbb{E}(e^{\beta X_1}\mathbb{1}\{X_1 > R\}) \\
&\leq \mathbb{E}(e^{\beta X_1}\mathbb{1}\{X_1 \leq R\}) + \mathbb{E}(e^{\beta X_1}e^{(\varepsilon-\beta)(X_1-R)}) \\
&\leq \mathbb{E}(e^{\beta X_1}\mathbb{1}\{X_1 \leq R\}) + \mathbb{E}(e^{\varepsilon X_1})e^{(\beta-\varepsilon)R} \\
&\leq \mathbb{E}(e^{\beta X_1}\mathbb{1}\{X_1 \leq R\}) + \mathbb{E}(e^{\varepsilon X_1})e^{-\varepsilon R/2}.
\end{aligned}$$

Nun wählen wir $R > 0$ so groß, dass der zweite Summand nicht größer ist als $e^{-\varepsilon R/3}$. Wir können eine Taylorentwicklung benutzen, um für $\beta \downarrow 0$ zu erhalten:

$$\mathbb{E}(e^{\beta X_1}\mathbb{1}\{X_1 \leq R\}) = \mathbb{E}\big((1 + \beta X_1 + \mathcal{O}(\beta^2))\mathbb{1}\{X_1 \leq R\}\big) \leq 1 + \mathcal{O}(\beta^2),$$

wobei wir auch $\mathbb{E}(X_1\mathbb{1}\{X_1 \leq R\}) \leq \mathbb{E}(X_1) = 0$ abgeschätzt haben. Dann haben wir die folgende untere Schranke für den Term in $[\ldots]$ in (6.1.2):

$$\beta\varepsilon - \log \mathbb{E}(e^{\beta X_1}) \geq \beta\varepsilon - \log\Big(1 + \mathcal{O}(\beta^2) + e^{-\varepsilon R/3}\Big) \geq \beta\varepsilon - \mathcal{O}(\beta^2) - e^{-\varepsilon R/3},$$

wobei wir die Ungleichung $\log(1 + x) \leq x$ benutzten. Nun sehen wir, dass dies für genügend kleines $\beta > 0$ und etwa $R = 1/\beta$ positiv ist, und dies beendet den Beweis. □

## 6.2 Der Zentrale Grenzwertsatz

Wir gehen zurück zu der im vorigen Abschnitt beschriebenen Situation von oftmaligen Ausführungen eines Zufallsexperiments und fragen uns: Wenn also $\frac{1}{n}S_n - E$ gegen Null geht, mit welcher Rate passiert denn das?[2] Ist diese Größe typischerweise von der Ordnung $n^{-1}$ oder $1/\log n$ oder $e^{-n}$ oder von welcher sonst? Eine grobe Antwort wird gegeben durch eine etwas trickreichere Anwendung der Tschebyscheffschen Ungleichung, als dies im Beweis des Gesetzes der Großen Zahlen geschehen ist: Sie liefert für jedes $\alpha > 0$ die Abschätzung

$$\mathbb{P}\big(n^\alpha|\tfrac{1}{n}S_n - E| > \varepsilon\big) \leq n^{2\alpha-1}\frac{V}{\varepsilon^2},$$

[2] Man beachte, dass diese Frage sehr verschieden ist von der Frage, mit welcher Rate die Wahrscheinlichkeit $\mathbb{P}(|\frac{1}{n}S_n - E| \geq \varepsilon)$ gegen Null geht.

und dies konvergiert für jedes $\alpha < \frac{1}{2}$ gegen Null. Dies legt die Vermutung nahe, dass $\frac{1}{n}S_n - E$ von der Größenordnung $n^{-1/2}$ sein sollte und dass die 'aufgeblähte' Zufallsvariable $\sqrt{n}(\frac{1}{n}S_n - E)$ gegen etwas Nichttriviales konvergieren könnte. Dies stellt sich auch als korrekt heraus, wie wir im folgenden Satz 6.2.2 sehen werden. Tatsächlich kommt sogar in einer Vielzahl von Fällen immer die selbe Grenzverteilung heraus, und zwar die Normalverteilung (siehe Beispiel 5.3.7). Der Sinn, in dem die Konvergenz stattfindet, ist der folgende.

**Definition 6.2.1 (Schwache Konvergenz).** *Es seien Zufallsgrößen $X$ und $X_1, X_2, \ldots$ gegeben mit Verteilungsfunktionen $F$ bzw. $F_1, F_2, \ldots$. Wir sagen, dass $X_n$* in Verteilung *oder* schwach gegen $X$ konvergiert*, falls für jedes $t \in \mathbb{R}$, in dem $F$ stetig ist, gilt*

$$\lim_{n\to\infty} F_n(t) = F(t).$$

*In diesem Fall schreiben wir $X_n \xrightarrow{\mathrm{w}} X$.*

Die Notation $X_n \xrightarrow{\mathrm{w}} X$ lehnt sich an dem englischen Ausdruck 'weak convergence' für schwache Konvergenz an. Natürlich konvergiert $X_n$ genau dann schwach gegen $X$, wenn $X_n - X$ schwach gegen Null konvergiert. Wie beim Begriff der Konvergenz in Wahrscheinlichkeit im vorigen Abschnitt braucht man streng genommen keine Zufallsgrößen, um die Konvergenz zu formulieren, sondern nur deren Verteilungen, und zwar nur die von $X_n - X$ für jedes $n \in \mathbb{N}$. Man kann leicht zeigen, dass $X_n \xrightarrow{\mathrm{w}} X$ genau dann gilt, wenn für alle $a < b$, so dass die Verteilungsfunktion von $X$ in $a$ und in $b$ stetig ist, gilt

$$\lim_{n\to\infty} \mathbb{P}(X_n \in [a,b]) = \mathbb{P}(X \in [a,b]).$$

Wir formulieren nun das zentrale Ergebnis dieses Abschnitts, das wir allerdings nur in einem Spezialfall beweisen werden, siehe Satz 6.2.5.

**Satz 6.2.2 (Zentraler Grenzwertsatz).** *Für jedes $n \in \mathbb{N}$ seien unabhängige und identisch verteilte Zufallsgrößen $X_1, \ldots, X_n$ gegeben, die alle den gleichen Erwartungswert $E$ und die gleiche Varianz $V$ besitzen. Sei $\frac{1}{n}S_n = \frac{1}{n}(X_1 + \ldots, X_n)$ der Mittelwert. Dann gilt für jedes $t \in \mathbb{R}$*

$$\lim_{n\to\infty} \mathbb{P}\Big(\sqrt{\frac{n}{V}}\Big(\frac{1}{n}S_n - E\Big) \le t\Big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}}\,\mathrm{d}x. \tag{6.2.1}$$

*Mit anderen Worten, $\sqrt{\frac{n}{V}}(\frac{1}{n}S_n - E)$ konvergiert schwach gegen eine standardnormalverteilte Zufallsgröße:*

$$\sqrt{\frac{n}{V}}\Big(\frac{1}{n}S_n - E\Big) \xrightarrow{\mathrm{w}} \mathcal{N}(0,1).$$

**Beweis.** Siehe die Vorlesung *Stochastik I.* □

Man beachte, dass der Erwartungswert von $\sqrt{\frac{n}{V}}(S_n - E)$ gleich Null und ihre Varianz gleich Eins ist, genau wie die der Standardnormalverteilung $\mathcal{N}(0,1)$. Solche Zufallsvariable nennt man

*standardisiert*. Man beachte auch, dass gilt

$$\sqrt{\frac{n}{V}}\Big(\frac{1}{n}S_n - E\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - \mathbb{E}(X_i)}{S(X_i)},$$

wobei $S(X_i)$ die Standardabweichung von $X_i$ ist. Auch die Zufallsgrößen $\frac{X_i - \mathbb{E}(X_i)}{S(X_i)}$ sind standardisiert, und man kann den Zentralen Grenzwertsatz auch (ein wenig schlampig) formulieren, indem man sagt: *Die Summe von $n$ unabhängigen standardisierten Zufallsgrößen gleicher Verteilung ist asymptotisch verteilt wie $\sqrt{n}$ mal eine Standardnormalvariable.*

**Bemerkung 6.2.3.** In der Situation des Satzes 6.2.2 haben wir also für jedes $C > 0$:

$$\lim_{n\to\infty}\mathbb{P}\Big(\Big|\frac{1}{n}S_n - E\Big| > \frac{CV}{\sqrt{n}}\Big) = 2(1 - \Phi(C)),$$

und dies konvergiert für $C \to \infty$ gegen Null. Insbesondere kann man leicht die Gültigkeit des Schwachen Gesetzes der Großen Zahlen beweisen, indem man bei gegebenem $\varepsilon > 0$ für genügend großes $n$ abschätzt: $CV/\sqrt{n} < \varepsilon$. Die Aussage des Zentralen Grenzwertsatzes ist also stärker als die des Schwachen Gesetzes der Großen Zahlen. ◇

**Bemerkung 6.2.4 (Der Zentrale Grenzwertsatz als Fixpunktsatz).** Es ist bemerkenswert, dass die Grenzverteilung nicht abhängt von den Details der Verteilung der $X_i$ und immer gleich der Normalverteilung $\mathcal{N}(0,1)$ ist. Ein wenig Plausibilität dafür kommt von der speziellen Eigenschaft der Normalverteilung in Lemma 5.3.8 her: Falls alle $X_i$ exakt $\mathcal{N}(0,1)$-verteilt sind, so ist auch $\sqrt{\frac{n}{V}}(\frac{1}{n}S_n - E) = n^{-1/2}\sum_{i=1}^{n} X_i$ exakt $\mathcal{N}(0,1)$-verteilt. Dies heißt, dass die Normalverteilung ein Fixpunkt ist unter der Abbildung, die eine Wahrscheinlichkeitsverteilung $\mathbb{P}$ abbildet auf die Verteilung von $2^{-1/2}$ Mal die Summe zweier unabhängiger Zufallsgrößen mit Verteilung $\mathbb{P}$. In Termen der zugehörigen Dichten ist dies die Abbildung $\varphi \mapsto \varphi\star\varphi(\,\cdot\, 2^{-1/2})2^{-1/2}$. Man kann zumindest die Teilfolge $2^{-n/2}\sum_{i=1}^{2^n} X_i$ (wobei $X_1, X_2, \ldots$ standardisierte unabhängige identisch verteilte Zufallsgrößen im $\mathcal{L}^2$ sind) auffassen als die Iterationsfolge unter der oben beschriebenen Abbildung. Da die Standardnormalverteilung ein Fixpunkt dieser Abbildung ist, ist es zumindest nicht unplausibel, dass ein 'Fixpunktsatz' wie der Zentrale Grenzwertsatz gelten könnte. ◇

Tatsächlich werden wir den Zentralen Grenzwertsatz in der Form von Satz 6.2.2 nicht beweisen, sondern nur den Spezialfall für Bernoulli-Zufallsgrößen:

**Satz 6.2.5 (Satz von de Moivre-Laplace).** *Es seien $X_1, \ldots, X_n$ Bernoulli-Zufallsgrößen mit Parameter $p \in (0,1)$, d. h. $S_n = \sum_{i=1}^{n} X_i$ ist binomialverteilt mit Parametern $n$ und $p$. Dann gilt für alle $a, b \in \mathbb{R}$ mit $a < b$*

$$\lim_{n\to\infty}\mathbb{P}\Big(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\Big) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\,\mathrm{d}x.$$

**Beweis.** Wir schreiben $q$ statt $1 - p$.

Wir schreiben zunächst die betrachtete Wahrscheinlichkeit als eine Summe von Einzelwahrscheinlichkeiten und skalieren diese Summe zu einem Integral, wobei wir die Randeffekte bei $a$

und bei $b$ pauschal mit $o(1)$ für $n \to \infty$ beschreiben:

$$\begin{aligned}\mathbb{P}\Big(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\Big) &= \sum_{k \approx np + a\sqrt{npq}}^{\approx np + b\sqrt{npq}} \mathrm{Bi}_{n,p}(k) + o(1) \\ &= \int_{np+a\sqrt{npq}}^{np+b\sqrt{npq}} \mathrm{Bi}_{n,p}(\lfloor t \rfloor)\,\mathrm{d}t + o(1) \\ &= \int_a^b \sqrt{npq}\,\mathrm{Bi}_{n,p}\big(\lfloor np + x\sqrt{npq}\rfloor\big)\,\mathrm{d}x + o(1).\end{aligned}$$

Der Rest des Beweises besteht darin zu zeigen, dass der Integrand gleichmäßig in $x \in [a,b]$ gegen $\varphi_{0,1}(x) = (2\pi)^{-1/2} e^{-x^2/2}$ konvergiert. Wir schreiben zunächst die Definition von $\mathrm{Bi}_{n,p}(k)$ aus und setzen Stirling's Formel (siehe (4.3.1)) asymptotisch für $n \to \infty$ ein. Man beachte, dass diese und alle folgenden Approximationen gleichmäßig in $x \in [a,b]$ gelten.

$$\begin{aligned}\sqrt{npq}\,\mathrm{Bi}_{n,p}\big(\lfloor np + x\sqrt{npq}\rfloor\big) &\sim \sqrt{npq}\,\frac{\big(\frac{n}{e}\big)^n \sqrt{2\pi n}}{\big(\frac{np+x\sqrt{npq}}{ep}\big)^{np+x\sqrt{npq}} \sqrt{2\pi np}\big(\frac{nq-x\sqrt{npq}}{eq}\big)^{nq-x\sqrt{npq}} \sqrt{2\pi nq}} \\ &= \frac{1}{\sqrt{2\pi}}\Big(1 + \frac{x}{\sqrt{n}}\sqrt{\frac{q}{p}}\Big)^{-np-x\sqrt{npq}} \Big(1 - \frac{x}{\sqrt{n}}\sqrt{\frac{p}{q}}\Big)^{-nq+x\sqrt{npq}}.\end{aligned} \tag{6.2.2}$$

Nun schreiben wir die letzten beiden Terme mit Hilfe von $\exp\{\dots\}$ und benutzen die Asymptotik $(1 + \frac{c_1}{\sqrt{n}})^{c_2\sqrt{n}} \to e^{c_1 c_2}$. Also ist die rechte Seite von (6.2.2) asymptotisch gleich

$$\frac{1}{\sqrt{2\pi}} \exp\Big\{-np\log\Big(1 + \frac{x}{\sqrt{n}}\sqrt{\frac{q}{p}}\Big) - nq\log\Big(1 - \frac{x}{\sqrt{n}}\sqrt{\frac{p}{q}}\Big)\Big\} e^{-x^2\sqrt{q/p}\sqrt{pq}} e^{-x^2\sqrt{p/q}\sqrt{pq}}. \tag{6.2.3}$$

Nun benutzen wir eine Taylor-Approximation für den Logarithmus: $\log(1+h) = h - h^2/2 + \mathcal{O}(h^3)$ für $h \to 0$. Also ist der Term in (6.2.3) asymptotisch äquivalent zu

$$\frac{1}{\sqrt{2\pi}} \exp\Big\{-np\frac{x}{\sqrt{n}}\sqrt{\frac{q}{p}} + np\frac{x^2}{2n}\frac{q}{p} + nq\frac{x}{\sqrt{n}}\sqrt{\frac{p}{q}} + nq\frac{x^2}{2n}\frac{p}{q} - x^2 q - x^2 p\Big\}. \tag{6.2.4}$$

Elementares Zusammenfassen zeigt, dass dies identisch ist mit $\varphi_{0,1}(x)$. □

**Beispiel 6.2.6 (Irrfahrt).** Wenn $S_n$ der Endpunkt der $n$-schrittigen Irrfahrt aus Abschnitt 4.3 ist, so wissen wir nun, dass die Verteilung von $S_n n^{-1/2}$ schwach gegen die Standardnormalverteilung konvergiert. Insbesondere wissen wir nun, dass der Endpunkt $S_n$ des Pfades mit hoher Wahrscheinlichkeit in einem Intervall der Größenordnung $\sqrt{n}$ um die Null herum zu finden ist. Das Gesetz der Großen Zahlen sagte uns nur, dass er in dem (weit größeren) Intervall der Ordnung $\varepsilon n$ ist, allerdings mit höherer Wahrscheinlichkeit. ◇

**Beispiel 6.2.7 (Normalapproximation).** Eine typische Anwendung des Satzes von de Moivre-Laplace ist die Folgende. Wenn man einen fairen Würfel 1200 Mal wirft, wie groß ist dann die Wahrscheinlichkeit, dass dabei zwischen 190 und 200 Sechsen fallen? Wir nehmen natürlich an, dass die 1200 Würfe unabhängig voneinander geschehen.

Wir können die Zufallsgrößen $X_i$ interpretieren als 1, falls der $i$-te Wurf eine Sechs hervor bringt und als 0 andernfalls. Dann ist die Anzahl der gewürfelten Sechsen gleich $S_{1200} = X_1 + \cdots +$

$X_{1200}$, und wir sind in der Situation des Satzes von de Moivre-Laplace mit $p = \frac{1}{6}$ und $n = 1200$. Insbesondere sind $np = 200$ und $\sqrt{np(1-p)} = \sqrt{\frac{500}{3}} \approx 13$. Die gesuchte Wahrscheinlichkeit müssen wir wie folgt umschreiben:

$$\begin{aligned}\mathbb{P}(190 \le S_{1200} \le 200) &\approx \mathbb{P}\Big(\frac{190-200}{13} \le \frac{S_n - np}{\sqrt{np(1-p)}} \le 0\Big) \\ &= \mathbb{P}\Big(-\frac{10}{13} \le \frac{S_n - np}{\sqrt{np(1-p)}} \le 0\Big) \approx \Phi(0) - \Phi(-\tfrac{10}{13}).\end{aligned}$$

Es ist klar, dass $\Phi(0) = \frac{1}{2}$, und in einer Tabelle finden wir den Wert $\Phi(-\frac{10}{13}) = 1 - \Phi(\frac{10}{13}) \approx 1 - \Phi(0,77) \approx 1 - 0,7794 = 0,2306$. Also ist die gesuchte Wahrscheinlichkeit ungefähr gleich $0,2794$. ◇

**Beispiel 6.2.8 (Wahlprognose).** Bei einer Wahl erhält Kandidat $A$ einen unbekannten Anteil $p \in (0,1)$ der Stimmen. Um den Wert von $p$ zu ermitteln, werten wir die ersten $n$ Wahlzettel aus (bzw. befragen wir $n$ zufällig gewählte Wähler). Wir groß sollte $n$ sein, damit die Wahrscheinlichkeit eines Irrtums von mehr als einem Prozent nicht größer als $0,05$ ist?

Wenn wir $n$ Zettel auswerten bzw. $n$ Personen befragen, dann bekommen wir $S_n$ Stimmen für den Kandidaten $A$, und $S_n$ ist binomialverteilt mit Parametern $n$ und $p$. Die Wahrscheinlichkeit des Ereignisses $\{|\frac{1}{n}S_n - p| > 0,01\}$ soll unter $0,05$ liegen, also

$$\begin{aligned}0,95 &\approx \mathbb{P}\Big(-\frac{0,01n}{\sqrt{np(1-p)}} \le \frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{0,01n}{\sqrt{np(1-p)}}\Big) \\ &\approx \Phi\Big(0,01\sqrt{\frac{n}{p(1-p)}}\Big) - \Phi\Big(-0,01\sqrt{\frac{n}{p(1-p)}}\Big) \\ &= 2\Phi\Big(0,01\sqrt{\frac{n}{p(1-p)}}\Big) - 1.\end{aligned}$$

Also wollen wir ein (möglichst kleines) $n$ bestimmen mit $\Phi\Big(0,01\sqrt{\frac{n}{p(1-p)}}\Big) \approx 0,975$. Einer Tabelle entnehmen wir, dass $\Phi(1,96) \approx 0,975$. Also sollte $n \approx p(1-p)10000 \cdot 1,96^2$ ausreichen. Wenn man keine Informationen über $p$ besitzt, dann kann man (indem man den maximalen Wert $p = \frac{1}{2}$ einsetzt) davon ausgehen, dass $n \approx \frac{1}{4}10000 \cdot 1,96^2 \approx 9600$ ausreicht. ◇

# Kapitel 7

# Einführung in die Schätztheorie

Eine Grundaufgabe der Statistik lautet: 'Gegeben zufällige Beobachtungen, was ist das wahrscheinlichkeitstheoretische Modell, das am besten diese Beobachtungen beschreibt?' Natürlich müssen wir mittels einer geschulten Intuition zunächst einen Vorrat an sinnvollen Modellen zur Verfügung stellen, unter denen wir dann das beste oder ein möglichst gutes heraus suchen. Das bedeutet, wir werden es mit einer ganzen *Familie* von Wahrscheinlichkeitsräumen zu tun haben, unter denen wir aussuchen werden. Auch werden verschiedene Qualitätskriterien diskutiert und verglichen werden müssen, nach denen wir auswählen werden.

