

Vorlesungsskript  
Einführung in die Wahrscheinlichkeitstheorie  
und Mathematische Statistik

*Wintersemester 2020/21*  
Friedrich-Schiller-Universität Jena

Michael H. Neumann

Dieses Skript ist nur für den persönlichen Gebrauch bestimmt. Unerlaubte Weitergabe oder Veröffentlichung jeder Art ist streng untersagt.

## Wichtige Informationen

- Vorlesungszeit: 02.11.2020 – 12.02.2021
  - Bis zu einer eventuellen Aufnahme des Präsenzunterrichts erhalten Sie wöchentlich einen Teil des Vorlesungsskriptes. Das Material wird unter folgendem Link abgelegt:  
<https://users.fmi.uni-jena.de/~jschum/lehre/lectures.php?name=Neumann>
  - In den jeweiligen Vorlesungswochen werden spätestens donnerstags die zu bearbeitenden Übungsaufgaben unter dem Link  
<https://users.fmi.uni-jena.de/~jschum/lehre/lectures.php?name=Neumann> bereitgestellt. Die schriftlichen Lösungen sind in der jeweils folgenden Woche bis Mittwoch 10:00 Uhr beim entsprechenden Übungsleiter abzugeben oder in elektronischer Form zu übermitteln. Die Lösungen werden durch die jeweiligen Übungsleiter korrigiert und in der Übung in der darauffolgenden Woche besprochen.
  - Die **Zulassung zur Klausur** wird erteilt, wenn mindestens **50% der möglichen Punkte** in den Übungsreihen erreicht werden.
- Klausurtermin:           02.03.2021, 12-15 Uhr, Hörsaal 2, Carl-Zeiß-Str. 3  
 Wiederholungsklausur: 08.04.2021, 12-15 Uhr
- Kontakt
  - Email: michael.neumann@uni-jena.de
  - Übungsleiter:
    - \* jamal.drewlo@uni-jena.de       (Gruppe 1, Mi 10-12)
    - \* daniel.max@uni-jena.de       (Gruppe 2, Do 10-12)
    - \* engelhardt.stefan@uni-jena.de   (Gruppe 3, Do 14-16)

## Literatur

- [1] Behrends, E. (2013). *Elementare Stochastik: Ein Lernbuch - von Studierenden mitentwickelt*. Springer Spektrum, Wiesbaden.
- [2] Hesse, Chr. (2003). *Angewandte Wahrscheinlichkeitstheorie*. Vieweg, Braunschweig/Wiesbaden.
- [3] Krengel, U. (2005). *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. 8., erweiterte Auflage. Vieweg, Wiesbaden.

### Alternative:

Stöbern Sie doch einfach in der Mathebibliothek nach Büchern mit der Signatur MAT:PA::1000:....

# Inhalt

1	Einführende Beispiele	4
2	Mathematische Modellierung von Zufallsexperimenten	10
3	Elementare bedingte Wahrscheinlichkeiten	21
4	Stochastische Unabhängigkeit	27
5	Zufallsvariable	31
6	Gesetze der großen Zahlen	49
7	Wahrscheinlichkeitsmaße mit Dichten	59
8	Approximationen der Binomialverteilung	76
9	Grundbegriffe der Mathematischen Statistik	85

# 1 Einführende Beispiele

Die Wahrscheinlichkeitstheorie beschäftigt sich mit der Beschreibung von zufälligen Phänomenen und entsprechenden Gesetzmäßigkeiten. Doch was ist eigentlich „Zufall“? Einer Reihe von Ereignissen würde man sicherlich ohne zu zögern das Prädikat „zufällig“ zugehen, z.B.

- Augenzahlen beim Würfeln
- Ergebnis (Zahl oder Wappen) beim Münzwurf
- Lottozahlen

Aber auch andere Ereignisse werden durch Zufallsmodelle beschrieben:

- verbleibende Lebenszeit eines Individuums oder die Zeit bis zum Ausfall eines technischen Gerätes
- Börsenkurse
- Erfolg/Misserfolg einer medizinischen Behandlungsmethode

An dieser Stelle scheint es nun weniger klar zu sein, wie weit der Begriff des Zufalls geht. Die Schwankungen von Börsenkursen können im Rückblick oft auf konkrete Ursachen zurückgeführt werden. Positive Unternehmensnachrichten (gestiegene Gewinne, höhere Verkaufszahlen,...) führen so zu steigenden Kursen der Aktien des entsprechenden Unternehmens. Auch wenn eine Behandlungsmethode bei verschiedenen Patienten zu unterschiedlichen Ergebnissen führt, so kann doch der jeweilige Behandlungserfolg auf die Wirkung vieler Aspekte (Alter, allgemeiner Gesundheitszustand,...) zurückgeführt werden. Was nun rechtfertigt die Verwendung des Begriffs „Zufall“? Charakteristisch ist, dass sich die Ergebnisse (Augenzahlen, zukünftige Aktienkurse,...) nicht deterministisch vorausbestimmen lassen. Hierbei ist unerheblich, ob „reiner Zufall“ vorliegt oder ob ein Ereignis aufgrund einer zu komplexen Ursache-Wirkungsbeziehung nicht exakt vorhergesagt werden kann. Ob überhaupt „reiner Zufall“ existiert ist eher eine philosophische Frage, deren Beantwortung an dieser Stelle nicht nachgegangen wird. Vielmehr soll festgehalten werden, dass Zufallsmodelle auf vielen Gebieten seit langer Zeit erfolgreich eingesetzt werden, z.B. bei der Beschreibung der Bewegung von Aktienkursen. Ebenfalls charakteristisch ist, dass zufällige Phänomene gewissen Gesetzmäßigkeiten folgen. Beim oftmaligen, wiederholten Würfeln wird man sicher feststellen, dass der Anteil der Würfe mit der Augenzahl 6 gegen  $1/6$  konvergiert. Ein mathematisch rigoroser Beweis dieser Gesetzmäßigkeit wird im Rahmen dieser Vorlesung geführt werden.