Im vorliegenden, einführenden Kapitel kümmern wir uns zunächst nur um die Schätzung einer unbekannten Größe, die wir als zufällig annehmen.

## 7.1 Grundbegriffe

An einem klassischen Beispiel führen wir in die Problematik ein.

**Beispiel 7.1.1 (Schätzung eines Fischbestandes).** In einem Teich tummelt sich eine unbekannte Anzahl $N$ von Fischen, die wir schätzen wollen. Dazu gehen wir in zwei Schritten vor: Zunächst fischen wir $W$ Fische heraus, markieren sie und setzen sie wieder im Teich aus. Nach einer gewissen Wartezeit (damit sich die markierten und unmarkierten Fische gut miteinander mischen können) fischen wir $n$ Fische und zählen darunter genau $x$ markierte. Auf Grund dieses Ergebnisses wollen wir nun eine plausible Schätzung für $N$ abgeben. (Selbstverständlich machen wir stillschweigend wieder eine Reihe von vereinfachenden Annahmen, die wir nicht im Einzelnen erwähnen wollen.)

Eine simple, plausible Möglichkeit ist es anzunehmen, dass der Anteil der markierten Fische in der Stichprobe so groß ist wie im ganzen Teich, also schätzen wir, dass es $N_1 = Wn/x$ Fische im Teich gibt.

Eine kompliziertere, genauere Möglichkeit ist, zu Grunde zu legen, dass die Zahl $x$ der markierten gefangenen Fische hypergeometrisch verteilt sein sollte mit Parametern $N - W$, $W$ und $n - x$; siehe Beispiel 1.3.1. Wenn wir für eine Weile mit dem unbekannten Parameter $N$ rechnen (die anderen Parameter sind ja bekannt), dann besitzt unser Ergebnis von $x$ markierten Fischen die Wahrscheinlichkeit

$$p_N(x) = \mathrm{Hyp}_{n,N-W,W}(x) = \frac{\binom{N-W}{n-x}\binom{W}{x}}{\binom{N}{n}}.$$

Der zweite Ansatz besteht nun darin, denjenigen Wert von $N$ als Schätzung zu wählen, der diese Wahrscheinlichkeit $p_N(x)$ maximiert. Eine simple Rechnung zeigt, dass

$$\frac{p_N(x)}{p_{N-1}(x)} = \frac{(N-W)(N-n)}{N(N-W-n+x)} = 1 + \frac{Wn-Nx}{N(N-W-n+x)},$$

und daher ist $p_N(x)$ maximal genau für $N = N_2 = nW/x$. (Also läuft die zweite Methode auf das Ergebnis der ersten hinaus.) Wir wählen also denjenigen Wert von $N$, der das mathematische Modell (hier die hypergeometrische Verteilung) am besten 'passend' macht, also die größte Ähnlichkeit mit der Messung hervor bringt. Daher nennt man $N_2$ auch den *Maximum-Likelihood-Schätzer*; siehe Definition 7.3.1. ◇

Aus diesem Beispiel ziehen wir schon einige Lehren:

(i) Es gibt im Allgemeinen mehrere plausible Schätzer.

(ii) Man braucht eine wahrscheinlichkeitstheoretische Intuition bzw. eine einschlägige Schulung, um eine Klasse von geeigneten Modellen zu wählen.

(iii) Eine plausible Methode ist es, unter einer ganzen Familie von Modellen durch Optimierung dasjenige herauszusuchen, das das beobachtete Ergebnis mit höchster Wahrscheinlichkeit versieht.

(iv) Das Optimieren ist relativ einfach, wenn die Familie von einem Parameter abhängt, so dass man ein eindimensionales Maximierungsproblem lösen muss.

(v) Der zu optimierende Parameterwert hat *nicht* die Interpretation des 'wahrscheinlichsten' Wertes der zu schätzenden Zufallsgröße (auf der Menge der Parameter ist ja keine Verteilung definiert), sondern optimiert die Übereinstimmung des gewählten Modells mit den Messwerten.

Die grundlegenden Begriffe sind die folgenden. Nach wie vor werden wir auf Benutzung von Maßtheorie verzichten. Alle auftretenden Wahrscheinlichkeitsmaße sind also im Sinne eines diskreten Maßes wie in Definition 1.1.2 zu verstehen oder als ein Maß, das mit einer Dichte definiert wird, wie in Abschnitt 5.1.

**Definition 7.1.2 (statistisches Modell).** *Ein* statistisches Modell *ist ein Paar* $(\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$*, bestehend aus einer Menge* $\mathfrak{X}$*, die entweder höchstens abzählbar ist oder eine Teilmenge eines* $\mathbb{R}^n$*, und einer Familie von Wahrscheinlichkeitsmaßen* $\mathbb{P}_\vartheta$ *auf* $\mathfrak{X}$ *mit einer Indexmenge* $\Theta$*. Man nennt* $\mathfrak{X}$ *einen* Stichprobenraum*. Das Modell heißt* parametrisch*, falls* $\Theta \subset \mathbb{R}^d$ *für ein* $d \in \mathbb{N}$*, und es heißt in diesem Fall* einparametrig*, wenn* $d = 1$*.*

Im Falle von diskretem $\mathfrak{X}$ ist also für jedes $\vartheta \in \Theta$ das Tupel $(\mathfrak{X}, p_\vartheta)$ ein Wahrscheinlichkeitsraum, wobei $p_\vartheta(x) = \mathbb{P}_\vartheta(\{x\})$. Im Fall, dass $\mathbb{P}_\vartheta$ eine Dichte hat, haben wir den Begriff eines Wahrscheinlichkeitsraums nicht spezifiziert. Was wir ein statistisches Modell genannt haben, wird in der Literatur oft ein statistisches *Standard*modell genannt. Die zu $\mathbb{P}_\vartheta$ gehörige Erwartung und Varianz werden mit $\mathbb{E}_\vartheta$ bzw. $\mathbb{V}_\vartheta$ bezeichnet. Die Menge $\mathfrak{X}$ spielt also mathematisch die selbe Rolle wie die Grundmenge $\Omega$ eines Wahrscheinlichkeitsraums. Die zu Grunde liegende

Vorstellung ist, dass ein abstraktes $\Omega$ im Hintergrund existiert, aber eine Zufallsgröße $X\colon \Omega \to \mathfrak{X}$ die tatsächlich beobachtbaren Ereignisse beschreibt.

Im Beispiel 7.1.1 benutzten wir also das einparametrige Modell mit $\mathfrak{X} = \mathbb{N}$, und die Familie $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ mit $\Theta = \mathbb{N}$ war eine Familie gewisser hypergeometrischen Verteilungen.

Für Beobachtungen, die aus ganzen Serien unabhängiger Experimente gewonnen werden, brauchen wir noch den folgenden Begriff.

**Definition 7.1.3 (Produktmodell).** *Es seien $\mathcal{M} = (\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $n \in \mathbb{N}$. Dann heisst $\mathcal{M}^{\otimes n} = (\mathfrak{X}^n, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ das zugehörige $n$-fache* Produktmodell*, wobei für jedes $\vartheta \in \Theta$ mit $(\mathfrak{X}^n, \mathbb{P}_\vartheta^{\otimes n})$ der $n$-fache Produktraum von $(\mathfrak{X}, \mathbb{P}_\vartheta)$ bezeichnet wird (siehe Definition 2.3.1 bzw. Lemma 5.2.4). In diesem Fall bezeichnen wir mit $X_i\colon \mathfrak{X}^n \to \mathfrak{X}$ die $i$-te Koordinate.*

Insbesondere sind dann $X_1, \dots, X_n$ unabhängig bezüglich jedes $\mathbb{P}_\vartheta^{\otimes n}$ mit Verteilung $\mathbb{P}_\vartheta$.

## 7.2 Beispiele für Schätzer

Die Denkweise der Statistik spiegelt sich auch in der Begriffsbildung wider:

**Definition 7.2.1 (Statistik, Schätzer).** *Es sei $(\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\Sigma$ ein Ereignisraum.*

*(i) Jede Zufallsvariable $S\colon \mathfrak{X} \to \Sigma$ heißt eine* Statistik*.*

*(ii) Sei $\tau\colon \Theta \to \Sigma$ eine Abbildung, dann heißt jede Statistik $T\colon \mathfrak{X} \to \Sigma$ ein* Schätzer *für $\tau$.*

Das mathematische Objekt 'Zufallsvariable' hat für den Statistiker also nichts Zufälliges mehr an sich, sondern besitzt die Interpretation des gemessenen Ergebnisses eines wohldurchdachten Experimentes. Die Abbildung $\tau$ spielt die Rolle einer Kenngröße $\tau(\vartheta)$ für den Parameter $\vartheta$; meist werden wir $\tau(\vartheta) = \vartheta$ wählen. Dass ein Schätzer für $\tau$ *a priori* nichts mit $\tau$ zu tun haben muss, soll nur eine zu starke Einengung der Definition verhindern.

Es folgen Beispiele, in denen wir mehrere verschiedene Schätzer kurz vorstellen. Wir werden Schätzer kennen lernen, die nach unterschiedlichen Kriterien plausibel oder sogar optimal sind. Der theoretische Hintergrund wird in den folgenden Abschnitten behandelt werden.

**Beispiel 7.2.2 (Taxiproblem).** In einer großen Stadt gibt es $N$ Taxis, deren Anzahl wir schätzen wollen, indem wir an einer belebten Kreuzung stehen und über eine gewisse Zeit hinweg die Nummern der vorbeifahrenden Taxis registrieren, wobei wir Wiederholungen ignorieren. Wir nehmen dabei an, dass jedes Taxi eine Nummer zwischen 1 und $N$ leicht sichtbar trägt, dass alle Taxis zu dieser Zeit in Betrieb sind, ihre Fahrten gleichmäßig über die ganze Stadt verteilen und insbesondere an unserer Kreuzung mit etwa der gleichen Wahrscheinlichkeit vorbei fahren. Wir registrieren, bis wir genau $n$ verschiedene Taxis gesehen haben. Das Ergebnis unserer Beobachtung ist eine Teilmenge $\{x_1, \dots, x_n\}$ von $\{1, \dots, N\}$ mit $1 \le x_1 < x_2 < \dots < x_n \le N$. Auf Grund dieses Ergebnisses soll nun $N$ geschätzt werden.

Eine simple (nicht zu plausible) Möglichkeit ist es, den Schätzer $N_1 = x_n$, die höchste aufgetretene Nummer, zu wählen. Dies ist sogar ein Maximum-Likelihood-Schätzer, denn wenn

wir annehmen, dass alle $\binom{N}{n}$ Teilmengen von $\{1,\dots,N\}$ mit $n$ Elementen gleich wahrscheinlich sind, dann ist die Wahrscheinlichkeit unserer Beobachtung gerade $\binom{N}{n}^{-1}$, und diese Zahl ist maximal für minimale $N$. (Formal gesehen, haben wir dieser Überlegung das statistische Modell $(\mathfrak{X},(\mathbb{P}_N)_{N\in\mathbb{N}})$ mit $\mathfrak{X}=\{A\subset\mathbb{N}\colon |A|=n\}$ und $\mathbb{P}_N=$ Gleichverteilung auf $\{A\subset\{1,\dots,N\}\colon |A|=n\}$ zu Grunde gelegt.) Dieser Schätzer unterschätzt den gesuchten Wert systematisch, so dass wir nach besseren Schätzern Ausschau halten.

Eine bessere Idee ist zu bemerken, dass die Zahl der nicht registrierten Taxinummern unter $x_1$ aus Symmetriegründen gleich der Zahl der nicht registrierten Nummern $> x_n$ sein sollte, also wählt man den Schätzer $N_2 = x_n + x_1 - 1$.

Eine noch bessere Idee ist es, die Lücke zwischen $N$ und $x_n$ mit der mittleren Länge zwischen den Zahlen $0, x_1, x_2, \dots, x_n$ zu schätzen, was auf den Schätzer $N_3 = x_n + (x_n - n)/n$ hinaus läuft. $\diamond$

**Beispiel 7.2.3 (Raten des Bereichs von Zufallszahlen).** Ein Showmaster produziert mit einer Maschine Zufallszahlen, die im Intervall $[0,\vartheta]$ gleichförmig verteilt sind, wobei der Parameter $\vartheta > 0$ vom Showmaster geheim eingestellt worden ist. Man soll nun auf Grund der Beobachtung von $n$ zufällig von der Maschine ausgegebenen Werten $X_1,\dots,X_n$ den Parameter $\tau(\vartheta)=\vartheta$ schätzen.

Unter der Annahme, dass die $n$ Ergebnisse der Maschine unabhängig sind, bietet sich also das statistische Produktmodell $((0,\infty)^n,(\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ an, wobei $\Theta=(0,\infty)$, und $\mathbb{P}_\vartheta$ ist die gleichförmige Verteilung auf dem Intervall $[0,\vartheta]$; siehe Beispiel 5.3.1.

Eine Möglichkeit ist, sich an das Gesetz der Großen Zahlen, Satz 6.1.4, zu erinnern und zu hoffen, dass schon für das in der Show konzedierte $n$ der Mittelwert der Ergebnisse nahe beim Erwartungswert $\mathbb{E}(\mathbb{P}_\vartheta)=\vartheta/2$ liegt. Der Schätzer $T_1(n)=\frac{2}{n}\sum_{i=1}^n X_i$ ist also plausibel.

Eine zweite Möglichkeit ist das beobachtete Maximum $T_2(n)=\max\{X_1,\dots,X_n\}$, denn zwar liegt man mit dieser Schätzung immer unter dem wahren Wert, aber für große $n$ sollte man nahe daran liegen. In der Tat gilt für alle $\varepsilon>0$

$$\mathbb{P}_\vartheta^{\otimes n}\big(|T_2(n)-\vartheta|\ge\varepsilon\big)=\mathbb{P}_\vartheta^{\otimes n}\big(T_2(n)\le\vartheta-\varepsilon\big)=\mathbb{P}_\vartheta^{\otimes n}\big(X_1\le\vartheta-\varepsilon,\dots,X_n\le\vartheta-\varepsilon\big)=\big(\tfrac{\vartheta-\varepsilon}{\vartheta}\big)^n\overset{n\to\infty}{\longrightarrow}0.$$

Also konvergieren beide Schätzer $T_1(n)$ und $T_2(n)$ in Wahrscheinlichkeit gegen den zu schätzenden Wert $\vartheta$ unter $\mathbb{P}_\vartheta^{\otimes n}$. Welcher der beiden Schätzer ist der bessere? Um dies zu beantworten, müssen wir zunächst Gütekriterien festlegen. Je nach Gütekriterium werden wir die Frage verschieden beantworten müssen.

*Erwartungstreue:* Der erste Schätzer ist *erwartungstreu* in dem Sinne, dass $\mathbb{E}_\vartheta^{\otimes n}(T_1(n))=\vartheta$ für alle $n\in\mathbb{N}$ und $\vartheta>0$. Der zweite ist 'asymptotisch' erwartungstreu, denn nach Lemma 5.3.4 gilt $\mathbb{E}_\vartheta^{\otimes n}(T_2(n))=\frac{n}{n+1}\vartheta$. Das bringt uns auf die Idee, $T_2(n)$ zu verbessern, indem wir statt seiner den Schätzer $T_3(n)=\frac{n+1}{n}T_2(n)$ betrachten, der ja erwartungstreu ist.

*Minimale Varianz:* Es ist sicher eine gute Eigenschaft eines Schätzers, eine geringe Varianz zu haben. Die Varianz des ersten Schätzers ist gleich $\mathbb{V}_\vartheta^{\otimes n}(T_1(n))=\frac{\vartheta^2}{3n}$ (siehe Beispiel 5.3.1), und die des zweiten errechnet man leicht als Übungsaufgabe zu $\mathbb{V}_\vartheta^{\otimes n}(T_2(n))=\frac{n\vartheta^2}{(n+1)^2(n+2)}$, und die seiner erwartungstreuen Modifikation zu $\mathbb{V}_\vartheta^{\otimes n}(T_3(n))=\frac{\vartheta^2}{n(n+2)}$. Also sind für große $n$ die letzten beiden Schätzer im Sinne der Varianzminimierung dem ersten hoch überlegen. $T_2(n)$ streut zwar etwas weniger als $T_3(n)$, aber um den leicht falschen Wert $\frac{n}{n+1}\vartheta$ herum. Seine mittlere quadratische Abweichung von $\vartheta$ ist $\mathbb{E}_\vartheta^{\otimes n}((T_2(n)-\vartheta)^2)=\frac{2\vartheta^2}{(n+1)(n+2)}$, also etwa doppelt so groß

wie die von $T_3(n)$. Im Sinne der mittleren quadratischen Abweichung von der zu schätzenden Größe ist also $T_3(n)$ der beste Schätzer von den drei betrachteten. ◇

Man kann also die Qualität von Schätzern nach verschiedenen Kriterien messen. In den voran gegangenen Beispielen haben wir die Kriterien Maximum Likelihood, asymptotische Konvergenz gegen den zu schätzenden Parameter (man nennt dies *Konsistenz*), Erwartungstreue, minimale Varianz und minimale mittlere quadratische Abweichung kennen gelernt. In den nächsten Abschnitten behandeln wir diese Kriterien einzeln systematisch.

## 7.3 Das Maximum-Likelihood-Prinzip

In diesem Abschnitt behandeln wir das Prinzip, das bei der Schätzung des Fischbestandes im Teich angewendet wurde. Abstrakt lässt sich dieses Prinzip wie folgt zusammen fassen: Unter einer geeigneten Familie von Modellen wählen wir dasjenige aus, das unserer beobachteten Messung die höchste Wahrscheinlichkeit zuordnet. Dem liegt die Idee zu Grunde, dass die Messung 'typisch' sein sollte, d. h. repräsentativ für den 'wahren' Mechanismus, und kein 'statistischer Ausreißer'.

**Definition 7.3.1 (Maximum-Likelihood-Schätzer).** *Sei $(\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell.*

*(i) Die Abbildung*

$$\varrho\colon \mathfrak{X}\times\Theta\to[0,\infty),\qquad \varrho(x,\vartheta)=\varrho_\vartheta(x)=\begin{cases}\mathbb{P}_\vartheta(\{x\}) & \text{falls } \mathfrak{X} \text{ diskret,}\\ \frac{\mathrm{d}\mathbb{P}_\vartheta(x)}{\mathrm{d}x} & \text{falls } \mathfrak{X} \text{ kontinuierlich,}\end{cases}$$

*heißt* Likelihood-Funktion *oder* Plausibilitätsfunktion. *Die Abbildung $\varrho_x(\cdot)=\varrho(x,\cdot)\colon\Theta\to[0,\infty)$ nennt man die* Likelihood-Funktion zum Beobachtungswert $x\in\mathfrak{X}$.

*(ii) Ein Schätzer $T\colon\mathfrak{X}\to\Theta$ für $\tau(\vartheta)=\vartheta$ heißt ein* Maximum-Likelihood-Schätzer*, falls*

$$\varrho(x,T(x))=\max_{\vartheta\in\Theta}\varrho(x,\vartheta),\qquad x\in\mathfrak{X}.$$

Mit $\frac{\mathrm{d}\mathbb{P}_\vartheta(x)}{\mathrm{d}x}$ bezeichnen wir im nichtdiskreten Fall eine Dichte des Wahrscheinlichkeitsmaßes $\mathbb{P}_\vartheta$, deren Existenz wir ja immer voraus setzen. Die Likelihood-Funktion $\varrho$ ist also als Funktion von $x$ eine Dichte von $\mathbb{P}_\vartheta$, und als Funktion von $\vartheta$ wird sie optimiert, um einen Maximum-Likelihood-Schätzer zu ermitteln.

**Beispiel 7.3.2 (Reißnagel).** Ein auf den Boden geworfener Reißnagel fällt mit unbekannter Wahrscheinlichkeit $\vartheta\in[0,1]$ auf die Spitze, ansonsten auf den Rücken. Wir werfen $n$ Mal und sehen, dass der Reißnagel genau $x$ Mal auf der Spitze zu liegen kommt. Wir wollen $\vartheta$ schätzen.

Wir gehen natürlich von Unabhängigkeit der Experimente aus und betrachten das Binomialmodell $(\{0,\dots,n\},(\mathrm{Bi}_{n,\vartheta})_{\vartheta\in\Theta})$ mit Parametermenge $\Theta=[0,1]$ und Likelihood-Funktion $\varrho_x(\vartheta)=\mathrm{Bi}_{n,\vartheta}(x)=\binom{n}{x}\vartheta^x(1-\vartheta)^{n-x}$. Mit der Log-Likelihood-Funktion $\log\varrho_x$ lässt sich besser rechnen:

$$\frac{\mathrm{d}}{\mathrm{d}\vartheta}\log\varrho_x(\vartheta)=\frac{x}{\vartheta}-\frac{n-x}{1-\vartheta},$$

und dies ist fallend in $\vartheta$ und nimmt den Wert Null genau in $\vartheta = x/n$ an. Dies bedeutet, dass $T = x/n$ (also der intuitiv naheliegendste Schätzer) der Maximum-Likelihood-Schätzer für dieses Modell ist. ◇

**Beispiel 7.3.3 (Bereich von Zufallszahlen).** Die Likelihood-Funktion im Beispiel 7.2.3 ist gegeben als

$$\varrho_x(\vartheta) = \begin{cases} \vartheta^{-n}, & \text{falls } x_1, \dots, x_n \leq \vartheta, \\ 0 & \text{sonst,} \end{cases}$$

wobei $x = (x_1, \dots, x_n)$. Der Schätzer $T_2(n) = \max\{X_1, \dots, X_n\}$ ist also genau der Maximum-Likelihood-Schätzer, denn für gegebene $x_1, \dots, x_n$ ist $\max\{x_1, \dots, x_n\}$ die kleinste Zahl $\vartheta$ mit $x_1 \leq \vartheta, \dots, x_n \leq \vartheta$. ◇

Ein wichtiges statistisches Modell mit zwei Parametern ist das folgende.

**Satz 7.3.4 (Maximum-Likelihood-Schätzer im Gaußmodell).** *Für $n \in \mathbb{N}$ betrachten wir das Produkt-Gauß-Modell $(\mathbb{R}^n, (\mathcal{N}(\mu, \sigma)^{\otimes n})_{\mu \in \mathbb{R}, \sigma^2 \in (0,\infty)})$, wobei $\mathcal{N}(\mu, \sigma)$ die in Beispiel 5.3.7 behandelte Gaußverteilung mit Erwartungswert $\mu$ und Varianz $\sigma^2$ ist. Dann ist der Maximum-Likelihood-Schätzer für $\tau(\mu, \sigma^2) = (\mu, \sigma^2)$ gegeben durch $T = (M, V)$, wobei*

$$M = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{und} \qquad V = \frac{1}{n} \sum_{i=1}^{n} (X_i - M)^2.$$

**Beweis.** Übungsaufgabe. (Eventuell ist Lemma 3.4.7 hilfreich.) □

Man nennt auch $M$ den *empirischen* Mittelwert und $V$ die *empirische* Varianz, manchmal auch den *Stichprobenmittelwert* bzw. die *Stichprobenvarianz* der Zufallsgrößen $X_1, \dots, X_n$.

## 7.4 Erwartungstreue und quadratischer Fehler

**Definition 7.4.1 (erwartungstreu, Bias).** *Es seien $(\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell und $\tau \colon \Theta \to \mathbb{R}$ eine reelle Kenngröße. Ein Schätzer $T \colon \mathfrak{X} \to \mathbb{R}$ heißt* erwartungstreu, *falls $\mathbb{E}_\vartheta(T) = \tau(\vartheta)$ für alle $\vartheta \in \Theta$ gilt. Die Differenz $\mathbb{B}_\vartheta(T) = \mathbb{E}_\vartheta(T) - \tau(\vartheta)$ heißt der* Bias *oder der* systematische Fehler *von $T$.*

In Beispiel 7.2.3 (siehe auch Beispiel 7.3.3) haben wir gesehen, dass Erwartungstreue und das Maximum-Likelihood-Prinzip nicht immer mit einander vereinbar sind: Der Maximum-Likelihood-Schätzer $T_2(n)$ ist nicht erwartungstreu (aber immerhin asymptotisch erwartungstreu). Auch der Schätzer der Varianz aus Satz 7.3.4 muss korrigiert werden, damit er erwartungstreu wird:

**Satz 7.4.2 (Erwartungstreue Schätzung von Erwartungswert und Varianz).** *Es sei $n \in \mathbb{N} \setminus \{1\}$ und $(\mathbb{R}^n, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta \in \Theta})$ ein reelles $n$-faches Produktmodell. Für jedes $\vartheta \in \Theta$ seien der Erwartungswert $m(\vartheta) = \mathbb{E}(\mathbb{P}_\vartheta)$ und die Varianz $v(\vartheta) = \mathbb{V}(\mathbb{P}_\vartheta)$ definiert. Dann sind der Stichprobenmittelwert und die korrigierte Stichprobenvarianz,*

$$M = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{und} \qquad V^* = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M)^2,$$

*erwartungstreue Schätzer für $m$ und $v$.*

**Beweis.** Mit Hilfe der Linearität des Erwartungswertes sieht man problemlos, dass $\mathbb{E}_\vartheta^{\otimes n}(M) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\vartheta(X_i) = m(\vartheta)$. Den Erwartungswert von $V = \frac{n-1}{n}V^*$ berechnen wir unter Benutzung von $\mathbb{E}_\vartheta^{\otimes n}(X_i - M) = 0$ und mit Hilfe des Satzes 3.5.3 von Bienaymé wie folgt:

$$\begin{aligned}\mathbb{E}_\vartheta^{\otimes n}(V) &= \frac{1}{n}\sum_{i=1}^n \mathbb{V}_\vartheta^{\otimes n}(X_i - M) = \mathbb{V}_\vartheta^{\otimes n}(X_1 - M) = \mathbb{V}_\vartheta^{\otimes n}\Big(\frac{n-1}{n}X_1 - \frac{1}{n}\sum_{i=2}^n X_i\Big)\\ &= \Big[\Big(\frac{n-1}{n}\Big)^2 + (n-1)\frac{1}{n^2}\Big]v(\vartheta) = \frac{n-1}{n}v(\vartheta),\end{aligned}$$

woraus die Behauptung folgt. □

Ein brauchbares Maß für die Qualität von Schätzern ist das folgende.

**Definition 7.4.3 (mittlerer quadratischer Fehler).** *Es seien $(\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\tau\colon \Theta \to \mathbb{R}$ eine reelle Kenngröße. Der* mittlere quadratische Fehler *eines Schätzers $T\colon \mathfrak{X} \to \mathbb{R}$ für $\tau$ ist die Größe*

$$\mathbb{F}_\vartheta(T) = \mathbb{E}_\vartheta\big((T - \tau(\vartheta))^2\big).$$

Man errechnet leicht, dass

$$\mathbb{F}_\vartheta(T) = \mathbb{V}_\vartheta(T) + \mathbb{B}_\vartheta(T)^2.$$

Um den quadratischen Fehler eines Schätzers möglichst klein zu halten, muss man also die Summe aus der Varianz und dem Quadrat des Bias klein halten. Dabei kann es zweckmäßig sein, einen Bias zuzulassen, wie das folgende Beispiel zeigt.

**Beispiel 7.4.4 (Ein guter Schätzer mit Bias).** Wir betrachten wie in Beispiel 7.3.2 das Binomialmodell $(\{0,\dots,n\}, (\mathrm{Bi}_{n,\vartheta})_{\vartheta\in\Theta})$ mit Parametermenge $\Theta = [0,1]$ und Likelihood-Funktion $\varrho_x(\vartheta) = \mathrm{Bi}_{n,\vartheta}(x) = \binom{n}{x}\vartheta^x(1-\vartheta)^{n-x}$. Für $\tau(\vartheta) = \vartheta$ betrachten wir die beiden Schätzer $T(x) = x/n$ und $S(x) = (x+1)/(n+2)$. Wir wissen aus Beispiel 7.3.2, dass $T$ erwartungstreu ist, und man sieht leicht, dass $S$ nicht erwartungstreu ist. Andererseits zeigt man als Übungsaufgabe, dass die mittleren quadratischen Fehler von $T$ und $S$ sich errechnen zu

$$\mathbb{F}_\vartheta(T) = \frac{\vartheta(1-\vartheta)}{n} \qquad \text{und} \qquad \mathbb{F}_\vartheta(S) = \frac{n\vartheta(1-\vartheta) + (1-2\vartheta)^2}{(n+2)^2}.$$

Insbesondere gilt $\mathbb{F}_\vartheta(S) \le \mathbb{F}_\vartheta(T)$ für alle $\vartheta$ mit $|\vartheta - \frac{1}{2}| \le 2^{-3/2}$, d. h. $S$ hat für zentrale Werte des Parameters einen kleineren mittleren quadratischen Fehler als $T$. ◇

## 7.5 Varianzminimierende Schätzer

Wir betrachten nun Schätzer, die erwartungstreu sind und unter allen erwartungstreuen Schätzern optimal sind in dem Sinne, dass sie am wenigsten streuen.

**Definition 7.5.1 (varianzminimierend).** *Sei $(\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell. Ein erwartungstreuer Schätzer $T$ für eine reelle Kenngröße $\tau(\vartheta)$ heißt* varianzminimierend *oder* (gleichmäßig) bester Schätzer*, wenn für jeden erwartungstreuen Schätzer $S$ gilt*

$$\mathbb{V}_\vartheta(T) \le \mathbb{V}_\vartheta(S), \qquad \vartheta \in \Theta.$$

Wir werden uns im Folgenden auf einparametrige Modelle beschränken und das Optimierungsproblem der Varianzminimierung explizit lösen, zunächst aber nur für Modelle, die die folgende Regularitätsannahme erfüllen.

**Definition 7.5.2 (reguläres Modell, Fisher-Information).** *Ein einparametriges statistisches Modell $(\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ heißt* regulär*, wenn $\Theta$ ein offenes Intervall in $\mathbb{R}$ ist und folgende zwei Bedingungen erfüllt sind:*

- *Die Likelihood-Funktion $\varrho$ ist auf $\mathfrak{X}\times\Theta$ strikt positiv und nach $\vartheta$ stetig differenzierbar. Insbesondere existiert die* Scorefunktion

$$U_\vartheta(x) = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\log\varrho(x,\vartheta) = \frac{\varrho_x'(\vartheta)}{\varrho_x(\vartheta)}.$$

- *Für jedes $\vartheta\in\Theta$ existiert die Varianz $I(\vartheta) = \mathbb{V}_\vartheta(U_\vartheta)$ und ist nicht 0, und es gilt die Vertauschungsrelation*

$$\int_{\mathfrak{X}} \frac{\mathrm{d}}{\mathrm{d}\vartheta}\varrho(x,\vartheta)\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\int \varrho(x,\vartheta)\,\mathrm{d}x.$$

  *(Für diskretes $\mathfrak{X}$ ist wie üblich das Integral durch eine Summe zu ersetzen.)*

*Die Funktion $I\colon\Theta\to[0,\infty)$ heißt die* Fisher-Information *des Modells.*

Die Vertauschungsrelation ist erfüllt, wenn jedes $\vartheta_0\in\Theta$ eine Umgebung $N(\vartheta_0)$ besitzt mit

$$\int_{\mathfrak{X}} \sup_{\vartheta\in N(\vartheta_0)} \Big|\frac{\mathrm{d}}{\mathrm{d}\vartheta}\varrho(x,\vartheta)\Big|\,\mathrm{d}x < \infty.$$

Wenn die Vertauschungsrelation erfüllt ist, so ergibt sich

$$\mathbb{E}_\vartheta(U_\vartheta) = \int \frac{\mathrm{d}}{\mathrm{d}\vartheta}\varrho(x,\vartheta)\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\int\varrho(x,\vartheta)\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\vartheta}1 = 0,$$

also ist die Scorefunktion bezüglich $\mathbb{P}_\vartheta$ zentriert. Insbesondere gilt

$$I(\vartheta) = \mathbb{E}_\vartheta(U_\vartheta^2) = \int \frac{\varrho_x'(\vartheta)^2}{\varrho_x(\vartheta)}\,\mathrm{d}x.$$

Große Werte von $I(\vartheta)$ bedeuten also große Änderungen der Dichte $\varrho(\cdot,\vartheta)$ im Parameter $\vartheta$, also sollte man besonders gut die Einflüsse verschiedener Werte von $\vartheta$ von einander unterscheiden können. Falls etwa $I(\vartheta) = 0$ für alle $\vartheta$ in einem Intervall, so kann man die Einflüsse der Parameter in diesem Intervall überhaupt nicht von einander unterscheiden.

**Lemma 7.5.3 (Additivität der Fisher-Information).** *Wenn $\mathcal{M} = (\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein reguläres Modell mit Fisher-Information $I$ ist, so besitzt das $n$-fache Produktmodell $\mathcal{M}^{\otimes n} = (\mathfrak{X}^n, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ die Fisher-Information $nI$.*

**Beweis.** Auf Grund des Produktsatzes für Dichten unabhängiger Zufallsgrößen (siehe Lemma 5.2.4) ist die Likelihood-Funktion $\varrho^{\otimes n}$ von $\mathcal{M}^{\otimes n}$ gegeben durch $\varrho^{\otimes n}(x,\vartheta) = \prod_{i=1}^n \varrho(x_i,\vartheta)$ für $x = (x_1,\dots,x_n)$. Also ist die zugehörige Scorefunktion gleich

$$U_\vartheta^{\otimes n}(x) = \frac{\frac{\mathrm{d}}{\mathrm{d}\vartheta}\prod_{i=1}^n \varrho(x_i,\vartheta)}{\varrho^{\otimes n}(x,\vartheta)} = \sum_{i=1}^n \frac{\frac{\mathrm{d}}{\mathrm{d}\vartheta}\varrho(x_i,\vartheta)}{\varrho(x_i,\vartheta)} = \sum_{i=1}^n U_\vartheta(x_i).$$

Wegen der Unabhängigkeit der Komponenten kann man den Satz von Bienaymé anwenden und erhält die Fisher-Information $I^{\otimes n}$ von $\mathcal{M}^{\otimes n}$ als

$$I^{\otimes n}(\vartheta) = \mathbb{V}_\vartheta^{\otimes n}(U_\vartheta^{\otimes n}) = \sum_{i=1}^n \mathbb{V}_\vartheta(U_\vartheta) = nI(\vartheta).$$

□

Die Bedeutung der Fisher-Information wird offenbar in dem folgenden Satz. Wir nennen einen erwartungstreuen Schätzer $T$ für $\tau(\vartheta) = \vartheta$ *regulär*, wenn für jedes $\vartheta$ die Vertauschungsrelation

$$\int_{\mathfrak{X}} T(x)\frac{\mathrm{d}}{\mathrm{d}\vartheta}\varrho(x,\vartheta)\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\int T(x)\varrho(x,\vartheta)\,\mathrm{d}x$$

erfüllt ist.

**Satz 7.5.4 (Informationsungleichung).** *Gegeben seien ein reguläres statistisches Modell $(\mathfrak{X},(\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$, eine zu schätzende stetig differenzierbare Funktion $\tau\colon\Theta\to\mathbb{R}$ mit $\tau'\neq 0$ und ein regulärer erwartungstreuer Schätzer $T$ für $\tau$. Dann gilt*

$$\mathbb{V}_\vartheta(T) \geq \frac{\tau'(\vartheta)^2}{I(\vartheta)}, \qquad \vartheta\in\Theta. \tag{7.5.1}$$

*Gleichheit für alle $\vartheta\in\Theta$ gilt genau dann, wenn*

$$T - \tau(\vartheta) = \frac{\tau'(\vartheta)U_\vartheta}{I(\vartheta)}, \qquad \vartheta\in\Theta,$$

*d. h. wenn die Likelihood-Funktion die Gestalt*

$$\varrho(x,\vartheta) = e^{a(\vartheta)T(x)-b(\vartheta)}\,h(x) \tag{7.5.2}$$

*besitzt mit einer Stammfunktion $a\colon\Theta\to\mathbb{R}$ von $I/\tau'$, einer beliebigen (genügend integrierbaren) Funktion $h\colon\mathfrak{X}\to(0,\infty)$ und einer Normierungsfunktion $b(\vartheta) = \log\int_{\mathfrak{X}} e^{a(\vartheta)T(x)}h(x)\,\mathrm{d}x$.*

**Beweis.** Wir berechnen zunächst die Kovarianz von $T$ und $U_\vartheta$ und benutzen dabei die Zentriertheit von $U_\vartheta$ sowie die Regularität und die Erwartungstreue von $T$:

$$\mathrm{cov}_\vartheta(T,U_\vartheta) = \mathbb{E}_\vartheta(TU_\vartheta) = \int_{\mathfrak{X}} T(x)\frac{\mathrm{d}}{\mathrm{d}\vartheta}\varrho(x,\vartheta)\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\int T(x)\varrho(x,\vartheta)\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\mathbb{E}_\vartheta(T) = \tau'(\vartheta).$$

Hieraus ergibt sich mit $c(\vartheta) = \tau'(\vartheta)/I(\vartheta)$:

$$0 \leq \mathbb{V}_\vartheta\big(T - c(\vartheta)U_\vartheta\big) = \mathbb{V}_\vartheta(T) + c(\vartheta)^2\mathbb{V}_\vartheta(U_\vartheta) - 2c(\vartheta)\mathrm{cov}_\vartheta(T,U_\vartheta) = \mathbb{V}_\vartheta(T) - \frac{\tau'(\vartheta)^2}{I(\vartheta)},$$

also die Informationsungleichung in (7.5.1). Gleichheit gilt genau dann, wenn die Zufallsgröße $T - c(\vartheta)U_\vartheta$ bezüglich $\mathbb{P}_\vartheta$ konstant ist, und diese Konstante muss natürlich gleich ihrem Erwartungswert $\tau(\vartheta)$ sein. Da (im stetigen Fall) $\mathbb{P}_\vartheta$ die positive Dichte $\varrho(\cdot,\vartheta)$ besitzt bzw. (im diskreten Fall) jeden Punkt $x \in \mathfrak{X}$ mit dem positiven Gewicht $\varrho(x,\vartheta)$ versieht, folgt

$$\int_{\mathfrak{X}} \mathbb{1}_{\{x\colon T(x)-\tau(\vartheta)\neq c(\vartheta)U_\vartheta(x)\}}(x)\,\mathrm{d}x = 0$$

für jedes $\vartheta \in \Theta$. Im diskreten Fall folgt sofort, dass $T(x) - \tau(\vartheta) = c(\vartheta)U_\vartheta(x)$ für alle $x \in \mathfrak{X}$ und alle $\vartheta \in \Theta$. Im stetigen Fall benutzt man ein wenig Maßtheorie, um die Gleichung $T(x) - \tau(\vartheta) = c(\vartheta)U_\vartheta(x)$ für alle $x \in \mathfrak{X}$ und alle $\vartheta \in \Theta$ zu folgern. Diese Gleichung ist äquivalent zu

$$\frac{\mathrm{d}}{\mathrm{d}\vartheta}\log\varrho(x,\vartheta) = \big(T(x) - \tau(\vartheta)\big)\frac{I(\vartheta)}{\tau'(\vartheta)},$$

also folgt durch unbestimmte Integration die Gleichung (7.5.2) mit $b(\vartheta) = -\int \tau(\vartheta)I(\vartheta)/\tau'(\vartheta)\,\mathrm{d}\vartheta$. Hierbei ist $\log h(x)$ die Integrationskonstante, und $b(\vartheta)$ wird festgelegt durch die Tatsache, dass $\varrho(\cdot,\vartheta)$ eine Wahrscheinlichkeitsdichte ist. Also hat $\varrho$ die angegebene Form.

Die umgekehrte Richtung ist evident. □

Wir nennen einen regulären erwartungstreuen Schätzer $T$, der die Gleichheit in der Informationsungleichung (7.5.1) erfüllt, *Cramér-Rao-effizient*. Wir halten folgende Folgerungen aus Satz 7.5.4 und Lemma 7.5.3 fest:

(i) Bei $n$-facher unabhängiger Wiederholung eines regulären Experiments ist die Varianz eines erwartungstreuen Schätzers für $\tau$ mindestens von der Größenordnung $1/n$. (Im – nicht regulären – Modell in Beispiel 7.2.3 (Zufallszahlen) hatten wir Schätzer, deren Varianzen wie $1/n^2$ fallen.)

(ii) Ein Cramér-Rao-effizienter Schätzer existiert nur, wenn die Likelihood-Funktion $\varrho$ von der Form (7.5.2) ist. Dann ist er aber ein varianzminimierender Schätzer, zumindest in der Klasse aller regulären Schätzer für $\tau$.