Bevor wir nun einige einfache Beispiele für zufällige Ereignisse und entsprechende mathematische Modelle betrachten, soll noch kurz der für diese Vorlesung zentrale Begriff der Wahrscheinlichkeit erwähnt werden. Unter der **Wahrscheinlichkeit** eines gewissen Ereignisses verstehen wir ein **Maß für die Bestimmtheit des Eintretens dieses Ereignisses**. Dem **sicheren Ereignis** (z.B. „Ergebnis des Würfels liegt zwischen 1 und 6“) wird hierbei die Wahrscheinlichkeit 1 zugeordnet, dem **unmöglichen Ereignis** (z.B. „Augenzahl 7“) die Wahrscheinlichkeit 0. Man beachte dass, anders als in der Umgangssprache,

Wahrscheinlichkeiten durch Zahlen zwischen 0 und 1 und nicht in Prozenten angegeben werden. Diese Festlegung hat ihren Ursprung darin, dass die Wahrscheinlichkeitstheorie auf der sogenannten Maßtheorie basiert, wobei Teilmengen eines gewissen Grundraumes nichtnegative reelle Zahlen zugeordnet werden.

Im Folgenden betrachten wir einige Beispiele von einfachen Zufallsexperimenten. Neben dem Wecken von etwas Intuition soll dabei auch schrittweise die in der Wahrscheinlichkeitstheorie übliche Terminologie eingeführt werden.

**Beispiel 1.1.** Würfeln bei „Die Siedler von Catan“

Dieses Brettspiel wurde im Jahr 1995 sowohl mit dem Kritikerpreis „Spiel des Jahres“ als auch mit dem „Deutschen Spielepreis“ ausgezeichnet. Bei diesem Spiel würfeln die Spieler reihum jeweils gleichzeitig mit zwei Zahlenwürfeln sowie einem weiteren Würfel. Für den Fortgang des Spiels ist unter anderem die Augensumme der beiden Zahlenwürfel relevant. Um zu Spielbeginn eine strategisch sinnvolle Entscheidung zu treffen, sollten die Spieler die Wahrscheinlichkeiten für das Vorkommen der möglichen Augensummen in Betracht ziehen. Wir wollen einen Weg skizzieren, wie man beispielsweise die Wahrscheinlichkeit, dass diese Augensumme 10, ist bestimmen kann.

**Lösung:** Auch wenn die Antwort auf die Frage mit etwas Intuition leicht gefunden werden kann, werden wir uns im Hinblick auf kompliziertere Aufgabenstellungen die Lösung systematisch erarbeiten. Zunächst beschreiben wir die möglichen Versuchsausgänge (sogenannte **Elementarereignisse**) in einer geeigneten Form. Dabei werden wir abstrahieren und unwesentliche Details (z.B. exakte Lage der Würfel nach dem Werfen, Ergebnis des dritten Würfels,...) nicht berücksichtigen. So könnte man folgende Mengen zur Beschreibung der möglichen Versuchsausgänge ansetzen:

$$\begin{aligned}\Omega^{(1)} &= \{2, 3, \dots, 12\} && \text{(Augensummen),} \\ \Omega^{(2)} &= \{(1, 1), \dots, (1, 6), (2, 2), \dots, (2, 6), \dots, (6, 6)\} && \text{(kleinere/größere Augenzahl).}\end{aligned}$$

Nun soll die gesuchte Wahrscheinlichkeit des Ereignisses „die Augensumme ist gleich 10“ bestimmt werden.

- 1) Wir versuchen,  $\Omega^{(1)}$  als Beschreibung der möglichen Versuchsausgänge heranzuziehen. Es ist zu vermuten, dass beispielsweise die Augensumme 10 mit größerer Wahrscheinlichkeit eintritt als die 12. Dies ist heuristisch klar, denn die Augensumme 12 tritt nur auf, wenn beide Zahlenwürfel 6 anzeigen. Dagegen ist die Augensumme 10 auf verschiedene Weise erreichbar. (Diese Heuristik wird übrigens bei der Gestaltung von entsprechenden Spielkärtchen dadurch unterstützt, dass die 12 wie auch die 2 kleiner gedruckt sind als beispielsweise die 8 oder die 6.) Trotz der hier beschriebenen Heuristik scheint eine exakte Bestimmung der gesuchten Wahrscheinlichkeit nicht trivial zu sein.
- 2) Die Menge  $\Omega^{(2)}$  bietet eine etwas detailliertere Beschreibung der möglichen Versuchsausgänge. Das interessierende Ereignis „10“ tritt ein, falls beide Würfel 5 anzeigen oder falls die kleinere Augenzahl 4 und die größere 6 ist. Da sich (5, 5) und (4, 6) gegenseitig ausschließen, so gilt für die gesuchte Wahrscheinlichkeit (eine axiomatische