Wenn man eine Likelihood-Funktion in der Gestalt (7.5.2) vorliegen hat, muss man nach Satz 7.5.4 prüfen, ob das Modell regulär ist und ob $a' = I/\tau'$ gilt, um schließen zu können, dass Gleichheit in der Informationsungleichung gilt. Der folgende Satz 7.5.6 nimmt uns diese Arbeit im Allgemeinen ab. Die Modelle, auf die man diesen Satz anwenden kann, stellen eine interessante Klasse dar:

**Definition 7.5.5 (exponentielles Modell).** *Es sei $\mathcal{M} = (\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit einem offenen Intervall $\Theta$. $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ heißt eine* exponentielle Familie *bezüglich einer Statistik $T\colon \mathfrak{X} \to \mathbb{R}$, wenn die Likelihood-Funktion die Gestalt* (7.5.2) *besitzt mit einer stetig differenzierbaren Funktion $a\colon \Theta \to \mathbb{R}$ mit $a' \neq 0$ und einer Funktion $h\colon \mathfrak{X} \to \mathbb{R}$, die über jedes Intervall Riemann-integrierbar ist. (Die Normierungsfunktion $b$ ist durch $a$ und $h$ eindeutig fest gelegt.)*

Man beachte also, dass die Regularität und die Beziehung $a' = I/\tau'$ nicht in der Definition enthalten ist; es ist ja nicht einmal die Rede von irgend einer zu schätzenden Kenngröße $\tau(\vartheta)$. Der folgende Satz aber sagt, dass diese beiden Eigenschaften fü jede exponentielle Familie erfüllt sind für eine geeignete Funktion $\tau$.

**Satz 7.5.6 (Eigenschaften exponentieller Familien).** [1] *Für jedes exponentielle Modell $\mathcal{M}$ gelten:*

*(i) $b$ ist stetig differenzierbar mit $b'(\vartheta) = a'(\vartheta)\mathbb{E}_\vartheta(T)$ für jedes $\vartheta \in \Theta$.*

*(ii) Jede Statistik $S\colon \mathfrak{X} \to \mathbb{R}$ mit existierenden Varianzen $\mathbb{V}_\vartheta(S)$ ist regulär. Insbesondere sind $\mathcal{M}$ und $T$ regulär, und $\tau(\vartheta) = \mathbb{E}_\vartheta(T)$ ist stetig differenzierbar mit $\tau'(\vartheta) = a'(\vartheta)\mathbb{V}_\vartheta(T) \neq 0$.*

*(iii) Für die Fisher-Information gilt $I(\vartheta) = a'(\vartheta)\tau'(\vartheta)$.*

**Beweisskizze.** Wir dürfen annehmen, dass $a(\vartheta) = \vartheta$, denn der allgemeine Fall ergibt sich durch Reparametrisierung und Anwendung der Kettenregel.

Man zeigt leicht, dass die Funktion $u(\vartheta) = e^{b(\vartheta)} = \int_{\mathfrak{X}} e^{\vartheta T(x)} h(x)\,\mathrm{d}x$ unendlich oft differenzierbar ist mit

$$u'(\vartheta) = \int_{\mathfrak{X}} T(x) e^{\vartheta T(x)} h(x)\,\mathrm{d}x = e^{b(\vartheta)} \int_{\mathfrak{X}} T(x)\varrho(x,\vartheta)\,\mathrm{d}x = u(\vartheta)\mathbb{E}_\vartheta(T)$$

und analog $u''(\vartheta) = u(\vartheta)\mathbb{E}_\vartheta(T^2)$. Daraus folgt für $b = \log u$, dass $b'(\vartheta) = \mathbb{E}_\vartheta(T) = \tau(\vartheta)$, also die Ausage in (i). Weiter hat man

$$\tau'(\vartheta) = b''(\vartheta) = \frac{u''(\vartheta)}{u(\vartheta)} - \Big(\frac{u'(\vartheta)}{u(\vartheta)}\Big)^2 = \mathbb{V}_\vartheta(T),$$

womit auch die letzte Aussage in (ii) gezeigt ist. Mit ein wenig Maßtheorie zeigt man, dass jede Statistik $S$ mit endlichen Varianzen regulär ist und die Beziehung $\mathbb{E}_\vartheta(SU_\vartheta) = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\mathbb{E}_\vartheta(S)$ erfüllt (dies ist ähnlich zum entsprechenden Beweisteil in Satz 7.5.4). Insbesondere ist $T$ regulär, und die Regularität von $\mathcal{M}$ ist die von $S \equiv 1$. Wegen

$$U_\vartheta(x) = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\log\varrho(x,\vartheta) = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\big(\vartheta T(x) - b(\vartheta) + \log h(x)\big) = T(x) - b'(\vartheta)$$

folgt $I(\vartheta) = \mathbb{V}_\vartheta(U_\vartheta) = \mathbb{V}_\vartheta(T) = \tau'(\vartheta)/a'(\vartheta)$, und dies beendet die Beweisskizze. □

**Bemerkung 7.5.7.** Nach Satz 7.5.6(iii) ist für jedes exponentielle Modell die zu Grunde liegende Statistik $T$ ein bester Schätzer für die Kenngröße $\tau(\vartheta) = \mathbb{E}_\vartheta(T)$, zunächst unter den regulären erwartungstreuen Schätzern. Allerdings ist nach Satz 7.5.6(ii) jeder erwartungstreue Schätzer mit endlichen Varianzen automatisch regulär, so dass $T$ sogar die kleinste Varianz unter *allen* erwartungstreuen Schätzern besitzt. Damit ist also $T$ varianzminimierend im Sinne der Definition 7.5.2. ◇

**Beispiel 7.5.8 (Poisson-Modell).** Die Likelihood-Funktion

$$\varrho(x,\vartheta) = e^{-\vartheta}\frac{\vartheta^x}{x!} = \exp\big[(\log\vartheta)x - \vartheta\big]\frac{1}{x!}, \qquad x \in \mathfrak{X} = \mathbb{N}_0, \vartheta \in \Theta = (0,\infty),$$

der Poisson-Verteilungen $\mathrm{Po}_\vartheta$ hat die Form (7.5.2) mit $T(x) = x$ und $a(\vartheta) = \log\vartheta$, also handelt es sich um eine exponentielle Familie. Zur Übung rechnen wir nach, dass $a' = I/\tau' = I$ gilt (was ja auch aus Satz 7.5.6 folgt):

$$\begin{aligned} I(\vartheta) &= \sum_{x=0}^{\infty} \frac{\varrho_x'(\vartheta)^2}{\varrho_x(\vartheta)} = \sum_{x=0}^{\infty} \frac{1}{x!} e^{\vartheta}\vartheta^{-x}\Big[-e^{-\vartheta}\vartheta^x + e^{-\vartheta}x\vartheta^{x-1}\Big]^2 = \sum_{x=0}^{\infty}\frac{\vartheta^x}{x!}e^{-\vartheta}\Big(\frac{x}{\vartheta} - 1\Big)^2 = \frac{1}{\vartheta^2}\mathbb{V}(\mathrm{Po}_\vartheta) \\ &= \frac{1}{\vartheta} = a'(\vartheta). \end{aligned}$$

[1] Diesem Satz und seinem Beweis (den wir nur skizzieren) liegt tatsächlich nicht der Riemann- sondern der Lebesgue-Integralbegriff zu Grunde.

$T$ ist ein erwartungstreuer Schätzer für $\tau(\vartheta) = \vartheta$, also sogar varianzminimierend. Insbesondere erhalten wir nochmals die Gleichung $\mathbb{V}(\mathrm{Po}_\vartheta) = \mathbb{V}_\vartheta(T) = 1/\frac{1}{\vartheta} = \vartheta$; siehe Beispiele 3.3.6 und 3.4.5. ◇

**Beispiel 7.5.9 (Binomialmodell).** Für festes $n \in \mathbb{N}$ bilden auf $\mathfrak{X} = \{0, \dots, n\}$ die Binomialverteilungen $\mathrm{Bi}_{n,\vartheta}$ mit $\vartheta \in (0,1)$ eine Familie, für die die Likelihood-Funktion $\varrho(x,\vartheta) = \binom{n}{x}\vartheta^x(1-\vartheta)^{n-x}$ die Form (7.5.2) besitzt: Wir setzen $T = x/n$, $a(\vartheta) = n\log\frac{\vartheta}{1-\vartheta}$, $b(\vartheta) = -n\log(1-\vartheta)$ und $h(x) = \binom{n}{x}$. $T$ ist ein erwartungstreuer Schätzer für $\tau(\vartheta) = \vartheta$, also insbesondere varianzminimierend, und es gilt

$$I(\vartheta) = a'(\vartheta) = n\Big(\frac{1}{\vartheta} + \frac{1}{1-\vartheta}\Big) = \frac{n}{\vartheta(1-\vartheta)},$$

also insbesondere $\mathbb{V}_\vartheta(T) = \tau'(\vartheta)^2/I(\vartheta) = \vartheta(1-\vartheta)/n$. ◇

**Beispiel 7.5.10 (Normalverteilungen).** *(a) Schätzung des Erwartungswerts.* Bei festem Varianzparameter $v > 0$ hat die Familie der Normalverteilungen $\{\mathcal{N}(\vartheta, v)\colon \vartheta \in \mathbb{R}\}$ auf $\mathfrak{X} = \mathbb{R}$ die Likelihood-Funktion

$$\varrho(x,\vartheta) = (2\pi v^2)^{-1/2}\exp\Big[-\frac{(x-\vartheta)^2}{2v^2}\Big],$$

die von der Form (7.5.2) ist mit $T(x) = x$, $a(\vartheta) = \vartheta/v^2$, $b(\vartheta) = \vartheta^2/(2v^2) + \frac{1}{2}\log(2\pi v^2)$ und $h(x) = \exp[-x^2/(2v^2)]$. Da $T$ ein erwartungstreuer Schätzer für $\tau(\vartheta) = \vartheta$ ist, ist $T$ varianzminimierend, und es gilt $\mathbb{V}_\vartheta(T) = v^2$.

*(b) Schätzung der Varianz.* Bei festem Erwartungswert $m \in \mathbb{R}$ hat die Familie $\{\mathcal{N}(m,\vartheta)\colon \vartheta \in (0,\infty)\}$ der zugehörigen Normalverteilungen auf $\mathfrak{X} = \mathbb{R}$ die Likelihood-Funktion

$$\varrho(x,\vartheta) = \exp\Big[-\frac{T(x)}{2\vartheta^2} - \frac{1}{2}\log(2\pi\vartheta^2)\Big]$$

mit $T(x) = (x-m)^2$. $T$ ist ein erwartungstreuer Schätzer für die Varianz $\tau(\vartheta) = \vartheta^2$, also auch varianzminimierend. ◇

## 7.6 Konsistenz

Wenn man eine Schätzung auf oft unabhängig wiederholten Experimenten basieren lässt, so sollte man erwarten dürfen, dass die Schätzung immer genauer wird, also immer 'näher' liegt am 'wahren' Wert der zu schätzenden Größe. Diese Eigenschaft nennt man die Konsistenz des Schätzers, genauer: der benutzten Folge von Schätzern. Wir geben eine allgemeine Definition der Konsistenz, geben ein einfaches Kriterium in einem Spezialfall und diskutieren ein paar Beispiele.

**Definition 7.6.1 (konsistente Schätzer).** *Für jedes $n \in \mathbb{N}$ sei $(\mathfrak{X}_n, (\mathbb{P}_{\vartheta,n})_{\vartheta\in\Theta})$ ein statistisches Modell mit gemeinsamer Parametermenge $\Theta$. Ferner sei $\tau\colon \Theta \to \mathbb{R}$ eine reelle Kenngröße und $T_n\colon \mathfrak{X}_n \to \mathbb{R}$ ein Schätzer für $\tau$ im $n$-ten Modell. Die Schätzerfolge $(T_n)_{n\in\mathbb{N}}$ heißt* konsistent, *falls für jedes $\vartheta \in \Theta$ gilt:*

$$\lim_{n\to\infty} \mathbb{P}_{\vartheta,n}(|T_n - \tau(\vartheta)| > \varepsilon) = 0, \qquad \varepsilon > 0,$$

*wenn also für jedes $\vartheta \in \Theta$ der Schätzer $T_n$ in $\mathbb{P}_{\vartheta,n}$-Wahrscheinlichkeit gegen $\tau(\vartheta)$ konvergiert (siehe Definition 6.1.3).*

**Beispiel 7.6.2.** Wir erinnern an das Beispiel 7.2.3, in dem in einer Quizshow der Bereich von gleichverteilten Zufallszahlen geraten werden sollte. Wir zeigten dort explizit, dass die beiden Schätzer $T_1(n) = \frac{2}{n}\sum_{i=1}^n X_i$ und $T_2(n) = \max\{X_1, \dots, X_n\}$ konsistent sind. Während die Konsistenz von $T_1(n)$ ein Spezialfall des folgenden Satzes 7.6.4 ist, ist $T_2(n)$ ein Beispiel für einen konsistenten Maximum-Likelihood-Schätzer. ◇

Es ist klar, dass sich die Konsistenz in natürlicher Weise aus dem Schwachen Gesetz der Großen Zahlen ergibt, wenn es sich um den empirischen Mittelwert bei oftmaliger Wiederholung von unabhängigen Experimenten handelt. Die Konsistenz der beiden erwartungstreuen Schätzer aus Satz 7.4.2 ist also nahezu evident, wie wir gleich sehen werden. Wir schicken ein elementares allgemeines Lemma über Konvergenz in Wahrscheinlichkeit voraus:

**Lemma 7.6.3.** *Es sei $(\mathbb{P}_n)_{n\in\mathbb{N}}$ eine Folge von Wahrscheinlichkeitsmaßen auf $\mathfrak{X}$, und es seien $(X_n)_n$ und $(Y_n)_n$ zwei Folgen von Zufallsgrößen mit $X_n \overset{\mathbb{P}_n}{\to} 0$ und $Y_n \overset{\mathbb{P}_n}{\to} 0$. Ferner sei $(a_n)_n$ eine beschränkte Folge reeller Zahlen. Dann gelten $X_n + Y_n \overset{\mathbb{P}_n}{\to} 0$ und $a_n X_n \overset{\mathbb{P}_n}{\to} 0$.*

**Beweis.** Übungsaufgabe. □

Nun folgt die angekündigte Konsistenz der Schätzer aus Satz 7.4.2.

**Satz 7.6.4 (Konsistenz von empirischem Mittel und Varianz).** *Es sei $\mathcal{M} = (\mathbb{R}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, und für jedes $\vartheta \in \Theta$ existiere sowohl der Erwartungswert $m(\vartheta) = \mathbb{E}(\mathbb{P}_\vartheta)$ als auch die Varianz $v(\vartheta) = \mathbb{V}(\mathbb{P}_\vartheta)$. Wir setzen voraus, dass auch das vierte Moment von $\mathbb{P}_\vartheta$ endlich ist, also $\int x^4 \mathbb{P}_\vartheta(\mathrm{d}x) < \infty$. Für jedes $n \in \mathbb{N}$ betrachten wir auf dem $n$-fachen Produktmodell $\mathcal{M}^{\otimes n} = (\mathbb{R}^n, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ die Schätzer*

$$M_n = \frac{1}{n}\sum_{i=1}^n X_i \qquad \text{und} \qquad V_n^* = \frac{1}{n-1}\sum_{i=1}^n (X_i - M_n)^2$$

*für $m$ und $v$. (Wir erinnern, dass $X_i \colon \mathbb{R}^n \to \mathbb{R}$ die $i$-te Koordinate ist.) Dann sind die Folgen $(M_n)_n$ und $(V_n^*)_n$ konsistent.*

**Beweis.** Die Konsistenz von $(M_n)_n$, also die Konvergenz in Wahrscheinlichkeit von $M_n$ gegen $\mathbb{E}(\mathbb{P}_\vartheta)$ unter $\mathbb{P}_\vartheta^{\otimes n}$, folgt direkt aus dem Schwachen Gesetz der Großen Zahlen, Satz 6.1.4.

Auf $(V_n^*)_n$ können wir diesen Satz zwar nicht direkt anwenden, aber auf

$$\widetilde{V}_n = \frac{1}{n}\sum_{i=1}^n (X_i - m(\vartheta))^2.$$

Man errechnet leicht, dass $V_n = \widetilde{V}_n - (M_n - m(\vartheta))^2$, wobei $V_n = \frac{n-1}{n} V_n^*$. Wegen $\widetilde{V}_n \overset{\mathbb{P}_\vartheta^{\otimes n}}{\to} v(\vartheta)$ nach Satz 6.1.4 und $(M_n - m(\vartheta))^2 \overset{\mathbb{P}_\vartheta^{\otimes n}}{\to} 0$ folgt nun leicht mit Hilfe von Lemma 7.6.3, dass auch $V_n^* \overset{\mathbb{P}_\vartheta}{\to} v(\vartheta)$ gilt. □

Auch Maximum-Likelihood-Schätzer sind im Allgemeinen konsistent, worauf wir allerdings nicht eingehen wollen. Es folgen ein paar Beispiele von Schätzern, deren Konsistenz sich aus Satz 7.6.4 ergibt.

**Beispiel 7.6.5 (Poisson-Parameterschätzung).** Wir wollen die mittlere Anzahl von Versicherungsfällen bei einer Kraftfahrzeug-Versicherung pro Jahr schätzen. Diese Anzahl kann man als ungefähr Poisson-verteilt annehmen, also betrachten wir das $n$-fache Produktmodell $(\mathbb{N}_0^n, (\mathrm{Po}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ mit $\Theta = (0,\infty)$ und der Poisson-Verteilung $\mathrm{Po}_\vartheta$ mit Parameter $\vartheta$. Der Schätzer $T_n = M_n = \frac{1}{n}\sum_{i=1}^n X_i$ für den Parameter $\vartheta$, also den Erwartungswert von $\mathbb{P}_\vartheta$, ist also nach Satz 7.6.4 konsistent. ◇

**Beispiel 7.6.6 (Exponentialparameter-Schätzung).** Wir wollen die mittlere Wartezeit eines Kunden in einer Warteschlange (oder die Lebensdauer einer Glühbirne) schätzen. Wir dürfen solche Wartezeiten als exponential-verteilt annehmen, also wollen wir den Kehrwert des zugehörigen Parameters schätzen. Also betrachten wir das $n$-fache Produktmodell des Modells $((0,\infty), (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$, wobei $\mathbb{P}_\vartheta$ die Dichte $\varrho_\vartheta(x) = \vartheta e^{-\vartheta x}$ besitzt. Also liefert Satz 7.6.4 die Konsistenz des Schätzers $T_n = 1/M_n = n/\sum_{i=1}^n X_i$ für $\vartheta$, die auf $n$ unabhängigen Beobachtungen $X_1, \ldots, X_n$ basiert. ◇

# Kapitel 8

# Konfidenzbereiche

Ein Schätzer gibt nur einen groben Anhaltspunkt für den wahren Wert einer unbekannten zufälligen Größe, aber keine Aussage über die Zuverlässigkeit dieser Schätzung, d. h. über die Abweichung der Schätzung vom wahren Wert. Besser werden die Launen des Zufalls berücksichtigt, wenn man sogar einen ganzen (von der Beobachtung abhängigen) Bereich angibt, in dem die Größe mit gegebener Sicherheit liegt. Diese Bereiche nennt man Konfidenz- oder Vertrauensbereiche; sie werden in diesem Kapitel behandelt.

## 8.1 Definition

Bei der experimentellen Bestimmung der Wahrscheinlichkeit, dass ein Reißnagel auf die Spitze fällt (siehe Beispiel 7.3.2), kann der eine Experimentator auf Grund seiner 20 Versuche auf den Schätzwert $\vartheta = \frac{2}{5}$ kommen, ein anderer mit seinen 30 Würfen aber vielleicht auf $\vartheta = \frac{3}{10}$. Natürlich kann man dann von keiner der beiden Schätzungen sagen, sie seien 'richtig', nicht einmal, wenn irgend ein höheres Wesen uns verlässlich mitteilen würde, dass $\vartheta = \frac{2}{5}$ der wahre Wert sei, denn auch die erste Schätzung kam eben nur von einem Zufall her.

Seriöser ist eine Aussage, die Abweichungen zulässt und Irrtumswahrscheinlichkeiten angibt, etwa indem man sagt, mit einer Sicherheit von 95 Prozent liege $\vartheta$ im Intervall $[0.37, 0.45]$.

**Definition 8.1.1 (Konfidenzbereich).** *Es sei $(\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $\tau\colon \Theta \to \mathbb{R}$ eine Kenngröße und $\alpha \in (0,1)$ eine Fehlerschranke. Eine Familie $(C(x))_{x\in\mathfrak{X}}$ von Mengen $C(x) \subset \mathbb{R}$ heißt ein* Konfidenz- *oder* Vertrauensbereich *für $\tau$ zum* Irrtumsniveau *$\alpha$ (oder zum* Sicherheitsniveau *$1-\alpha$), falls für jedes $\vartheta \in \Theta$ gilt*

$$\mathbb{P}_\vartheta\big(x \in \mathfrak{X}\colon \tau(\vartheta) \in C(x)\big) \geq 1-\alpha. \tag{8.1.1}$$

*Falls jedes $C(x)$ ein Intervall ist, so spricht man auch von einem* Konfidenzintervall.

Diese Definition erfordert ein paar Kommentare:

**Bemerkung 8.1.2.** (i) Die Familie $(C(x))_{x\in\mathfrak{X}}$ ist also tatsächlich eine einzige Zufallsgröße, die ihre Werte in der Menge der Teilmengen von $\mathbb{R}$, also der Potenzmenge von $\mathbb{R}$, annimmt. Der Konfidenzbereich $C(x)$ wird aus dem Beobachtungsergebnis $x$ heraus konstruiert. Die

Bedingung lautet also, dass für jedes $\vartheta \in \Theta$ diese zufällige Menge $C(x)$ mit einer $\mathbb{P}_\vartheta$-Wahrscheinlichkeit nicht unter $1-\alpha$ die zu schätzende Kenngröße $\tau(\vartheta)$ enthält.

(ii) Die Bedingung ist trivialerweise erfüllt, wenn $C(x) = \mathbb{R}$ ist für jedes $x \in \mathfrak{X}$. Sinnvoll wird ein Konfidenzbereich allerdings nur, wenn die Mengen $C(x)$ möglichst *klein* gewählt werden. Ferner möchte man gerne eine möglichst kleine Fehlerschranke $\alpha$ haben, aber diese beiden Ziele streiten wider einander: Je kleiner $\alpha$ ist, desto größer wird man die Mengen $C(x)$ wählen müssen. Das Ausbalancieren dieser beiden Ziele ist eine komplizierte Aufgabe, die man im jeweiligen konkreten Fall vornehmen muss.

(iii) Man hüte sich vor dem folgenden Missverständnis. Wenn beim Reißnagelexperiment etwa der Schätzwert $T = \frac{2}{5}$ erhalten wird und daher das Konfidenzintervall $(0.3, 0.5)$ zum Niveau 95 % angegeben wird, heißt das *nicht*, dass in 95 Prozent aller Fälle, in denen ein Schätzwert in $(0.3, 0.5)$ erhalten wird, auch der Wert von $\tau(\vartheta)$ in diesem Intervall liegt. Damit würde man $\vartheta$ wie eine Zufallsgröße behandeln, aber auf der Parametermenge $\Theta$ wird keinerlei Wahrscheinlichkeitsmaß betrachtet. Hingegen ist die Menge $C(x)$ zufällig, und korrekt ist die Formulierung, dass in 95 Prozent aller Fälle (egal, mit welchen $\mathbb{P}_\vartheta$ wir messen) dieses Zufallsintervall die zu schätzende Größe $\tau(\vartheta)$ enthält.

◇

## 8.2 Konstruktion

In diesem Abschnitt geben wir eine allgemeine Konstruktion von Konfidenzbereichen, allerdings nur für die Kenngröße $\tau(\vartheta) = \vartheta$. Der Ausgangspunkt ist die Betrachtung der zweidimensionalen Menge

$$\mathcal{C} = \big\{(x, \vartheta) \in \mathfrak{X} \times \Theta \colon \vartheta \in C(x)\big\}.$$

Wenn man $x \in \mathfrak{X}$ auf der waagerechten Achse abträgt und $\vartheta \in \Theta$ auf der senkrechten, dann ist $C(x)$ gerade der vertikale $x$-Schnitt durch $\mathcal{C}$, also bestimmt $\mathcal{C}$ die Familie $(C(x))_{x\in\mathfrak{X}}$ eindeutig. Für jedes $\vartheta \in \Theta$ ist das Ereignis

$$C_\vartheta = \big\{C(\cdot) \ni \vartheta\big\} = \big\{x \in \mathfrak{X} \colon \vartheta \in C(x)\big\} = \big\{x \in \mathfrak{X} \colon (x, \vartheta) \in \mathcal{C}\big\}$$

der horizontale $\vartheta$-Schnitt durch $\mathcal{C}$. Die Bedingung (8.1.1) ist äquivalent zu $\mathbb{P}_\vartheta(C_\vartheta) \geq 1-\alpha$. Das bedeutet, dass bei gegebenem $\alpha \in (0,1)$ für jedes $\vartheta \in \Theta$ eine (möglichst kleine) Menge $C_\vartheta$ definiert werden soll, sodass die Bedingung $\mathbb{P}_\vartheta(C_\vartheta) \geq 1-\alpha$ erfüllt ist. Dies machen wir im diskreten Fall, indem wir die Wahrscheinlichkeiten $\varrho_\vartheta(x)$ der Größe nach ordnen und, beginnend mit dem größten, so lange die zugehörigen $x$ in der Menge $C_\vartheta$ versammeln, bis die Summe ihrer Wahrscheinlichkeiten $\varrho_\vartheta(x)$ den Wert $1-\alpha$ übersteigt. Im stetigen Fall verfahren wir analog, indem wir Intervalle um die maximalen Werte der Dichten $\varrho_\vartheta(\cdot)$ geeignet legen, bis die Gesamtmasse dieser Intervalle den Wert $1-\alpha$ übersteigt. Die Vereinigung dieser Intervalle wählen wir dann als $C_\vartheta$. Auf diese Weise konstruieren wir eine Familie $(C_\vartheta)_{\vartheta\in\Theta}$ von horizontalen Schnitten, also auch eine Menge $\mathcal{C}$. Dann sind die vertikalen Schnitte $C(x)$ geeignete Konfidenzbereiche.

Eine kleine Zusammenfassung dieses Konstruktionsverfahrens lautet wie folgt. Es sei die Likelihood-Funktion $\varrho(x, \vartheta)$ gegeben sowie eine Fehlerschranke $\alpha \in (0,1)$.

**Konstruktion eines Konfidenzbereiches.**
(1) Für jedes $\vartheta \in \Theta$ bestimme man eine Menge $C_\vartheta$ der Gestalt

$$C_\vartheta = \{x \in \mathfrak{X} \colon \varrho_\vartheta(x) \geq c_\vartheta\},$$

wobei $c_\vartheta > 0$ so bestimmt wird, dass die Bedingung $\mathbb{P}_\vartheta(C_\vartheta) > 1 - \alpha$ möglichst knapp erfüllt ist.

(2) Man setze $\mathcal{C} = \{(x, \vartheta) \colon x \in C_\vartheta\}$. Dann bilden die $x$-Schnitte von $\mathcal{C}$, also

$$C(x) = \{\vartheta \in \Theta \colon C_\vartheta \ni x\},$$

einen Konfidenzbereich zum Niveau $\alpha$.

Um in Schritt (1) möglichst *kleine* Mengen $C_\vartheta$ zu erhalten, wählen wir solche $x$, die möglichst *große* Werte $\varrho_\vartheta(x)$ haben. In vielen Fällen wird dies der Fall sein, wenn wir in der Nähe des Erwartungswertes von $\mathbb{P}_\vartheta$ zu suchen beginnen. Wenn also $m(\vartheta)$ den Erwartungswert von $\mathbb{P}_\vartheta$ bezeichnet, so wird man oft $C_\vartheta$ ansetzen als $\{x \in \mathfrak{X} \colon |x - m(\vartheta)| \leq s\}$, wobei $s > 0$ so gewählt ist, dass die Verteilungsfunktion der Zufallsgröße $|X - m(\vartheta)|$ gerade den Wert $1 - \alpha$ überschreitet (falls $X$ die Verteilung $\mathbb{P}_\vartheta$ hat). Daher ist der folgende Begriff im Zusammenhang mit Konfidenzbereichen recht nützlich.

**Definition 8.2.1 (Quantil, Fraktil).** *Es sei $Q$ ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ und $\alpha \in (0, 1)$. Jede Zahl $q \in \mathbb{R}$ mit $Q((-\infty, q]) \geq \alpha$ und $Q([q, \infty)) \geq 1 - \alpha$ heißt ein* $\alpha$-Quantil *von $Q$. Ein* $1/2$*-Quantil heißt ein* Median*, und ein* $(1-\alpha)$*-Quantil heißt ein* $\alpha$-Fraktil*. Quantile einer Zufallsgröße sind die Quantile der zugehörigen Verteilung.*

Quantile sind im Allgemeinen nicht eindeutig bestimmt, es sei denn, dass die Verteilungsfunktion streng monoton wächst. Falls $Q$ eine Dichte $\varrho$ besitzt, sodass die Menge $\{x \colon \varrho(x) > 0\}$ ein Intervall ist, dann ist das $\alpha$-Quantil gerade die Zahl $q$ mit $\int_{-\infty}^{q} \varrho(x)\,\mathrm{d}x = \alpha$.

Nun diskutieren wir drei wichtige Beispiele der Konstruktion von Konfidenzbereichen.

## 8.3 Beispiele

### Emissionskontrolle

Von $N = 10$ Kraftwerken überprüfen wir $n = 4$ zufällig ausgewählte auf ihre Emissionswerte. Gesucht ist ein Konfidenzbereich für die Anzahl $\vartheta$ der Kraftwerke mit zu hohen Emissionswerten. Mathematisch gesehen, betrachten wir ein Stichprobe vom Umfang 4 ohne Zurücklegen aus einer Urne mit 10 Kugeln, von denen eine unbekannte Zahl $\vartheta$ schwarz ist. Also legen wir das statistische Modell $(\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ mit $\mathfrak{X} = \{0, \dots, 4\}$, $\Theta = \{0, \dots, 10\}$ und $\mathbb{P}_\vartheta = \mathrm{Hyp}_{4,\vartheta,10-\vartheta}$ zu Grunde. Es sei die Fehlerschranke $\alpha = \frac{1}{5}$ gegeben.

In der folgenden Tabelle sind die Werte für $\binom{10}{4}\varrho(x, \vartheta) = \binom{10-\vartheta}{4-x}\binom{\vartheta}{x}$ für alle $x$ und alle $\vartheta \leq 5$ dargestellt. Die Werte für $\vartheta > 5$ ergeben sich aus der Symmetrie; zum Beispiel steht in der Zeile für $\vartheta = 6$ die Zeile für $\vartheta = 4$ in umgekehrter Reihenfolge.

| | | | | | |
|---|---|---|---|---|---|
| $\vartheta = 5$ | 5 | $\underline{50}$ | $\underline{100}$ | $\underline{50}$ | 5 |
| $\vartheta = 4$ | 15 | $\underline{80}$ | $\underline{90}$ | 24 | 1 |
| $\vartheta = 3$ | 35 | $\underline{105}$ | $\underline{63}$ | 7 | 0 |
| $\vartheta = 2$ | $\underline{70}$ | $\underline{112}$ | 28 | 0 | 0 |
| $\vartheta = 1$ | $\underline{126}$ | $\underline{84}$ | 0 | 0 | 0 |
| $\vartheta = 0$ | $\underline{210}$ | 0 | 0 | 0 | 0 |
| | $x = 0$ | $x = 1$ | $x = 2$ | $x = 3$ | $x = 4$ |

In jeder Zeile sind (beginnend mit dem jeweils größten Wert) gerade soviele Werte unterstrichen, dass ihre Summe den Wert 168 übersteigt, denn $168\binom{10}{4}^{-1} = \frac{168}{210} = \frac{4}{5} = 1 - \alpha$. Die zugehörigen $x$-Werte gehören in die Menge $C_\vartheta$; so ist zum Beispiel $C_3 = \{1, 2\}$. Also ist somit die Menge $\mathcal{C}$ bestimmt worden. Die $x$-Schnitte ergeben sich, indem man jeweils in der $x$-ten Spalte diejenigen $\vartheta$ sammelt, in deren Zeile der Eintrag unterstrichen ist, also ergibt sich

$$\begin{array}{llllll} C(0) & = \{0,1,2\}, & C(1) & = \{1,\dots,5\}, & C(2) & = \{3,\dots,7\}, \\ C(3) & = \{5,\dots,9\}, & C(4) & = \{8,9,10\}. & & \end{array}$$

Obwohl wir $\alpha$ recht groß gewählt haben, sind die Konfidenzbereiche recht groß, was an der geringen Stichprobenzahl liegt.

**Binomialmodell**

Wir betrachten das Binomialmodell $(\{0,\dots,n\}, (\mathrm{Bi}_{n,\vartheta})_{\vartheta\in(0,1)})$, das in Beispiel 7.3.2 bei der experimentellen Bestimmung einer gewissen unbekannten Wahrscheinlichkeit auftauchte. Gesucht ist ein Konfidenzintervall für die 'Erfolgswahrscheinlichkeit' $\vartheta$. Es sei eine Fehlerschranke $\alpha \in (0,1)$ vorgegeben. Es ist ja $T = x/n$ ein varianzminimierender Schätzer. Wir machen den Ansatz

$$C(x) = \Big(\frac{x}{n} - \varepsilon, \frac{x}{n} + \varepsilon\Big),$$

wobei wir die Wahl von $\varepsilon > 0$ noch so treffen müssen, dass die Bedingung (8.1.1) erfüllt ist, die hier lautet:

$$\mathrm{Bi}_{n,\vartheta}\Big(x\colon \Big|\frac{x}{n} - \vartheta\Big| \geq \varepsilon\Big) = \sum_{x\colon |\frac{x}{n}-\vartheta|\geq\varepsilon} \mathrm{Bi}_{n,\vartheta}(x) \leq \alpha,$$

wobei wir $\mathrm{Bi}_{n,\vartheta}$ sowohl für die Binomialverteilung als auch für ihre Einzelwahrscheinlichkeiten geschrieben haben. Mit anderen Worten, wir suchen das $\alpha$-Quantil $\varepsilon$ für $|X/n - \vartheta|$, wenn $X$ eine $\mathrm{Bi}_{n,\vartheta}$-verteilte Zufallsgröße ist. Wir diskutieren zwei Herangehensweisen: eine Anwendung der Tschebyscheff-Ungleichung und eine der Normalapproximation.

Auf Grund der *Tschebyscheff-Ungleichung* (siehe Korollar 6.1.2) ist die linke Seite nicht kleiner als $\mathbb{V}(\mathrm{Bi}_{n,\vartheta})/(n^2\varepsilon^2)$, und dies ist gleich $\vartheta(1-\vartheta)/(n\varepsilon^2)$. Da wir $\vartheta$ nicht kennen, schätzen wir weiter ab gegen $1/(4n\varepsilon^2)$. Also ist die Bedingung (8.1.1) erfüllt, sobald $\varepsilon > 1/\sqrt{4n\varepsilon^2}$. Für $n = 1000$ und $\alpha = 0.025$ braucht man also beispielsweise $\varepsilon > 0.1$.

Mit der Tschebyscheff-Ungleichung kommt man also mit geringem Aufwand an sichere Abschätzungen, aber diese Ungleichung ist sehr allgemein und nutzt keine besonderen Eigenschaften der Binomialverteilung aus, ist also recht grob. Das errechnete $\varepsilon$ könnte also (und ist es tatsächlich) unnötigerweise zu groß sein.

Im zweiten Ansatz nehmen wir an, dass $n$ so groß ist, dass wir den *zentralen Grenzwertsatz* (siehe Satz 6.2.2 oder 6.2.5) als Approximation einsetzen können, wie etwa in Beispiel 6.2.7. Also erhalten wir

$$\mathrm{Bi}_{n,\vartheta}\Big(x\colon \Big|\frac{x}{n}-\vartheta\Big|<\varepsilon\Big)=\mathrm{Bi}_{n,\vartheta}\Big(x\colon \Big|\frac{x-n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}\Big|<\varepsilon\sqrt{\frac{n}{\vartheta(1-\vartheta)}}\Big)\approx 2\Phi\Big(\varepsilon\sqrt{\frac{n}{\vartheta(1-\vartheta)}}\Big)-1,$$

wobei $\Phi$ wie üblich die Verteilungsfunktion der Standardnormalverteilung ist. Im Beispiel $n = 1000$ und $\alpha = 0.025$ muss also $\varepsilon$ die Bedingung

$$\Phi\Big(\varepsilon\sqrt{\frac{n}{\vartheta(1-\vartheta)}}\Big)\geq\frac{1}{2}\big(1+0.975+0.02\big)=0.9975$$

erfüllen, wobei wir eine willkürliche Sicherheitsmarge 0.02 für den Approximationsfehler eingefügt haben. Einer Tabelle entnimmt man, dass $\Phi(2.82) \approx 0.9975$, also reicht die Bedingung $\varepsilon > 2.82/\sqrt{4000} \approx 0.0446$ aus. Wir haben also das Ergebnis der ersten Methode um den Faktor 2 verbessert.

### Mittelwert im Gaußschen Produktmodell

Wir betrachten das $n$-fache Gaußsche Produktmodell $(\mathbb{R}^n, (\mathcal{N}(m,v)^{\otimes n})_{m\in\mathbb{R},v\in(0,\infty)})$, wobei wie üblich $\mathcal{N}(m,v)$ die Normalverteilung mit Erwartungswert $m$ und Varianz $v^2$ ist. Die Parametermenge $\Theta = \mathbb{R}\times(0,\infty)$ hat also zwei Komponenten, und wir wollen die erste Komponente $m(\vartheta)$ von $\vartheta = (m(\vartheta), v(\vartheta))$ schätzen. Wir wollen Konfidenzintervalle zum Niveau $\alpha \in (0,1)$ angeben.

In offensichtlicher Verallgemeinerung des in Abschnitt 8.2 gegebenen Verfahrens müssen wir also für jedes $m \in \mathbb{R}$ eine (möglichst kleine) Menge $C_m \subset \mathbb{R}^n$ bestimmen mit $\mathbb{P}_\vartheta(C_{m(\vartheta)}) \geq 1-\alpha$ für jedes $\vartheta \in \Theta$, wobei wir $\mathbb{P}_\vartheta$ für $\mathcal{N}(m,v)^{\otimes n}$ schreiben. Der gesuchte Konfidenzbereich für $m(\vartheta)$ hat dann die Gestalt $C(x) = \{m \in \mathbb{R}\colon C_m \ni x\}$, wie wir weiter unten noch konkretisieren werden.

Für die Menge $C_m$ machen wir den Ansatz $C_m = \{x \in \mathbb{R}^n\colon |M(x)-m| \leq s(x)\}$ (wobei $M(x) = \frac{1}{n}\sum_{i=1}^n x_i$ der empirische Mittelwert ist) für ein geeignetes $s(x) > 0$, denn die Dichte von $M = M(X)$ (wenn $X_1,\dots,X_n$ unabhängig und $\mathcal{N}(m,v)$-verteilt sind) sollte maximal sein genau in $m(\vartheta)$ (und ist es auch). Die Abweichung $s(x)$ setzen wir an als $s(x) = t\sqrt{V^*(x)/n}$, wobei $V^*(x) = \frac{1}{n-1}\sum_{i=1}^n (x_i - M(x))^2$ die korrigierte empirische Varianz ist, denn die Abweichung vom Mittelwert sollte für große $n$ von gerade dieser Größenordnung sein. (Man sieht leicht, dass die Varianz von $M(X)$ von der Ordnung $\frac{1}{n}$ ist, und aus Satz 7.4.2 wissen wir, dass der Erwartungswert von $V^* = V^*(X)$ konstant ist.)

Nun ist also nur noch der Wert von $t$ abhängig von $\alpha$ (und von $n$) zu bestimmen, und es wird sich heraus stellen, dass er tatsaechlich von nichts Anderem abhängt. Wir formulieren unseren Ansatz für $C_m$ mit Hilfe der Statistik

$$T_m = T_m(X_1,\dots,X_n) = \sqrt{n}\frac{M-m}{\sqrt{V^*}},$$

wobei wie in Lemma 7.4.2 $M = \frac{1}{n}\sum_{i=1}^n X_i$ und $V^* = \frac{1}{n-1}\sum_{i=1}^n (X_i - M)^2$ das Stichprobenmittel und die korrigierte Stichprobenvarianz sind, und $X_1,\dots,X_n$ sind unabhängig und $\mathcal{N}(m,v)$-verteilt. Dann können wir die oben angesetzte Menge $C_m$ schreiben als $C_m = \{x \in \mathbb{R}^n\colon |T_m(x)| \leq t\}$, und wir müssen $t$ so bestimmen, dass die Bedingung $1-\alpha \leq \mathbb{P}_\vartheta(\{x\colon |T_{m(\vartheta)}(x)| \leq t\}) = \mathbb{P}_\vartheta(|T_{m(\vartheta)}| \leq t)$ möglichst knapp erfüllt ist, also etwa mit Gleichheit.