Begründung folgt später), dass

$$P(\text{„Augensumme } 10\text{“}) = P(\text{„(5,5)“}) + P(\text{„(4,6)“}).$$

Hier wurde der Buchstabe  $P$  für die entsprechenden Wahrscheinlichkeiten (*probability*) verwendet. Wenn wir nun die Wahrscheinlichkeiten für das Auftreten der Augenpaare (5,5) und (4,6) kennen, so können wir mit obiger Gleichung die gesuchte Wahrscheinlichkeit berechnen. Ohne jegliche Erfahrung auf diesem Gebiet kann man aber bei der Bestimmung der Wahrscheinlichkeiten für diese Augenpaare dennoch fehlgreifen. Einsichtig sollte noch sein, dass die Wahrscheinlichkeit des Augenpaares (4,6) größer ist als die von (5,5). Wie gesagt, eine exakte Bestimmung kann immer noch unklar sein.

Ein Ausweg aus diesem Dilemma wird wiederum durch die Gestaltung des Spielmaterials bei „Die Siedler von Catan“ suggeriert. Zumindest in der Version aus dem vorigen Jahrtausend waren die beiden Zahlenwürfel verschiedenfarbig (rot bzw. holzfarben). Dementsprechend könnte man die möglichen Versuchsausgänge auch durch die folgende Menge beschreiben:

$$\Omega^{(3)} = \{(1, 1), \dots, (1, 6), (2, 1), \dots, (2, 6), \dots, (6, 1), \dots, (6, 6)\} \quad (\text{roter/holzf. Würfel}).$$

An dieser Stelle erscheint die Annahme gerechtfertigt, dass jeder der möglichen 36 Versuchsausgänge dieselbe Wahrscheinlichkeit besitzt. Eine Rechtfertigung dieser Annahme wird sich nach Einführung des Begriffes der stochastischen Unabhängigkeit ergeben. Für den Moment stelle man sich vor, dass die beiden Zahlenwürfel nicht gleichzeitig, sondern nacheinander geworfen werden und denke an die in der Schule verbreiteten Baumdiagramme. Unter dieser Annahme ergibt sich nun für die gesuchte Wahrscheinlichkeit

$$P(\text{„Augensumme } 10\text{“}) = \underbrace{P(\text{„(4,6)“})}_{=1/36} + \underbrace{P(\text{„(5,5)“})}_{=1/36} + \underbrace{P(\text{„(6,4)“})}_{=1/36} = \frac{1}{12}.$$

Der „Trick“ bestand darin, dass wir die möglichen Versuchsausgänge so beschrieben haben, dass diese jeweils (mehr oder weniger) offensichtlich die gleiche Wahrscheinlichkeit besitzen. Die Darstellung durch ein Zufallsexperiment mit gleichwahrscheinlichen Versuchsausgängen hat hier tatsächlich bei der Bestimmung der gesuchten Wahrscheinlichkeit geholfen und wird auch bei weiteren Problemstellungen hilfreich sein. Wir formalisieren dies durch die folgende Definition.

**Definition 1.2.** *Ein Zufallsexperiment möge  $\Omega = \{\omega_1, \dots, \omega_N\}$  als Menge der möglichen Versuchsausgänge besitzen und es gelte*

$$P(\{\omega\}) = \frac{1}{\#\Omega} = \frac{1}{N} \quad \forall \omega \in \Omega.$$

*Ein solches Experiment heißt **Laplace-Experiment**.*

Der aufmerksame Leser wird bemerkt haben, dass sich hier eine etwas gestelzt wirkende Bezeichnung eingeschlichen hat. Die Wahrscheinlichkeit, dass der Versuchsausgang  $\omega$

eintritt wird mit  $P(\{\omega\})$  und nicht mit  $P(\omega)$  bezeichnet. Diese in der Mathematik übliche Schreibweise wird gewählt, da im Allgemeinen Wahrscheinlichkeiten für **Mengen** von möglichen Versuchsausgängen angegeben werden, also beispielsweise für  $\{\omega_1, \omega_2, \omega_3\}$  oder eben auch für  $\{\omega_1\}$ . Dies wird auch im Zusammenhang mit dem später definierten Begriff eines **Wahrscheinlichkeitsmaßes** klar.

**Beispiel 1.3.** Urnenmodell I: Ziehen **ohne** Zurücklegen

Wir betrachten das folgende, zunächst verbal beschriebene Zufallsexperiment:

- In einer Urne befinden sich  $s$  schwarze und  $w$  weiße, ansonsten in ihrer Beschaffenheit gleichartige Kugeln.
- Aus dieser Urne werden nacheinander  $n$  Kugeln „rein zufällig“ und ohne Zurücklegen gezogen ( $n \leq s + w =: N$ ). (Wie vermutlich auch im Schulunterricht geschehen, wird das Zufallsexperiment verbal und ohne rigorose mathematische Modellierungen beschrieben. Mit der auf dieser Stufe üblichen Formulierung „rein zufällig“ ist gemeint, dass jede der verfügbaren Kugeln mit der gleichen Wahrscheinlichkeit gezogen wird.)

Gesucht ist die Wahrscheinlichkeit, dass genau  $k$  schwarze Kugeln gezogen werden.

**Lösung:** Wir beginnen wiederum mit Suche nach einer adäquaten Beschreibung der möglichen Versuchsausgänge. Naheliegend wären:

$$\begin{aligned}\Omega^{(1)} &= \{0, \dots, n\} && \text{(Anzahl der gezogenen schwarzen Kugeln)} \\ \Omega^{(2)} &= \{(\omega_1, \dots, \omega_n) : \omega_i \in \{0, 1\}\} && \text{(„1“ entspricht schwarz, „0“ weiß)}.\end{aligned}$$

Die obigen Mengen können ggf. verkleinert werden, z.B. kann  $\Omega^{(1)}$  ersetzt werden durch  $\{(n-w) \vee 0, \dots, s \wedge n\}$ . Die angegebene Wahl von  $\Omega^{(1)}$  ist aber ebenso möglich; ggf. gilt  $P(\{\omega\}) = 0$  für gewisse  $\omega \in \Omega^{(1)}$ . Wichtiger ist jedoch, dass die durch  $\Omega^{(1)}$  bzw.  $\Omega^{(2)}$  vorgenommenen Beschreibungen der möglichen Versuchsausgänge für die Bestimmung der gesuchten Wahrscheinlichkeit nicht hilfreich sind. Stattdessen benutzen wir wiederum den „Kunstgriff“ der Zurückführung auf ein Laplace-Experiment.

**Lösung:** Wir **nummerieren** (zumindest gedanklich) die schwarzen Kugeln mit  $1, \dots, s$  und die weißen mit  $s+1, \dots, N$  durch. Die Menge der möglichen Versuchsausgänge lässt sich dann so beschreiben:

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

An dieser Stelle kann man sich leicht zu einer adäquat scheinenden Bestimmung der Wahrscheinlichkeiten durchringen: Aus „**Symmetriegründen**“ setzt man nun passenderweise voraus, dass

$$P(\{\omega\}) = \frac{1}{\#\Omega} = \frac{1}{N(N-1)\cdots(N-n+1)} \quad \forall \omega \in \Omega.$$

Das interessierende Ereignis stellen wir nun als **Teilmenge** von  $\Omega$  dar:

$$\begin{aligned} E_k &:= \text{„Es werden genau } k \text{ schwarze Kugeln gezogen“} \\ &= \left\{ \omega \in \Omega : \sum_{i=1}^n \mathbb{1}(\omega_i \leq s) = k \right\}. \end{aligned}$$

Hierbei bezeichnen wir mit  $\mathbb{1}(\cdot)$  die Indikatorfunktion, welche im obigen Fall den Wert 1 annimmt, falls  $\omega_i \leq s$  ist und andernfalls den Wert 0 annimmt. Es gilt

$$\#E_k = \binom{n}{k} s(s-1) \cdots (s-k+1) w(w-1) \cdots (w-(n-k)+1),$$

wobei der Binomialkoeffizient  $\binom{n}{k}$  durch

$$\binom{n}{k} = \begin{cases} \frac{n!}{k!(n-k)!}, & \text{falls } 0 \leq k \leq n, \\ 0 & \text{sonst} \end{cases}$$

gegeben ist. Damit erhalten wir schließlich für die gesuchte Wahrscheinlichkeit

$$\begin{aligned} P(E_k) &= \sum_{\omega \in E_k} P(\{\omega\}) \\ &= \frac{1}{N(N-1) \cdots (N-n+1)} \#E_k \\ &= \frac{n!}{N(N-1) \cdots (N-n+1)} \frac{s(s-1) \cdots (s-k+1)}{k!} \frac{w(w-1) \cdots (w-(n-k)+1)}{(n-k)!} \\ &= \frac{\binom{s}{k} \binom{w}{n-k}}{\binom{N}{n}}. \end{aligned}$$

Ein entsprechendes „Verteilungsgesetz“ (Später!) wird in der Wahrscheinlichkeitstheorie **hypergeometrische Verteilung** genannt.

#### **Beispiel 1.4.** Urnenmodell II: Ziehen **mit** Zurücklegen

Wie beim vorangegangenen Beispiel wird vorausgesetzt, dass sich in einer Urne  $s$  schwarze,  $w$  weiße, ansonsten aber gleichartige Kugeln befinden. Es wird  $n$ -mal „rein zufällig“ eine Kugel gezogen, welche sofort wieder in die Urne zurückgelegt wird. Es soll wiederum die Wahrscheinlichkeit bestimmt werden, dass dabei genau  $k$ -mal eine schwarze Kugel gezogen wird.

**Lösung:** Wir stellen dieses Zufallsexperiment wiederum in der Form eines Laplaceexperimentes dar. Wir stellen uns vor, dass die schwarzen Kugeln mit  $1, \dots, s$  und die weißen mit  $s+1, \dots, N$  durchnummeriert sind. Die Menge der möglichen Versuchsausgänge lässt sich nun so beschreiben:

$$\Omega = \{(\omega_1, \dots, \omega_n) : \omega_i \in \{1, \dots, N\}\}.$$

An dieser Stelle sollte zumindest heuristisch nachvollziehbar sein, dass jedes mögliche  $n$ -Tupel mit der gleichen Wahrscheinlichkeit auftritt, also

$$P(\{\omega\}) = \frac{1}{\#\Omega} = \frac{1}{N^n} \quad \forall \omega \in \Omega.$$