Als Erstes überlegen wir uns, dass die Verteilung der Zufallsgröße $T_{m(\vartheta)}$ unter $\mathbb{P}_\vartheta$ nicht vom Parameter $\vartheta$ abhängt. Dies folgt aus der Überlegung, dass Zufallsgrößen $X_1, \dots, X_n$ genau dann $\mathcal{N}(m, v)$-verteilt sind, wenn die Zufallsgrößen $\widetilde{X}_1 = (X_1 - m)/v, \dots, \widetilde{X}_n = (X_n - m)/v$ die $\mathcal{N}(0,1)$-Verteilung haben. Ferner bemerke man, dass $T_m$ die selbe lineare Transformation der Zufallsvariablen $X_1, \dots, X_n$ ist wie $T_0$ sie für $\widetilde{X}_1, \dots, \widetilde{X}_n$ darstellt.

Also ist die Verteilung von $T_{m(\vartheta)}$ unter $\mathbb{P}_\vartheta$ gleich der von

$$T_0 = \frac{M}{\sqrt{\frac{1}{n}V^*}} = \frac{n^{-1/2}\sum_{i=1}^n X_i}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - M)^2}} \tag{8.3.1}$$

unter $\mathbb{P}_{0,1} = \mathcal{N}(0,1)^{\otimes n}$. Diese Verteilung wollen wir $Q$ nennen, also

$$Q(I) = \mathbb{P}_{0,1}\big(T_0 \in I\big) = \mathcal{N}(0,1)^{\otimes n}\big(T_0 \in I\big), \qquad I \subset \mathbb{R}.$$

Die Verteilung $Q$ ist bekannt als die *Studentsche t-Verteilung* mit $n-1$ Freiheitsgraden. In Abschnitt 8.4 stellen wir diese Verteilung näher vor. Es sei hier nur bemerkt, dass $Q$ eine Dichte besitzt, die positiv und symmetrisch ist und streng fallend im Intervall $[0, \infty)$.[1] Wenn $F_Q$ ihre Verteilungsfunktion bezeichnet, dann ist also $F_Q\colon \mathbb{R} \to (0,1)$ streng steigend und bijektiv und besitzt daher eine Umkehrfunktion $F_Q^{-1}\colon (0,1) \to \mathbb{R}$.

Wir betrachten nun das $\alpha/2$-Quantil der Studentschen $t_{n-1}$-Verteilung, also die Zahl

$$t = t_\alpha = F_Q^{-1}(1 - \alpha/2). \tag{8.3.2}$$

Sie hat die Eigenschaft, dass das Intervall $I = (-t, t)$ die $Q$-Wahrscheinlichkeit $1 - \alpha$ besitzt, denn aus Symmetriegründen gilt

$$Q\big((-t,t)\big) = Q\big((-\infty,t)\big) - Q\big((-\infty,-t)\big) = F_Q(t) - (1 - F_Q(t)) = 2(1-\alpha/2) - 1 = 1 - \alpha.$$

Die Menge $C_m = \{x \in \mathbb{R}^n\colon T_m(x) \in I\} = \{x \in \mathbb{R}^n\colon |T_m(x)| < t\}$ erfüllt dann

$$\mathbb{P}_\vartheta(C_{m(\vartheta)}) = \mathbb{P}_{0,1}(C_0) = \mathbb{P}_{0,1}\big(|T_0| < t\big) = \mathcal{N}(0,1)^{\otimes n}\big(|T_0| < t\big) = Q\big((-t,t)\big) = 1 - \alpha$$

für alle $\vartheta = (m(\vartheta), v(\vartheta))$. Also haben wir die Mengen $C_m$ geeignet konstruiert, und wir können nun die Konfidenzbereiche angeben. Man sieht leicht die Äquivalenzen

$$x \in C_m \qquad \Longleftrightarrow \qquad \big|M(x) - m\big| < t\sqrt{\frac{1}{n}V^*(x)} \qquad \Longleftrightarrow \qquad m \in C(x),$$

wobei

$$C(x) = \Big(M(x) - t\sqrt{\frac{1}{n}V^*(x)}, M(x) + t\sqrt{\frac{1}{n}V^*(x)}\Big).$$

Also ist das Intervall $C(x)$ der $x$-Schnitt der Menge $\mathcal{C} = \{(x,\vartheta)\colon x \in C_{m(\vartheta)}\}$, also ein Konfidenzintervall zum Niveau $\alpha$.

**Beispiel 8.3.1 (Vergleich zweier Schlafmittel).** Die Wirkung zweier Schlafmittel $A$ und $B$ soll verglichen werden. Dazu erhalten zehn Patienten in zwei auf einander folgenden Nächten jeweils zuerst $A$ und dann $B$ verabreicht, und man misst die jeweilige Schlafdauer. Es ergaben sich die folgenden Werte für die Differenzen der Schlafdauern:

[1] Tatsächlich wird sich später zeigen (siehe die Bemerkung zu Satz 8.4.7), dass für $n \to \infty$ die Verteilung $Q$ gegen die Standardnormalverteilung konvergiert.

| Patient | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Differenz | 1.2 | 2.4 | 1.3 | 1.3 | 0.0 | 1.0 | 1.8 | 0.8 | 4.6 | 1.4 |

Man sieht, dass das zweite Medikament wirkungsvoller ist, da alle Werte nicht negativ sind. Man errechnet aus diesem Datensatz $x = (x_1, \dots, x_{10})$, dass $M(x) = 1.58$ und $V^*(x) = 1.513$. Wenn man davon ausgeht, dass die Schlafdauer sich aus vielen kleinen unabhängigen Einflüssen zusammen setzt, kann man auf Grund des Zentralen Grenzwertsatzes annehmen, dass die Werte näherungsweise normalverteilt sind mit unbekannten Parametern $m$ und $v$. Dann sind wir in der obigen Situation und können die oben angegebenen Konfidenzintervalle benutzen. Für etwa die Schranke $\alpha = 0.025$ kann man Tabellen für die $t$-Verteilung den Wert $t = 2.72$ entnehmen und erhält für den Datensatz der Tabelle das Konfidenzintervall $C(x) = (0.52, 2.64)$ für $m$. ◇

## 8.4 Die $\chi^2$- und $t$-Verteilungen

Wie wir im Abschnitt 8.3 gesehen haben, tauchen bei der Erwartungswertschätzung im Produkt-Gaußmodell Summen der Quadrate unabhängiger normalverteilter Zufallsvariabler auf sowie Quotienten solcher Größen. In diesem Abschnitt identifizieren wir ihre Verteilungen. Dabei treten wichtige Wahrscheinlichkeitsmaße auf (wie die $\chi^2$-Verteilung und die Studentsche $\mathbf{t}$-Verteilung), die in der Schätz- und Testtheorie große Bedeutung haben.

Zunächst eine grundlegende Beobachtung. Es sei daran erinnert (siehe Beispiel 5.3.5), dass die Gamma-Verteilung mit Parametern $\alpha > 0$ und $r > 0$ die Dichte

$$\gamma_{\alpha,r}(t) = \frac{\alpha^r}{\Gamma(r)} t^{r-1} e^{-\alpha t} \mathbb{1}_{[0,\infty)}(t),$$

besitzt.

**Lemma 8.4.1.** *Wenn $X$ eine standardnormalverteilte Zufallsgröße ist, so hat $X^2$ die Gamma-Verteilung mit Parametern $\alpha = r = \frac{1}{2}$.*

**Beweis.** Für alle $x \in (0, \infty)$ ist

$$\begin{aligned}\mathbb{P}(X^2 \le x) &= 2\mathbb{P}\big(0 < X < \sqrt{x}\big) = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{x}} e^{-t^2/2}\,\mathrm{d}t = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-s/2} s^{-1/2}\,\mathrm{d}s \\ &= \int_0^x \gamma_{1/2,1/2}(s)\,\mathrm{d}x,\end{aligned}$$

wie eine Substitution $s = t^2$ ergibt. □

Da die Gamma-Verteilungen eine schöne Faltungseigenschaft haben (siehe Lemma 5.3.6), können wir auch sofort die Verteilung der Summe von Quadraten unabhängiger normalverteilter Zufallsgrößen identifizieren:

**Lemma 8.4.2.** *Wenn $X_1, \dots, X_n$ unabhängige, standardnormalverteilte Zufallsgrößen sind, so hat $\sum_{i=1}^n X_i^2$ die Gamma-Verteilung mit Parametern $\alpha = 1/2$ und $r = n/2$.*

**Beweis.** Man kombiniere Lemma 8.4.1 und Lemma 5.3.6(i). □

Die hier auftretende Verteilung ist so wichtig, dass sie einen eigenen Namen erhalten hat:

**Definition 8.4.3 (Chiquadrat-Verteilung).** *Für jedes $n \in \mathbb{N}$ heißt die Gamma-Verteilung mit Parametern $\alpha = 1/2$ und $r = n/2$ die* Chiquadrat-Verteilung mit $n$ Freiheitsgraden *und wird mit $\boldsymbol{\chi}^2_n$ bezeichnet. Sie besitzt die Dichte*

$$\chi^2_n(t) = \gamma_{1/2,n/2}(t) = \frac{t^{n/2-1}}{\Gamma(n/2)2^{n/2}} e^{-t/2}, \qquad t > 0. \tag{8.4.1}$$

Wie wir später sehen werden (siehe Satz 8.4.7), hat zum Beispiel die korrigierte Stichprobenvarianz $V^*$ im Produkt-Gaußmodell gerade die Chiquadrat-Verteilung. Für $n = 2$ ist die $\boldsymbol{\chi}^2_2$-Verteilung gerade die Exponential-Verteilung mit Parameter $\frac{1}{2}$. Die $\boldsymbol{\chi}^2_1$-Dichte ist streng fallend und explodiert bei Null, während die $\boldsymbol{\chi}^2_n$-Dichte für $n \geq 3$ bis zu ihrem Maximumpunkt $n-2$ streng steigt und danach streng fällt.

Als Nächstes wollen wir die Verteilung von Quotienten gewisser Zufallsgrößen identifizieren. Dazu zunächst eine allgemeine, elementare Beobachtung.

**Lemma 8.4.4.** *Seien $X$ und $Y$ zwei unabhängige Zufallsgrößen mit Dichten $f$ bzw. $g$, und sei $Y > 0$. Ferner sei $\alpha > 0$ eine Konstante.*

*(i) Dann hat $X/Y$ die Dichte $h$ mit*

$$h(t) = \int_0^\infty f(ty)g(y)y \,\mathrm{d}y, \qquad t \in \mathbb{R};$$

*insbesondere hat $X/\alpha$ hat die Dichte $t \mapsto \alpha f(\alpha t)$.*

*(ii) Ferner hat $\sqrt{Y}$ die Dichte $t \mapsto 2tg(t^2)$.*

**Beweis.** Übungsaufgabe. □

**Lemma 8.4.5.** *Es seien $X$ und $Y_1, \ldots, Y_n$ unabhängige, standardnormalverteilte Zufallsgrößen. Dann hat*

$$T = \frac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}}$$

*die Dichte*

$$\tau_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)\sqrt{n}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \qquad t \in \mathbb{R}. \tag{8.4.2}$$

**Beweis.** Wir kombinieren Lemma 8.4.2 und Lemma 8.4.4: Die Zufallsgröße $Y = \sum_{i=1}^n Y_i^2$ hat die Dichte $\gamma_{1/2,n/2}$, also hat $\frac{1}{n}Y$ die Dichte $t \mapsto n\gamma_{1/2,n/2}(nt)$, und daher hat der Nenner $\sqrt{\frac{1}{n}Y}$ die Dichte $t \mapsto 2tn\gamma_{1/2,n/2}(nt^2)$. Die Dichte von $T$ ist also

$$t \mapsto h(t) = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2y^2/2} 2y^2 n \gamma_{1/2,n/2}(ny^2) \,\mathrm{d}y.$$

Nun setzen wir (8.4.1) ein, benutzen die Substitution $s = y^2(t^2+n)/2$ und fassen zusammen:

$$\begin{aligned} h(t) &= \frac{2n^{n/2}}{\sqrt{2\pi}\Gamma(n/2)2^{n/2}} \int_0^\infty e^{-y^2(t^2+n)/2}\, y^n \; dy \\ &= \frac{1}{\Gamma(1/2)\Gamma(n/2)\sqrt{n}} \Big(1+\frac{t^2}{n}\Big)^{-\frac{n+1}{2}} n \int_0^\infty e^{-s} s^{(n-1)/2}\, \mathrm{d}s, \end{aligned}$$

wobei wir die Tatsache $\Gamma(1/2) = \sqrt{\pi}$ benutzten. Eine elementare Vollständige Induktion zeigt unter Benutzung der Funktionalgleichung $\Gamma(x+1) = x\Gamma(x)$, dass $\int_0^\infty e^{-s} s^{(n-1)/2}\, \mathrm{d}s = \frac{1}{n}\Gamma((n+1)/2)$, und dies beendet den Beweis. □

Auch diese Verteilung ist von großer Bedeutung in der Statistik:

**Definition 8.4.6 (Student-Verteilung).** *Die Wahrscheinlichkeitsverteilung auf* $\mathbb{R}$ *mit Dichte* $\tau_n$ *in* (8.4.2) *heißt die* Studentsche $t$-Verteilung mit $n$ Freiheitsgraden *oder kurz die* $t_n$-Verteilung.

Für $n = 1$ stimmt die Studentsche $t$-Verteilung überein mit der in Beispiel 5.3.10 eingeführten Cauchy-Verteilung.

Man sieht in (8.4.2), dass $\lim_{n\to\infty} \tau_n(t) = ce^{-t^2/2}$ für jedes $t \in \mathbb{R}$, wobei die Konstante $c$ natürlich nichts Anderes sein kann als $(2\pi)^{-1/2}$. Da diese Konvergenz sogar gleichmäßig auf kompakten Intervallen ist, konvergiert die $t_n$-Verteilung gegen die Standardnormalverteilung im schwachen Sinn (siehe Definition 6.2.1), wovon man sich leicht überzeugt. Als Korollar können wir also fest halten, dass das in Lemma 8.4.5 definierte $T$ schwach gegen $\mathcal{N}(0,1)$ konvergiert, eine einigermaßen überraschende Aussage. Das nächste Ergebnis zeigt, dass uns in Abschnitt 8.3 eine Zufallsgröße mit der Verteilung von $T$ (allerdings mit $n-1$ statt $n$) schon einmal unter die Augen gekommen ist, und zwar war dies die Zufallsgröße $T_0$, mit deren Hilfe wir Konfidenzintervalle im Produkt-Gauß-Modell konstruierten. Der folgende Satz identifiziert ihre Verteilung.

**Satz 8.4.7 (Die Verteilung von $T_0$ in (8.3.1)).** *Es seien* $n \in \mathbb{N}$ *und* $X_1, \dots, X_n$ *unabhängige standardnormalverteilte Zufallsgrößen. Wir betrachten das Stichprobenmittel und die korrigierte Stichprobenvarianz,*

$$M = \frac{1}{n}\sum_{i=1}^n X_i \qquad \text{und} \qquad V^* = \frac{1}{n-1}\sum_{i=1}^n (X_i - M)^2.$$

*Dann gelten:*

*(i)* $M$ *und* $V^*$ *sind unabhängig.*

*(ii)* $M$ *hat die Verteilung* $\mathcal{N}(0, n^{-1/2})$ *und* $(n-1)V^*$ *die Verteilung* $\chi^2_{n-1}$.

*(iii)* $T_0 = \sqrt{n}M/\sqrt{V^*} = n^{-1/2}\sum_{i=1}^n X_i / \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - M)^2}$ *hat die Verteilung* $t_{n-1}$.

Die Unabhängigkeit von $M$ und $V^*$ ist auf dem ersten Blick überraschend, denn $M$ taucht in der Definition von $V^*$ auf. Aber man erinnere sich daran (siehe Bemerkung 2.2.5(d)), dass Unabhängigkeit nichts mit kausaler Unabhängigkeit zu tun hat. Die Aussage in (i) benutzt sehr spezielle Eigenschaften der Normalverteilung.

Aus (iii) folgt also die in Abschnitt 8.3 erwähnte Tatsache, dass die Verteilung $Q$ von $T_0$ unter $\mathcal{N}(0,1)^{\otimes n}$ die Studentsche $t$-Verteilung mit $n-1$ Freiheitsgraden ist. Zusammen mit der Bemerkung nach Definition 8.4.6 ergibt sich sogar, dass $T_0$ für $n \to \infty$ schwach gegen die Standardnormalverteilung konvergiert.

**Beweis.** Wir schreiben $X = (X_1, \ldots, X_n)^{\mathrm{T}}$ für den normalverteilten Vektor, dessen Erwartungswertvektor 0 ist und dessen Kovarianzmatrix die $n \times n$-Einheitsmatrix $E_n$ ist. Sei $A$ eine orthogonale $n \times n$-Matrix, deren erste Zeile nur die Einträge $n^{-1/2}$ hat. Nach Beispiel 5.3.9 ist der Vektor $Y = AX$ normalverteilt mit Erwartungswertvektor $A0 = 0$ und Kovarianzmatrix $AA^{\mathrm{T}} = E_n$, also ist auch $Y$ wieder standardnormalverteilt. Insbesondere sind die Komponenten $Y_1, \ldots, Y_n$ unabhängig und standardnormalverteilt. Man beachte, dass (wegen der besonderen Form der ersten Zeile von $A$) $Y_1 = n^{-1/2} \sum_{i=1}^n X_i = \sqrt{n} M$. Ferner errechnet man leicht, dass

$$(n-1)V^* = \sum_{i=1}^n (X_i - M)^2 = \sum_{i=1}^n X_i^2 - nM^2 = \|A^{-1}Y\|^2 - Y_1^2 = \|Y\|^2 - Y_1^2 = \sum_{i=2}^n Y_i^2.$$

Da also $M$ nur von $Y_1$ abhängt und $(n-1)V^*$ nur von $Y_2 \ldots, Y_n$, ist die Unabhängigkeit bewiesen, also haben wir (i) gezeigt. Außerdem haben wir auch gezeigt, dass $M = n^{-1/2} Y_1$ normalverteilt ist mit Varianz $1/n$, und $(n-1)V^* = Y_2^2 + \cdots + Y_n^2$ hat nach Lemma 8.4.2 die $\boldsymbol{\chi}^2_{n-1}$-Verteilung, was (ii) impliziert. Nun folgt (iii) aus Lemma 8.4.5. □

# Kapitel 9

# Einführung in die Testtheorie

In Kapiteln 7 und 8 lernten wir im Rahmen der Schätztheorie Methoden kennen, um aus Beobachtungen den zu Grunde liegenden Zufallsmechanismus möglichst adäquat zu beschreiben. Im vorliegenden Kapitel behandeln wir die Situation, dass eine Entscheidung getroffen werden muss, welcher von gewissen in Frage kommenden Mechanismen vorliegt. Man formuliert eine Hypothese und entscheidet sich an Hand der Beobachtungen eines Experiments, ob man die Hypothese für zutreffend hält oder nicht. Wir werden Entscheidungsregeln, sogannte Tests, entwickeln und mit mathematischen Methoden auf ihre Qualität untersuchen, insbesondere auf eine möglichst geringe Irrtumswahrscheinlichkeit. Man erwarte allerdings nicht, dass die mathematische Theorie die Antwort geben kann, ob die Entscheidung letztendlich richtig war oder nicht.

## 9.1 Entscheidungsprobleme

Wir beginnen wieder mit einem einführenden Beispiel.

**Beispiel 9.1.1 (Qualitätskontrolle).** Ein Importeur erhält eine Lieferung von 10 000 Orangen. Den vereinbarten Preis muss er nur zahlen, wenn höchstens 5% davon faul sind. Da er nicht alle Orangen prüfen kann, untersucht er nur eine Stichprobe von 50 Orangen und setzt sich eine Grenze von $c$ faulen Orangen, die er bereit wäre zu akzeptieren. Falls er also höchstens $c$ faule Orangen in der Stichprobe entdeckt, akzeptiert er die Lieferung, sonst reklamiert er.