(Eine fundierte Begründung dieser Annahme wird erst möglich, nachdem der Begriff der „stochastischen Unabhängigkeit“ geklärt wurde.) Das interessierende Ereignis kann als Teilmenge von  $\Omega$  beschrieben werden,

$$\begin{aligned} E_k &:= \text{„Es werden genau } k \text{ schwarze Kugeln gezogen“} \\ &= \left\{ \omega \in \Omega : \sum_{i=1}^n \mathbb{1}(\omega_i \leq s) = k \right\}, \end{aligned}$$

und es gilt

$$\#E_k = \binom{n}{k} s^k w^{n-k}.$$

Damit ergibt sich für das interessierende Ereignis die folgende Wahrscheinlichkeit:

$$P(E_k) = \frac{\#E_k}{\#\Omega} = \binom{n}{k} \left(\frac{s}{N}\right)^k \left(1 - \frac{s}{N}\right)^{n-k} \quad \text{für } k = 0, 1, \dots, n.$$

**Beispiel 1.5.** Verallgemeinerung von Beispiel 1.4

Wir nehmen an, dass ein Zufallsexperiment  $n$ -mal hintereinander durchgeführt wird, wobei nur das Eintreten von „Erfolg“ interessiert und die einzelnen Versuchsausgänge sich nicht beeinflussen. Die jeweilige Erfolgswahrscheinlichkeit bei einem Teilversuch sei  $p \in [0, 1]$ . Es ist die Wahrscheinlichkeit gesucht, dass genau  $k$  Erfolge eintreten.

**Lösung:** An dieser Stelle bietet sich die Betrachtung eines adäquaten Laplaceexperimentes nicht direkt an. Falls jedoch  $p$  eine rationale Zahl wäre,  $p = s/N$  für ganzzahlige  $s$ ,  $N$  mit  $0 \leq s \leq N$ , so könnte man stattdessen das Urnenmodell II mit  $s$  schwarzen und  $w = N - s$  weißen Kugeln betrachten. Da sich wegen des Zurücklegens auch im Urnenmodell II die einzelnen Versuchsausgänge nicht beeinflussen (entsprechende Ereignisse sind „stochastisch unabhängig“), so ist zumindest heuristisch klar, dass die zu Beispiel 1.5 gesuchte Wahrscheinlichkeit gleich jener Wahrscheinlichkeit ist, dass bei Urnenmodell II genau  $k$ -mal eine schwarze Kugel gezogen wird. In diesem Fall gilt also

$$P(\text{„genau } k \text{ Erfolge treten ein“}) = \binom{n}{k} p^k (1-p)^{n-k} \quad \forall k = 0, 1, \dots, n. \quad (1.1)$$

Im Falle von irrationalem  $p$  lässt sich zwar kein äquivalentes Urnenmodell finden, jedoch kann man  $p$  durch rationale Zahlen approximieren und es erscheint naheliegend, dass (1.1) auch in diesem Fall gilt. Eine fundierte Begründung hierfür wird nachgereicht nachdem der Begriff der stochastischen Unabhängigkeit eingeführt wurde.

## 2 Mathematische Modellierung von Zufallsexperimenten

In ersten Kapitel hatten wir einige einfache Zufallsexperimente vorrangig verbal beschrieben und den in der Wahrscheinlichkeitstheorie üblichen Formalismus nur ansatzweise benutzt. Der Fokus auf konkrete Modelle hatte den Einstieg in dieses Gebiet vermutlich erleichtert. Andererseits hat das Fehlen eines Formalismus beispielsweise verhindert auszudrücken, was genau mit „stochastischer Unabhängigkeit“ gemeint ist; siehe Beispiele 1.4 und 1.5. Die Möglichkeit einer abstrakten Beschreibung von Zufallsexperimenten wird erlauben, Gesetze der Wahrscheinlichkeitstheorie nicht nur für spezielle Zufallsexperimente, sondern in einem allgemeinen Rahmen herzuleiten.

Welche Bestandteile soll nun ein jedes Modell für ein Zufallsexperiment besitzen? Zunächst sollte die Menge der möglichen Versuchsausgänge in einer möglichst kompakten, aber auch für die jeweiligen Zwecke geeigneten Form beschrieben werden. Die Menge der möglichen Versuchsausgänge wird in der Wahrscheinlichkeitstheorie üblicherweise mit  $\Omega$  bezeichnet und Grundraum genannt. Elemente von  $\Omega$  sind dann die einzelnen Versuchsausgänge, genannt **Elementarereignisse**, und werden mit Kleinbuchstaben wie z.B.  $\omega$  bezeichnet. Nun kommen wir zu den Wahrscheinlichkeiten. Bei einem Laplace-Experiment sind die Wahrscheinlichkeiten für das Eintreten eines der möglichen Versuchsausgänge gleich  $1/\#\Omega$ . So ist bei Beispiel 1.4 die Wahrscheinlichkeit, dass eine bestimmte Folge von Nummern  $(\omega_1, \dots, \omega_n)$  eintritt gleich  $1/(N(N-1)\cdots(N-n+1))$ . In diesem Fall waren wir aber gar nicht wirklich an der Wahrscheinlichkeit eines gewissen Musters an Nummern interessiert, sondern lediglich an der Wahrscheinlichkeit, dass genau  $k$  der gezogenen Kugeln rot sind. Da dies äquivalent dazu ist, dass genau  $k$  der gezogenen Nummern nicht größer als  $s$  sind, konnten wir ohne Mühe auch die Wahrscheinlichkeit des uns interessierenden Ereignisses berechnen. Die Wahrscheinlichkeiten von Ereignissen werden nun durch eine Funktion  $P$  beschrieben, welche den Ereignissen (Teilmengen von  $\Omega$ ) ihre jeweilige Wahrscheinlichkeit zuordnet. Im Falle eines Laplace-Experimentes kann man problemlos jeder Teilmenge von  $\Omega$  ihre Wahrscheinlichkeit zuordnen. In diesem Fall ist also  $P: 2^\Omega \rightarrow [0, 1]$  gegeben durch  $P(A) = \#A/\#\Omega$  für alle  $A \subseteq \Omega$ . (An dieser Stelle wird das Symbol  $2^\Omega$  anstelle anderer Schreibweisen wie z.B.  $\mathcal{P}(\Omega)$  für die Potenzmenge benutzt, damit eine zu große Ähnlichkeit mit dem Symbol  $P(A)$  für die Wahrscheinlichkeit eines Ereignisses  $A$  vermieden wird.) Wir werden jedoch später sehen, dass wir in gewissen, durchaus relevanten Fällen nicht jeder Teilmenge auf widerspruchsfreie Art eine Wahrscheinlichkeit zuordnen können. Wir werden daher geeignete, wohl strukturierte Systeme von Teilmengen von  $\Omega$  heranziehen, auf denen dies dann gut möglich ist. Ein mathematisches Modell für ein Zufallsexperiment wird aus den folgenden Bestandteilen bestehen:

- $\Omega$ : **Grundraum**, Menge der möglichen Versuchsausgänge, welche **Elementarereignisse** genannt werden
- $\mathcal{A}$ : Menge jener Teilmengen von  $\Omega$  („**Ereignisse**“), denen Wahrscheinlichkeiten zugeordnet werden
- $P: \mathcal{A} \rightarrow [0, 1]$ : „**Wahrscheinlichkeitsmaß**“.

Hier kommen nun die exakten Definitionen:

**Definition 2.1.**  $\Omega$  sei eine nichtleere Menge. Ein System  $\mathcal{A}$  von Teilmengen von  $\Omega$  heißt  **$\sigma$ -Algebra in  $\Omega$** , wenn es die folgenden Eigenschaften besitzt:

- (i)  $\Omega \in \mathcal{A}$ ,
- (ii) falls  $A \in \mathcal{A}$ , so folgt  $A^c \in \mathcal{A}$ ,
- (iii) falls  $A_1, A_2, \dots \in \mathcal{A}$ , so folgt  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ .

Falls statt (iii) nur die Eigenschaft

- (iii') falls  $A_1, A_2 \in \mathcal{A}$ , so folgt  $A_1 \cup A_2 \in \mathcal{A}$

gilt, so heißt  $\mathcal{A}$  **Algebra (Mengenalgebra)** in  $\Omega$ .

Die Struktur einer  $\sigma$ -Algebra  $\mathcal{A}$  ist so, dass Ereignisse, welche üblicherweise von Interesse sind, auch tatsächlich in  $\mathcal{A}$  enthalten sind. Falls  $A \in \mathcal{A}$  und das Elementarereignis  $\omega$  eingetreten ist, so bedeutet  $\omega \in A$ , dass auch das Ereignis  $A$  eingetreten ist. Umgekehrt bedeutet  $\omega \in A^c$ , dass das Ereignis  $A$  nicht eingetreten ist. Falls  $A_1, A_2, \dots$  Ereignisse aus  $\mathcal{A}$  sind und das Elementarereignis  $\omega$  eingetreten ist, so bedeutet  $\omega \in \bigcup_{i=1}^{\infty} A_i$ , dass mindestens eines der Ereignisse  $A_1, A_2, \dots$  eingetreten ist. Weiter unten wird dann gezeigt, dass aus  $A_1, A_2, \dots \in \mathcal{A}$  stets  $\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$  folgt.  $\omega \in \bigcap_{i=1}^{\infty} A_i$  bedeutet schließlich, dass alle Ereignisse  $A_1, A_2, \dots$  eingetreten sind.

Falls  $\Omega$  eine beliebige nichtleere Menge ist, so sind einfache Beispiele für  $\sigma$ -Algebren in  $\Omega$  gegeben durch:

- $\mathcal{A}_1 = \{\emptyset, \Omega\}$ ,
- $\mathcal{A}_2 = 2^\Omega$  (Potenzmenge von  $\Omega$ ).

Es ist leicht zu sehen, dass  $\mathcal{A}_1$  und  $\mathcal{A}_2$   $\sigma$ -Algebren in  $\Omega$  sind. Darüber hinaus ist klar, dass  $\mathcal{A}_1$  die kleinste und  $\mathcal{A}_2$  die größte  $\sigma$ -Algebra in  $\Omega$  ist.

**Definition 2.2.**  $\mathcal{A}$  sei eine  $\sigma$ -Algebra in einer nichtleeren Menge  $\Omega$ . Eine Abbildung  $P: \mathcal{A} \rightarrow [0, 1]$  heißt **Wahrscheinlichkeitsmaß** auf  $\mathcal{A}$ , falls

- (i)  $P(\Omega) = 1$ ,
- (ii) Falls  $A_1, A_2, \dots \in \mathcal{A}$  disjunkt sind, so folgt

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \quad (\text{„}\sigma\text{-Additivität“})$$

$(\Omega, \mathcal{A}, P)$  heißt **Wahrscheinlichkeitsraum**.