Also wählt der Importeur einen Test (eine Entscheidungsregel), dessen Ausführung ihm die Entscheidung abnehmen soll. Dieser Test hängt von der richtigen Wahl von $c$ ab. Zu kleine $c$ machen die Wahrscheinlichkeit groß, dass er ablehnen muss, obwohl die Qualität der Lieferung doch ganz ordentlich ist, und zu große $c$ lassen ihn eventuell eine schlechte Ware abnehmen. $\diamond$

Wir formulieren zunächst das Verfahren ein wenig allgemeiner in fünf Schritten:

1. Das statistische Modell wird aufgestellt. Im obigen Beispiel wäre das das hypergeometrische Modell mit $\mathfrak{X} = \{0, \dots, 50\}$, $\Theta = \{0, \dots, 10\,000\}$ und $\mathbb{P}_\vartheta = \mathrm{Hyp}_{50;\vartheta,10\,000-\vartheta}$.

2. Wir zerlegen die Parametermenge $\Theta$ in zwei Teilmengen $\Theta_0$ und $\Theta_1$ mit der Interpretation

$$\begin{array}{lll} \vartheta \in \Theta_0 & \Longleftrightarrow & \vartheta \text{ ist akzeptabel} \quad (\textit{Hypothese}), \\ \vartheta \in \Theta_1 & \Longleftrightarrow & \vartheta \text{ ist problematisch} \quad (\textit{Alternative}). \end{array}$$

Man sagt, dass die Hypothese $H_0 : \vartheta \in \Theta_0$ gegen die Alternative $H_1 : \vartheta \in \Theta_1$ getestet werden soll. Im obigen Beispiel hat man also $\Theta_0 = \{0, \dots, 500\}$ und $\Theta_1 = \{501, \dots, 10\,000\}$.

3. Wir wählen ein Irrtumsniveau $\alpha \in (0, 1)$ und wollen, dass die Wahrscheinlichkeit eines *Fehlers 1. Art* (d. h. einer Entscheidung für die Alternative, obwohl die Hypothese vorliegt) unter $\alpha$ liegt.

4. Wir wählen eine Entscheidungsregel, d. h. eine Statistik $\varphi\colon \mathfrak{X} \to [0, 1]$ mit der Interpretation, dass die Entscheidung für die Alternative mit der Wahrscheinlichkeit $\varphi(x)$ (wobei $x$ der Beobachtungswert ist) fällt, also

   $\varphi(x) = 0 \iff$ Festhalten an der Hypothese,
   $\varphi(x) = 1 \iff$ Verwerfen der Hypothese, Annahme der Alternative,
   $\varphi(x) \in (0, 1) \iff$ keine Klarheit: Durchführung eines Zufallsexperimentes, das mit Wahrscheinlichkeit $\varphi(x)$ die Alternative wählen lässt.

   Im obigen Beispiel kann der Importeur zum Beispiel die Statistik $\varphi(x) = \mathbb{1}_{\{x>c\}}$ wählen, die also zur Annahme der Hypothese rät, wenn nicht mehr als $c$ Orangen in der Stichprobe faul sind, und sonst die Hypothese verwirft. Bei dieser Entscheidungsregel wird also der Zufall nicht bemüht, denn es gibt keinen Wert von $x$ mit $0 < \varphi(x) < 1$. Er kann aber auch die Statistik

$$\varphi(x) = \begin{cases} 1, & \text{falls } x > c, \\ \frac{1}{2}, & \text{falls } x = c, \\ 0, & \text{falls } x < c, \end{cases}$$

   wählen, die bei genau $c$ faulen Orangen rät: ‘Wirf eine Münze!’.

5. Man führt das Experiment durch.

Ein paar Kommentare sind hilfreich:

**Bemerkung 9.1.2.** (i) Sehr wichtig ist, dass die Durchführung des Experiments als *letzter* Schritt erfolgt, wenn also alle anderen Festlegungen schon getroffen worden sind! Auf Grund der menschlichen Natur sind die Versuchungen zu groß, die Entscheidungsregeln den Beobachtungsdaten anzupassen oder auch Ausreißer zu eliminieren, sodass das Ergebnis des Tests vielleicht der voreingenommenen Erwartung besser entspricht.

(ii) Das obige Verfahren ist symmetrisch in der Hypothese und der Alternative, bis auf das Ziel, dass wir die Irrtumswahrscheinlichkeit eng begrenzen wollen, dass man die Alternative wählt, obwohl die Hypothese vorliegt, nicht anders herum. Dies trägt den meisten Anwendungen Rechnung, in der nämlich die Hypothese behauptet, dass es sich um den ‘normalen’ Fall handelt (also um ‘reinen Zufall’) und die Alternative sagt, dass da eine systematische signifikante Abweichung vorliegt. Erst wenn deutliche Hinweise auf das Vorliegen einer Abweichung vorliegen, wird man sich für die Alternative entscheiden, sonst bleibt man bei der Vermutung, das war alles nur Zufall. (Zum Beispiel wird man ein neues Medikament erst annehmen, wenn seine Vorzüge gegenüber den bekannten sehr überzeugend ist.) Deshalb nennt man die Hypothese auch oft die *Nullhypothese*. Eine Entscheidung für die Hypothese heißt nicht, dass die Hypothese als erwiesen gilt oder als einigermaßen sicher, sondern nur, dass das Ergebnis des Experiments keine genügende Rechtfertigung liefert für das Vorliegen der Alternative. Besser wird dies vielleicht durch die Formulierung

'die Hypothese wird nicht verworfen' ausgedrückt. Im Zweifelsfall kann man ja ein neues, verlängertes oder modifiziertes Experiment anstellen.

(iii) Wir unterscheiden also

| | |
|---|---|
| Fehler 1. Art: | $\vartheta \in \Theta_0$, aber die Hypothese wird verworfen, |
| Fehler 2. Art: | $\vartheta \in \Theta_1$, aber die Hypothese wird angenommen. |

Wie wir sehen werden, kann man die Wahrscheinlichkeit, einen Fehler 1. Art zu begehen, unter eine gegebene Schranke drücken, aber dann im Allgemeinen nicht mehr die Wahrscheinlichkeit für Fehler 2. Art.

◇

Der mathematische Kern ist der folgende.

**Definition 9.1.3 (Test, Hypothese, Niveau, Macht).** *Sei $(\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\Theta = \Theta_0 \cup \Theta_1$ eine Zerlegung der Parametermenge in die* (Null-)Hypothese $\Theta_0$ *und die* Alternative $\Theta_1$.

*(a) Jede Statistik $\varphi\colon \mathfrak{X} \to [0,1]$ (die als Entscheidungsregel interpretiert wird) heißt ein* Test *von $\Theta_0$ gegen $\Theta_1$. Sie heißt* nichtrandomisiert, *falls $\varphi(x) \in \{0,1\}$ für alle $x \in \mathfrak{X}$, anderenfalls* randomisiert. *Im ersten Fall heißt $\{x \in \mathfrak{X}\colon \varphi(x) = 1\}$ der* Ablehnungsbereich, Verwerfungsbereich *oder* kritische Bereich *des Tests $\varphi$.*

*(b) Die Größe $\sup_{\vartheta\in\Theta_0} \mathbb{E}_\vartheta(\varphi)$ (dies ist die maximale Wahrscheinlichkeit für einen Fehler 1. Art) heißt der* Umfang *oder das* effektive Niveau *des Tests $\varphi$. Man nennt $\varphi$ einen Test* zum (Irrtums-) Niveau $\alpha$, *wenn $\sup_{\vartheta\in\Theta_0} \mathbb{E}_\vartheta(\varphi) \le \alpha$.*

*(c) Die Funktion $G_\varphi\colon \Theta \to [0,1]$, $G_\varphi(\vartheta) = \mathbb{E}_\vartheta(\varphi)$, heißt* Gütefunktion *des Tests $\varphi$. Für $\vartheta \in \Theta_1$ heißt $G_\varphi(\vartheta)$ die* Macht, Stärke *oder* Schärfe *von $\varphi$ bei $\vartheta$. Die Macht ist also die Wahrscheinlichkeit, mit der die Alternative erkannt wird, wenn sie vorliegt, und $\beta_\varphi(\vartheta) = 1 - G_\varphi(\vartheta)$ ist die Wahrscheinlichkeit für einen Fehler 2. Art.*

Wir werden also natürlicherweise die zwei folgenden Forderungen an einen Test $\varphi$ stellen:

- $G_\varphi(\vartheta) \le \alpha$ für alle $\vartheta \in \Theta_0$, d. h. die Irrtumswahrscheinlichkeit 1. Art soll unter $\alpha$ liegen,
- Für $\vartheta \in \Theta_1$ soll die Macht $G_\varphi(\vartheta)$ möglichst groß sein, ein Fehler 2. Art also möglichst wenig wahrscheinlich.

**Definition 9.1.4.** *Ein Test $\varphi$ heißt* (gleichmäßig) bester Test zum Niveau $\alpha$, *wenn er vom Niveau $\alpha$ ist und für jeden anderen Test $\psi$ zum Niveau $\alpha$ und für jedes $\vartheta \in \Theta_1$ gilt: $G_\varphi(\vartheta) \ge G_\psi(\vartheta)$.*

Im folgenden Beispiel geht der Tester unsachgemäß vor:

**Beispiel 9.1.5 (Außersinnliche Wahrnehmungen).** Ein Medium behauptet, mit seinen außersinnlichen Fähigkeiten verdeckte Spielkarten identifizieren zu können. Zur Überprüfung legt man ihm 20 Mal die Herz-Dame und den Herz-König in zufälliger Anordnung verdeckt vor und lässt es jeweils die Herz-Dame aufdecken. Zum Schluss zählt man die 'Treffer', also die richtig identifizierten Karten.

Ein geeignetes Modell ist das Binomialmodell $(\{0,\dots,20\},(\mathrm{Bi}_{20,\vartheta})_{\vartheta\in\Theta})$ mit $\Theta=[\frac12,1]$. Man will die Nullhypothese $\Theta_0=\{\frac12\}$ gegen die Alternative $\Theta_1=(\frac12,1]$ testen und wählt das Irrtumsniveau $\alpha=0.05$. Als Test wählt man $\varphi=\mathbb{1}_{\{c,\dots,20\}}$ mit geeignetem $c$; wir werden in Satz 9.3.4 sehen, dass dies ein in gewissem Sinn optimaler Test ist. Durch einen Blick auf die konkreten Werte der Binomialverteilung $\mathrm{Bi}_{20,\vartheta}$ erkennt man, dass man $c=15$ wählen muss, um das Niveau 0.05 einzuhalten, denn für $\varphi=\mathbb{1}_{\{15,\dots,20\}}$ haben wir $G_\varphi(\frac12)=\mathrm{Bi}_{20,\frac12}(\{15,\dots,20\})\approx 0.02707$, während der Test $\psi=\mathbb{1}_{\{14,\dots,20\}}$ nur den Wert $G_\psi(\frac12)=\mathrm{Bi}_{20,\frac12}(\{14,\dots,20\})\approx 0.0577>\alpha$ hat.

Das Experiment wird durchgeführt, und das Medium erzielt 14 Treffer. Also hat man $\varphi(14)=0$, und der Test hat die besonderen Fähigkeiten nicht bestätigt. So weit, so gut.

Dieses Ergebnis wurmt den Tester, denn immerhin sind ja 14 Treffer nicht schlecht, und wenn man von vorne herein die Statistik $\psi$ gewählt hätte, dann hätte man ja das Irrtumsniveau nur um eine winzige Kleinigkeit verschlechtert. Also liest man im abschließenden Bericht über das Experiment, dass die medialen Fähigkeiten bei einem Irrtumsniveau von $\alpha'=0.6$ bestätigt werden konnten.

Dieser Schluss aus dem Experiment ist Betrug, denn der Tester hat tatsächlich nicht den Test $\psi=\mathbb{1}_{\{14,\dots,20\}}$ gewählt, sondern den Test $\widetilde{\psi}(x)=\mathbb{1}_{\{x,\dots,20\}}(x)$, d. h. er hat seine Testkriterien den Daten angepasst und damit genau den Fehler begangen, vor dem in Bemerkung 9.1.2(i) gewarnt wurde. Falls er den Test für unfair gegenüber dem Medium gehalten hätte (denn wegen der oben genannten Werte war $\varphi$ ja tatsächlich ein Test zum Niveau 0.03), dann hätte er die Anzahl der Versuche erhöhen sollen und dem Medium z. B. 40 Karten vorlegen sollen. ◇

Im Folgenden widmen wir uns der Frage nach Existenz und Konstruktion bester Tests.

## 9.2 Alternativtests

In diesem Abschnitt betrachten wir die besonders einfache Situation einer zweielementigen Parametermenge, sagen wir $\Theta=\{0,1\}$ mit einer Hypothese $\Theta_0=\{0\}$ und einer Alternative $\Theta_1=\{1\}$. Einelementige Hypothesen und Alternativen nennen wir *einfach*. Wir setzen voraus, dass die Summe der beiden Likelihood-Funktionen $\varrho_0$ und $\varrho_1$ für die betrachteten Wahrscheinlichkeitsmaße $\mathbb{P}_0$ bzw. $\mathbb{P}_1$ überall auf $\mathfrak{X}$ positiv ist, also $\varrho_0(x)+\varrho_1(x)>0$ für alle $x\in\mathfrak{X}$ (dies ist keine Einschränkung, denn anderenfalls verkleinern wir $\mathfrak{X}$ geeignet). Wir suchen einen (möglichst guten) Test $\varphi$ von $\mathbb{P}_0$ gegen $\mathbb{P}_1$ zu einem vorgegebenen Niveau $\alpha$.

Gemäß dem Maximum-Likelihood-Prinzip wird man sich für die Alternative $\mathbb{P}_1$ entscheiden, wenn $\varrho_1(x)$ hinreichend stark über $\varrho_0(x)$ dominiert, d. h. wenn der *Likelihood-Quotient*

$$R(x)=\begin{cases}\frac{\varrho_1(x)}{\varrho_0(x)}, & \text{falls } \varrho_0(x)>0,\\ \infty, & \text{falls } \varrho_0(x)=0<\varrho_1(x),\end{cases}$$

hinreichend groß ist. Der folgende, grundlegende Satz sagt, dass diese Intuition zu einem optimalen Testverfahren führt.

**Satz 9.2.1 (Neyman-Pearson-Lemma).** *In einem statistischen Modell $(\mathfrak{X}, \mathbb{P}_0, \mathbb{P}_1)$ mit einfacher Hypothese und einfacher Alternative gilt für jedes Irrtumsniveau $\alpha \in (0,1)$:*

*(a) Jeder beste Test $\psi$ von $\Theta_0 = \{0\}$ gegen $\Theta_1 = \{1\}$ zum Niveau $\alpha$ hat die Gestalt*

$$\psi(x) = \begin{cases} 1, & \text{falls } R(x) > c, \\ 0, & \text{falls } R(x) < c, \end{cases}$$

*für ein $c = c(\alpha) \geq 0$. Jeder solche Test heißt ein* Neyman-Pearson-Test.

*(b) Es gibt einen Neyman-Pearson-Test $\varphi$ mit $\mathbb{E}_0(\varphi) = \alpha$.*

*(c) Jeder Neyman-Pearson-Test $\varphi$ mit $\mathbb{E}_0(\varphi) = \alpha$ ist ein bester Test zum Niveau $\alpha$.*

**Beweis.** (a) Wir betrachten die Funktion $G^*\colon (0,1) \to [0,\infty)$,

$$G^*(\alpha) = \sup\bigl\{\mathbb{E}_1(\varphi)\colon \varphi \text{ ist ein Test mit } \mathbb{E}_0(\varphi) \leq \alpha\bigr\},$$

also die beim Niveau $\alpha$ bestenfalls erreichbare Macht. Es ist klar, dass $G^*$ monoton wächst, und als eine elementare Übungsaufgabe zeigt man, dass $G^*$ konkav ist. Ferner ist klar, dass $G^*(\mathbb{E}_0(\varphi)) \geq \mathbb{E}_1(\varphi)$ für jeden Test $\varphi$ gilt.

Sei $\psi$ ein bester Test zum Niveau $\alpha$, also $\mathbb{E}_1(\psi) = G^*(\alpha)$ und $\mathbb{E}_0(\psi) \leq \alpha$. Wegen der Monotonie von $G^*$ hat man also $G^*(\mathbb{E}_0(\psi)) \leq G^*(\alpha) = \mathbb{E}_1(\psi)$. Wegen Konkavität und Monotonie hat $G^*$ an der Stelle $\mathbb{E}_0(\psi)$ eine aufsteigende Tangente mit einer Steigung $c \geq 0$, d. h.

$$0 \geq G^*(s) - G^*\bigl(\mathbb{E}_0(\psi)\bigr) - c\bigl[s - \mathbb{E}_0(\psi)\bigr], \qquad s \in (0,1). \tag{9.2.1}$$

Wir zeigen nun, dass mit diesem $c$ der Test $\psi$ im Wesentlichen gleich dem Test $\varphi = \mathbb{1}_{\{R>c\}}$ ist und die angegebene Gestalt besitzt. Nutzen wir (9.2.1) für $s = \mathbb{E}_0(\varphi)$, so erhalten wir

$$\begin{aligned} 0 &\geq G^*\bigl(\mathbb{E}_0(\varphi)\bigr) - G^*\bigl(\mathbb{E}_0(\psi)\bigr) - c\bigl[\mathbb{E}_0(\varphi) - \mathbb{E}_0(\psi)\bigr] \\ &\geq \mathbb{E}_1(\varphi) - \mathbb{E}_1(\psi) - c\bigl[\mathbb{E}_0(\varphi) - \mathbb{E}_0(\psi)\bigr] \\ &= \mathbb{E}_1(\varphi - \psi) - c\mathbb{E}_0(\varphi - \psi) \\ &= \int_{\mathfrak{X}} f(x)\,\mathrm{d}x, \end{aligned}$$

wobei $f = (\varrho_1 - c\varrho_0)(\varphi - \psi)$. (Wie immer ist im diskreten Fall das Integral durch eine Summe zu ersetzen.) Aber es ist $f \geq 0$, denn für $\varrho_1 - c\varrho_0 \leq 0$ (also $R \leq c$) ist $\varphi - \psi = 0 - \psi \leq 0$, und für $\varrho_1 - c\varrho_0 > 0$ (also $R > c$) ist $\varphi - \psi = 1 - \psi \geq 0$. Daher haben die beiden Faktoren $\varrho_1 - c\varrho_0$ und $\varphi - \psi$ also das selbe Vorzeichen, und es gilt $f \geq 0$. Wegen $\int f(x)\,\mathrm{d}x \leq 0$ muss also (fast überall) $f = 0$ sein, also $\varphi(x) = \psi(x)$ für (fast) alle $x$ mit $R(x) \neq c$, was zu zeigen war. (Die mit dem Term *fast* angedeutete Ausnahmemenge ist statistisch bedeutungslos und kann ignoriert werden.)

(b) Es sei $c$ ein $\alpha$-Fraktil der Verteilung von $R$ unter $\mathbb{P}_0$ (siehe Definition 8.2.1), also $\mathbb{P}_0(R \geq c) \geq \alpha$ und $\mathbb{P}_0(R > c) \leq 1 - \alpha$. Insbesondere gilt $0 \leq \alpha - \mathbb{P}_0(R > c) \leq \mathbb{P}_0(R \geq c) - \mathbb{P}_0(R > c) = \mathbb{P}_0(R = c)$. Wir setzen nun

$$\gamma = \begin{cases} 0, & \text{falls } \mathbb{P}_0(R = c) = 0, \\ \frac{\alpha - \mathbb{P}_0(R>c)}{\mathbb{P}_0(R=c)}, & \text{falls } \mathbb{P}_0(R = c) > 0, \end{cases}$$

(beachte, dass $\gamma \in [0,1]$) und betrachten

$$\varphi(x) = \begin{cases} 1, & \text{falls } R(x) > c, \\ \gamma, & \text{falls } R(x) = c, \\ 0, & \text{falls } R(x) < c. \end{cases} \tag{9.2.2}$$

Dann ist $\varphi$ ein Neyman-Pearson-Test mit $\mathbb{E}_0(\varphi) = \mathbb{P}_0(R > c) + \gamma \mathbb{P}_0(R = c) = \alpha$.