An dieser Stelle sei noch einmal auf die Abstraktheit der Definition eines Wahrscheinlichkeitsmaßes hingewiesen. Es wird lediglich verlangt, dass die Funktion  $P$  auf einer  $\sigma$ -Algebra  $\mathcal{A}$  in einer beliebigen Menge  $\Omega$  definiert ist. Axiom (i) ist eine Normierungsbedingung und (ii) ist eine Eigenschaft, welche der Anschauung entspricht. Mit dem hier vorhandenen Grad der Allgemeinheit geht die breite Anwendbarkeit dieser Konzepte einher. Einige Eigenschaften einer  $\sigma$ -Algebra seien im Folgenden aufgezählt.

**Lemma 2.3.**  $\mathcal{A}$  sei eine  $\sigma$ -Algebra in  $\Omega$ . Dann gelten:

- (i)  $A_1, A_2, \dots \in \mathcal{A} \implies \bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$ ,
- (ii)  $A_1, \dots, A_n \in \mathcal{A} \implies \bigcup_{i=1}^n A_i \in \mathcal{A}, \bigcap_{i=1}^n A_i \in \mathcal{A}$ ,
- (iii)  $A, B \in \mathcal{A} \implies A \setminus B \in \mathcal{A}$ .

*Beweis.* Übungsaufgabe 5, Serie 2 □

Der folgende Satz fasst einige wichtige Eigenschaften von Wahrscheinlichkeitsmaßen zusammen.

**Satz 2.4.**  $(\Omega, \mathcal{A}, P)$  sei ein Wahrscheinlichkeitsraum. Dann gelten:

(i)  $P(\emptyset) = 0$  (Nulltreue)

(ii) Falls  $A_1, \dots, A_n \in \mathcal{A}$  disjunkt sind, so folgt

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i). \quad \text{(endliche Additivität)}$$

(iii) Falls  $A, B \in \mathcal{A}, A \subseteq B$ , so folgt

$$P(A) \leq P(B). \quad \text{(Isotonie)}$$

(iv) Falls  $A, B \in \mathcal{A}, A \subseteq B$ , so folgt

$$P(B \setminus A) = P(B) - P(A). \quad \text{(Subtraktivität)}$$

(v) Falls  $A_1, A_2, \dots \in \mathcal{A}, A_n \subseteq A_{n+1} \forall n \in \mathbb{N}$ , so folgt

$$P(A_n) \xrightarrow{n \rightarrow \infty} P\left(\bigcup_{i=1}^{\infty} A_i\right). \quad \text{(Stetigkeit von unten)}$$

(vi) Falls  $A_1, A_2, \dots \in \mathcal{A}, A_n \supseteq A_{n+1} \forall n \in \mathbb{N}$ , so folgt

$$P(A_n) \xrightarrow{n \rightarrow \infty} P\left(\bigcap_{i=1}^{\infty} A_i\right). \quad \text{(Stetigkeit von oben)}$$

(vii) Falls  $A_1, A_2, \dots \in \mathcal{A}$ , so folgt

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i). \quad (\text{Sub-}\sigma\text{-Additivitat})$$

*Beweis.* (i) Wir wahlen  $A_i := \emptyset \forall i \in \mathbb{N}$ . Da  $\Omega \in \mathcal{A}$ , so gilt auch  $A_i = \Omega^c \in \mathcal{A}$ . Aus der  $\sigma$ -Additivitat von  $P$  ergibt sich

$$P(\emptyset) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{\infty} P(\emptyset),$$

woraus  $P(\emptyset) = 0$  folgt.

(ii) Wir wahlen  $A_i := \emptyset \forall i > n$ . Dann folgt aus (i) und der  $\sigma$ -Additivitat von  $P$ , dass

$$P\left(\bigcup_{i=1}^n A_i\right) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \underbrace{P(A_i)}_{=0, \text{ fur } i > n} = \sum_{i=1}^n P(A_i).$$

(iii), (iv) Aus (ii) folgt, dass  $P(B) = P(A \cup (B \setminus A)) = P(A) + P(B \setminus A)$ . Daraus folgen jedoch (iii) und (iv).

(v) Zum Beweis dieser Aussage werden wir die Eigenschaft der  $\sigma$ -Additivitat heranziehen, welche jedoch fur disjunkte Mengen gilt. Daher werden wir  $\bigcup_{i=1}^{\infty} A_i$  durch eine Vereinigung disjunkter Mengen darstellen und definieren:

$$B_1 = A_1, B_2 = A_2 \setminus A_1, \dots, B_n = A_n \setminus (A_1 \cup \dots \cup A_{n-1}), \dots$$

Nun gilt

$$A_i = \bigcup_{j=1}^i B_j.$$

Daraus folgt

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} \bigcup_{j=1}^i B_j = \bigcup_{j=1}^{\infty} B_j.$$

Da die Mengen  $B_1, B_2, \dots$  nach Konstruktion disjunkt sind, so folgt aus der  $\sigma$ -Additivitat und (ii), dass

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} A_i\right) &= P\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i) \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n P(B_i) \stackrel{(ii)}{=} \lim_{n \rightarrow \infty} P\left(\bigcup_{i=1}^n B_i\right) = \lim_{n \rightarrow \infty} P(A_n). \end{aligned}$$