(c) Sei $\varphi$ ein Neyman-Pearson-Test mit $\mathbb{E}_0(\varphi) = \alpha$ und Schwellenwert $c$ sowie $\psi$ ein beliebiger Test zum Niveau $\alpha$. Die Funktion $f = (\varrho_1 - c\varrho_0)(\varphi - \psi)$ aus dem ersten Beweisteil ist wiederum nichtnegativ, also folgt

$$0 \le \int_{\mathfrak{X}} f(x)\,\mathrm{d}x = \mathbb{E}_1(\varphi - \psi) - c\mathbb{E}_0(\varphi - \psi) \le \mathbb{E}_1(\varphi) - \mathbb{E}_1(\psi),$$

wobei im letzten Schritt benutzt wurde, dass $\mathbb{E}_0(\varphi - \psi) = \alpha - \mathbb{E}_0(\psi) \ge 0$. Also folgt $\mathbb{E}_1(\varphi) \ge \mathbb{E}_1(\psi)$, was zu zeigen war. □

**Bemerkung 9.2.2.** Aus dem Beweis von Teil (b) sieht man, dass die beiden Parameter $c \in [0,\infty)$ und $\gamma \in [0,1]$ in der Definition (9.2.2) eines besten Tests $\varphi$ eindeutig durch die Bedingung

$$\alpha = \mathbb{P}_0(R > c) + \gamma \mathbb{P}_0(R = c) \tag{9.2.3}$$

fest gelegt sind, denn aus ihr folgt, dass die Abbildung $c \mapsto \mathbb{P}(R > c)$ gerade im Punkt $c$ den Wert $\alpha$ übersteigt (also $\mathbb{P}(R > c) < \alpha$ und $\mathbb{P}(R \ge c) \ge \alpha$), und im Fall $\mathbb{P}(R = c) > 0$ liegt dann $\gamma$ auf Grund von (9.2.3) eindeutig fest (sonst ist $\gamma$ irrelevant). ◇

## 9.3 Beste einseitige Tests

Aus dem Neyman-Pearson-Lemma wissen wir also für einelementige Hypothesen und einelementige Alternativen, welche Gestalt optimale Tests haben. In diesem Abschnitt erweitern wir dieses Wissen auf Tests mit größeren Hypothesen und Alternativen unter der Voraussetzung, dass sie gewisse Monotonieeigenschaften aufweisen. Das erläutern wir zunächst an dem Beispiel 9.1.1.

**Beispiel 9.3.1 (Qualitätskontrolle).** Wie in Beispiel 9.1.1 betrachten wir das hypergeometrische Modell, also $\mathfrak{X} = \{0,\dots,n\}$, $\Theta = \{0,\dots,N\}$ und $\mathbb{P}_\vartheta = \mathrm{Hyp}_{n;\vartheta,N-\vartheta}$, wobei $n < N$. (Also haben wir $\varrho_\vartheta(x) = \binom{\vartheta}{x}\binom{N-\vartheta}{n-x}\binom{N}{n}^{-1}$ für $\max\{0, n-N+\vartheta\} \le x \le \min\{n,\vartheta\}$.) Wir wollen die Nullhypothese $\Theta_0 = \{0,\dots,\vartheta_0\}$ gegen die Alternative $\Theta_1 = \{\vartheta_0+1,\dots,N\}$ testen. Sei also $\alpha \in (0,1)$ gegeben. Mit einem beliebigen $\vartheta_1 \in \Theta_1$ wählen wir einen Neyman-Pearson-Test $\varphi$ zum Niveau $\alpha$ von $\{\vartheta_0\}$ gegen $\{\vartheta_1\}$mit $\mathbb{E}_{\vartheta_0}(\varphi) = \alpha$. Wir werden nun zeigen, dass $\varphi$ sogar ein gleichmäßig bester Test der gesamten Hypothese $\Theta_0$ gegen die gesamte Alternative $\Theta_1$ zum Niveau $\alpha$ ist.

Der Beweis beruht auf der folgenden Monotonieeigenschaft: Für $\vartheta' > \vartheta$ ist der Likelihood-Quotient $R_{\vartheta':\vartheta}(x) = \varrho_{\vartheta'}(x)/\varrho_\vartheta(x)$ eine monoton wachsende Funktion von $x$, und zwar sogar streng wachsend in ihrem Definitionsbereich. Diese Eigenschaft prüft man leicht, indem man ausschreibt:

$$R_{\vartheta':\vartheta}(x) = \prod_{k=\vartheta}^{\vartheta'-1} \frac{\varrho_{k+1}(x)}{\varrho_k(x)} = \prod_{k=\vartheta}^{\vartheta'-1} \frac{(k+1)(N-k-n+x)}{(k+1-x)(N-k)}, \qquad x \le \vartheta,$$

und $R_{\vartheta':\vartheta}(x) = \infty$ für $x > \vartheta$. Durch Differenziation sieht man leicht, dass jeder der Quotienten in $x$ streng wächst.

Dies geht nun wie folgt in die Eigenschaften des Tests $\varphi$ ein. Wegen der strengen Monotonie der Funktion $R_{\vartheta_1:\vartheta_0}$ ist die Unterscheidung $R_{\vartheta_1:\vartheta_0}(x) > c$ bzw. $< c$ äquivalent zu der Unterscheidung $x > \widetilde{c}$ bzw. $< \widetilde{c}$ für ein geeignetes $\widetilde{c}$. Also hat der Test $\varphi$ die Gestalt

$$\varphi(x) = \begin{cases} 1, & \text{falls } x > c, \\ \gamma, & \text{falls } x = c, \\ 0, & \text{falls } x < c, \end{cases}$$

mit geeigneten $c$ und $\gamma$. Wie in Bemerkung 9.2.2 (mit $R(x)$ ersetzt durch $x$) legt die Niveaubedingung

$$\mathrm{Hyp}_{n;\vartheta_0,N-\vartheta_0}\big(\{c+1,\dots,n\}\big) + \gamma \mathrm{Hyp}_{n;\vartheta_0,N-\vartheta_0}\big(\{c\}\big) = G_\varphi(\vartheta_0) = \alpha$$

$c$ und $\gamma$ sogar eindeutig fest. Also hängt $\varphi$ gar nicht von $\vartheta_1$ ab, und nach Satz 9.2.1 ist $\varphi$ ein bester Test von $\vartheta_0$ gegen jedes $\vartheta_1 \in \Theta_1$ zum Niveau $\alpha$, also ein gleichmäßig bester Test von $\{\vartheta_0\}$ gegen $\Theta_1$.

Wir müssen noch zeigen, dass $\varphi$ auch als Test von ganz $\Theta_0$ gegen $\Theta_1$ das Niveau $\alpha$ hat, also dass $G_\varphi(\vartheta) \le \alpha$ für alle $\vartheta \in \Theta_0$ gilt. Sei also $\vartheta \in \Theta_0 \setminus \{\vartheta_0\}$, also $\vartheta < \vartheta_0$. Da $R_{\vartheta_0:\vartheta}$ streng monoton wächst, ist $\varphi$ auch ein Neyman-Pearson-Test von $\vartheta$ gegen $\vartheta_0$, also nach Satz 9.2.1 ein bester Test zum Niveau $\beta = G_\varphi(\vartheta)$. Insbesondere ist er besser als der konstante Test $\psi \equiv \beta$. Dies bedeutet, dass $\alpha \ge G_\varphi(\vartheta_0) \ge G_\psi(\vartheta_0) = \beta = G_\varphi(\vartheta)$, und das war zu zeigen.

Zusammenfassend ergibt sich, dass das intuitiv naheliegende Verfahren tatsächlich optimal ist. Es sei noch bemerkt, dass man mit geeigneter Software etwa für die in Beispiel 9.1.1 angegebenen Werte (also $n = 50$, $N = 10\,000$ und $\vartheta_0 = 500$) für $\alpha = 0.025$ die optimalen Werte von $c$ und $\gamma$ als $c = 6$ und $\gamma = 0.52$ bestimmen kann. ◇

Die mathematische Essenz des Beispiels 9.3.1 ist die folgende.

**Definition 9.3.2.** *Wir sagen, ein statistisches Modell $(\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit $\Theta \subset \mathbb{R}$ hat* wachsende Likelihood-Quotienten *bezüglich einer Statistik $T\colon \mathfrak{X} \to \mathbb{R}$, wenn für alle $\vartheta < \vartheta'$ der Dichtequotient $R_{\vartheta':\vartheta} = \varrho_{\vartheta'}/\varrho_\vartheta$ eine wachsende Funktion von $T$ ist, also wenn es eine monoton wachsende Funktion $f_{\vartheta':\vartheta}\colon \mathbb{R} \to [0,\infty)$ gibt mit $R_{\vartheta':\vartheta} = f_{\vartheta':\vartheta} \circ T$.*

**Beispiel 9.3.3 (Exponentielle Modelle).** Jedes einparametrige exponentielle Modell (siehe Definition 7.5.5) hat strikt wachsende Likelihood-Quotienten. Denn aus der Form (7.5.2) für $\varrho_\vartheta$ folgt für $\vartheta < \vartheta'$, dass

$$R_{\vartheta':\vartheta} = e^{[a(\vartheta')-a(\vartheta)]T+[b(\vartheta)-b(\vartheta')]},$$

und die Koeffizientenfunktion $\vartheta \mapsto a(\vartheta)$ ist nach Voraussetzung entweder strikt wachsend auf $\Theta$ oder strikt fallend auf $\Theta$. Im ersten Fall ist dann $a(\vartheta') - a(\vartheta) > 0$, also $R_{\vartheta':\vartheta}$ eine strikt wachsende Funktion von $T$, und im zweiten Fall ist $R_{\vartheta':\vartheta}$ eine strikt wachsende Funktion der Statistik $-T$. ◇

Nun können wir Beispiel 9.3.1 auf alle Modelle mit wachsenden Likelihood-Quotienten verallgemeinern und damit auch Satz 9.2.1 erweitern auf mehrfache Hypothesen und mehrfache Alternativen, zumindest im monotonen Fall:

**Satz 9.3.4 (Einseitiger Test bei monotonen Likelihood-Quotienten).** *Sei $(\mathfrak{X}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit wachsenden Likelihood-Quotienten bezüglich einer Statistik $T$. Ferner seien $\vartheta_0 \in \Theta$ und $\alpha \in (0,1)$. Dann existiert ein gleichmäßig bester Test $\varphi$ zum Niveau $\alpha$ für das einseitige Testproblem $H_0 : \vartheta \le \vartheta_0$ gegen $H_1 : \vartheta > \vartheta_0$. Dieser Test hat die Gestalt*

$$\varphi(x) = \begin{cases} 1, & \text{falls } T(x) > c, \\ \gamma, & \text{falls } T(x) = c, \\ 0, & \text{falls } T(x) < c, \end{cases} \tag{9.3.1}$$

*wobei sich $c$ und $\gamma$ aus der Bedingung $G_\varphi(\vartheta_0) = \alpha$ ergeben. Die Gütefunktion $G_\varphi$ is monoton wachsend.*

**Beweis.** Falls die Likelihood-Quotienten sogar streng monoton in $T$ sind, so überträgt man leicht die Argumentation aus Beispiel 9.3.1.

Im Fall der einfachen Monotonie konstruiert man zunächst einen Test $\varphi$ der Form (9.3.1) mit $G_\varphi(\vartheta_0) = \alpha$. Dies geht wie im Beweis des Neyman-Pearson-Lemmas (Satz 9.2.1): Man wählt $c$ als $\alpha$-Fraktil der Verteilung von $T$ unter $\mathbb{P}_{\vartheta_0}$. Für $\vartheta < \vartheta'$ gilt dann wegen der Monotonie der Likelihood-Quotienten: Falls $R_{\vartheta':\vartheta} = f_{\vartheta':\vartheta} \circ T > f_{\vartheta':\vartheta}(c)$, so folgt $T > c$ und daher $\varphi = 1$. Analog gilt im Fall $R_{\vartheta':\vartheta} < f_{\vartheta':\vartheta}(c)$ notwendigerweise $\varphi = 0$. Somit ist $\varphi$ ein Neyman-Pearson-Test von $\vartheta$ gegen $\vartheta'$, und man kann wieder wie im Beispiel 9.3.1 fortfahren. □

**Beispiel 9.3.5 (Einseitiger Gaußtest).** Auf Grund von $n$ unabhängigen Messungen soll fest gestellt werden, ob die Sprödigkeit eines Kühlrohres unter einem zulässigen Grenzwert $m_0$ liegt. Wir nehmen an, dass die Messungen $x_1, \dots, x_n$ mit vielen kleinen, zufälligen unabhängigen Fehlern behaftet sind, deren Gesamteinfluss sich zu einer Normalverteilung der $x_i$ auswirkt. Die Varianz $v^2 > 0$ hängt aber im Wesentlichen nur von der Präzision des Messgerätes ab und sei hier als bekannt voraus gesetzt. Nun möchten wir also testen, ob der Erwartungswert $m$ unter $m_0$ liegt oder nicht.

Wir betrachten also das $n$-fache Gaußsche Produktmodell $(\mathbb{R}^n, (\mathbb{P}_m)_{m\in\mathbb{R}})$, wobei $\mathbb{P}_m = \mathcal{N}(m, v)^{\otimes n}$ das $n$-fache Produktmaß der Normalverteilung mit Erwartungswert $m$ und (fester) Varianz $v^2 > 0$ sei. Wir wollen die Hypothese $H_0 : m \le m_0$ gegen die Alternative $H_1 : m > m_0$ testen. Man sieht leicht, dass für $m' > m$ der Likelihood-Quotient gegeben ist als

$$\begin{aligned} R_{m':m}(x) &= \exp\Big\{-\frac{1}{2v}\sum_{i=1}^n \Big((x_i - m')^2 - (x_i - m)^2\Big)\Big\} \\ &= \exp\Big\{-\frac{n}{2v}\Big(2(m - m')M(x) + (m')^2 + m^2\Big)\Big\}, \end{aligned}$$

wobei $M(x) = \frac{1}{n}\sum_{i=1}^n x_i$ wie üblich das Stichprobenmittel ist. Also ist $R_{m':m}$ eine strikt wachsende Funktion der Statistik $M$, und wir sind in der Situation von Satz 9.3.4. Da $\mathbb{P}_m$ eine Dichte hat, also auch die Verteilung von $M$ unter $\mathbb{P}_m$, ist $\mathbb{P}_m(M = c) = 0$ für alle $c$, also können wir die mittlere Zeile in (9.3.1) vernachlässigen. Der Wert von $c$ für den optimalen Test $\varphi$ in (9.3.1)

ergibt sich aus der Bedingung $\alpha = G_\varphi(m_0)$ wie folgt:

$$\begin{aligned} G_\varphi(m_0) = \mathbb{P}_{m_0}(M > c) &= \mathcal{N}(0,v)^{\otimes n}\Big(\sum_{i=1}^n X_i > (c-m_0)n\Big) \\ &= \mathcal{N}(0, v\sqrt{n})\Big(\big((c-m_0)n, \infty\big)\Big) = \mathcal{N}(0,1)\Big(\Big(\frac{(c-m_0)n}{v\sqrt{n}}, \infty\Big)\Big) \\ &= 1 - \Phi\Big(\frac{(c-m_0)\sqrt{n}}{v}\Big), \end{aligned}$$

da ja $\sum_{i=1}^n X_i$ die $\mathcal{N}(0, v\sqrt{n})$-Verteilung besitzt, wenn $X_1, \dots, X_n$ unabhängige $\mathcal{N}(0,v)$-verteilte Zufallsgrößen sind; siehe Lemma 5.3.8 (wie üblich ist $\Phi$ die Verteilungsfunktion von $\mathcal{N}(0,1)$). Die Gleichung $\alpha = G_\varphi(m_0)$ kann man nun nach $c$ auflösen und erhält

$$c = m_0 + \frac{v}{\sqrt{n}}\Phi^{-1}(1-\alpha).$$

Also ergibt Satz 9.3.4, dass der Ablehnungsbereich des optimalen Tests gleich der Menge

$$\Big\{x \in \mathbb{R}^n \colon M(x) > m_0 + \frac{v}{\sqrt{n}}\Phi^{-1}(1-\alpha)\Big\}$$

ist. Wenn also der Mittelwert der Stichprobe über $m_0 + \sqrt{v^2/n}\Phi^{-1}(1-\alpha)$ liegt, sollte man davon ausgehen, dass die Sprödigkeit über der Grenze $m_0$ liegt, also den Standard nicht erfüllt; anderenfalls sollte man vielleicht weitere Tests durchführen. ◇

**Beispiel 9.3.6 (Einseitiger Chiquadrattest).** Um die genetische Variabilität einer Getreidesorte zu ermitteln, soll auf Grund von $n$ unabhängigen Beobachtungen getestet werden, ob die Varianz der Halmlänge einen Mindestwert $v_0$ überschreitet. Wieder machen wir die Annahme, dass die Messungen vielen kleinen zufälligen unabhängigen Einflüssen unterliegen, die grob mit einer Normalverteilung angenähert werden können. Allerdings ist es plausibel anzunehmen, dass sich diese Einflüsse auf *multiplikative* Weise auf die Halmlänge auswirken, also auf additive Weise auf den Logarithmus der Halmlänge. Also nehmen wir an, dass die Logarithmen der Halmlängen normalverteilt sind mit einem bekannten Erwartungswert $m$ und einer unbekannten Varianz $v^2 > 0$. Als Modell wählen wir also das $n$-fache Produktmodell $(\mathbb{R}^n, (\mathbb{P}_v)_{v>0})$, wobei $\mathbb{P}_v = \mathcal{N}(m,v)^{\otimes n}$ das $n$-fache Produkt der Normalverteilung mit Erwartungswert $m$ und Varianz $v^2$ ist. Es soll die Hypothese $H_0 : v \geq v_0$ gegen die Alternative $H_1 : v < v_0$ getestet werden.

In direkter Verallgemeinerung des Beispiels 7.5.10 sieht man, dass $(\mathbb{P}_v)_{v>0}$ eine exponentielle Familie bezüglich der Statistik $T = \sum_{i=1}^n (X_i - m)^2$ ist. Hier sind nun die Seiten der Hypothese und Alternative gegenüber Satz 9.3.4 mit einander vertauscht, aber nach Multiplikation mit $-1$ ist man in der Situation von Satz 9.3.4 nach Übergang von '$<$' zu '$>$' und von $T$ zu $-T$. Wegen Beispiel 9.3.3 ist also Satz 9.3.4 anwendbar, und der optimale Test $\varphi$ hat bei gegebenem Niveau $\alpha$ also den Verwerfungsbereich

$$\Big\{x \in \mathbb{R}^n \colon \sum_{i=1}^n (X_i - m)^2 < v_0 t\Big\},$$

wobei $t$ das $\alpha$-Quantil der $\boldsymbol{\chi}^2_n$-Verteilung ist (also $\boldsymbol{\chi}^2_n([0,t]) = \alpha$; siehe Definition 8.4.3). Dies sieht man aus der Rechnung

$$G_\varphi(v_0) = \mathbb{P}_{v_0}(T < c) = \mathcal{N}(0,v_0)^{\otimes n}\Big(\sum_{i=1}^n X_i^2 < c\Big) = \mathcal{N}(0,1)^{\otimes n}\Big(\sum_{i=1}^n X_i^2 < \tfrac{c}{v_0}\Big) = \boldsymbol{\chi}^2_n\big([0, \tfrac{c}{v_0}]\big),$$

wobei wir uns daran erinnerten, dass $\sum_{i=1}^{n} X_i^2$ die Chiquadrat-Verteilung hat, wenn $X_1, \ldots, X_n$ unabhängig und $\mathcal{N}(0,1)$-verteilt sind. Also ist die Bedingung $\alpha = G_\varphi(v_0)$ äquivalent zu $c = v_0 t$ für den optimalen Test $\varphi$ aus (9.3.1), und der oben angegebene Verwerfungsbereich ist gleich der Menge $\{x\colon \varphi(x) = 1\} = \{x\colon T(x) < c\}$. $\diamond$

# Bibliographie


[Ba02] H. Bauer, *Wahrscheinlichkeitstheorie*, 5. Auflage, Walter de Gruyter, Berlin–New York, 2002.

[Ch78] K. L. Chung, *Elementare Wahrscheinlichkeitstheorie und stochastische Prozesse*, Springer, Berlin, 1978.

[Fe68] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. I, 3rd ed., Wiley, New York, 1968.

[Ge02] H.-O. Georgii, *Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik*, Walter de Gruyter, 2002.

[Kr02] U. Krengel, *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, 6. Auflage, Vieweg, 2002.

[Pe98] W.R. Pestman, *Mathematical Statistics*, Walter de Gruyter, Berlin–New York, 1998.

[Sch98] K. Schürger, *Wahrscheinlichkeitstheorie*, Oldenbourg, München, 1998.

[St94] D. Stirzaker, *Elementary Probability*, Cambridge University Press, Cambridge, 1994.