- (vi) Wir werden die soeben bewiesene Aussage (v) zum Beweis von (vi) benutzen. Dazu ist allerdings eine nichtfallende Folge von Ereignissen erforderlich;  $(A_1 \setminus A_i)_{i \in \mathbb{N}}$  ist eine solche. Somit gilt nach (v):

$$\lim_{i \rightarrow \infty} P(A_1 \setminus A_i) = P\left(\bigcup_{i=1}^{\infty} (A_1 \setminus A_i)\right). \quad (2.1)$$

Nun formen wir noch die Terme in (2.1) so um, dass wir die gewünschte Gleichheit erhalten. Zunächst gilt

$$\lim_{i \rightarrow \infty} P(A_1 \setminus A_i) = \lim_{i \rightarrow \infty} \{P(A_1) - P(A_i)\} = P(A_1) - \lim_{i \rightarrow \infty} P(A_i). \quad (2.2)$$

Wegen

$$\begin{aligned} \bigcup_{i=1}^{\infty} (A_1 \setminus A_i) &= \bigcup_{i=1}^{\infty} (A_1 \cap A_i^c) = A_1 \cap \left(\bigcup_{i=1}^{\infty} A_i^c\right) \\ &= A_1 \cap \left(\bigcap_{i=1}^{\infty} A_i\right)^c = A_1 \setminus \left(\bigcap_{i=1}^{\infty} A_i\right) \end{aligned}$$

erhalten wir

$$P\left(\bigcup_{i=1}^{\infty} (A_1 \setminus A_i)\right) = P\left(A_1 \setminus \left(\bigcap_{i=1}^{\infty} A_i\right)\right) = P(A_1) - P\left(\bigcap_{i=1}^{\infty} A_i\right). \quad (2.3)$$

Aus (2.1) bis (2.3) folgt schließlich (vi).

- (vii) Wir definieren wie beim Beweis von (v) disjunkte Menge, deren Vereinigung  $\bigcup_{i=1}^{\infty} A_i$  ergibt:

$$B_1 = A_1, B_2 = A_2 \setminus A_1, \dots, B_n = A_n \setminus (A_1 \cup \dots \cup A_{n-1}), \dots$$

Dann gilt

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \underbrace{P(B_i)}_{\leq P(A_i) \text{ wegen Isotonie}} \leq \sum_{i=1}^{\infty} P(A_i).$$

□

Im Folgenden wollen wir eine einfache Rechenregel herleiten, die Formel von Poicaré und Sylvester, auch als Siebformel bekannt. Dazu bietet sich die folgende (wohl etwas gekünstelte) Fragestellung an:

$n$  Teilnehmer einer Feier bringen jeweils ein Geschenk mit. Diese Geschenke werden anschließend „rein zufällig“ verteilt. (Dies ist eine wohl auch in Schulbüchern übliche Umschreibung der Tatsache, dass jede mögliche Verteilung mit derselben Wahrscheinlichkeit eintritt.)

Wie groß ist nun die Wahrscheinlichkeit, dass mindestens ein Gast sein eigenes Geschenk erhält? Wie verhält sich diese Wahrscheinlichkeit wenn  $n$  sehr groß ist?

Um möglichst „unfallfrei“ zur Lösung zu kommen, sollte man zunächst ein mathematisches Modell aufstellen. Damit lassen sich verbal getroffene Annahmen noch einmal klar formulieren und notwendige Überlegungen lassen sich auf solch einer abstrakten Ebene besser beschreiben. Die Menge der möglichen Versuchsausgänge lässt sich durch folgenden Grundraum in geeigneter Form beschreiben:

$$\Omega = \{(\omega_1, \dots, \omega_n): \omega_i \in \{1, 2, \dots, n\}, \omega_i \neq \omega_j \text{ für } i \neq j\}.$$

(Hierbei ist unerheblich, ob  $\omega_i = j$  das Ereignis „Der  $i$ -te Teilnehmer bekommt das  $j$ -te Geschenk.“ oder „Das  $i$ -te Geschenk geht an den  $j$ -ten Teilnehmer.“ beschreibt; in beiden Fällen erhält man schließlich dieselbe Lösung.)

Die verbal getroffene Annahme, dass die Geschenke „rein zufällig“ verteilt werden bedeutet, dass

$$P(\{\omega\}) = \frac{1}{\#\Omega} = \frac{1}{n!} \quad \forall \omega \in \Omega$$

gilt. Nun definieren wir Ereignisse, deren Wahrscheinlichkeiten sich leicht bestimmen lassen:

$$A_i = \text{„Der } i\text{-te Gast erhält sein eigenes Geschenk.“} = \{\omega \in \Omega: \omega_i = i\}.$$

Gesucht ist die Wahrscheinlichkeit, dass **mindestens** ein Gast sein eigenes Geschenk bekommt, also  $P(A_1 \cup \dots \cup A_n)$ . Die Bestimmung dieser Wahrscheinlichkeit ist nicht trivial, da die Kardinalität der Menge  $A_1 \cup \dots \cup A_n$  zumindest nicht offensichtlich ist. Andererseits lassen sich jedoch die Wahrscheinlichkeiten des gemeinsamen Eintretens von diesen Ereignissen bestimmen. So gelten  $\#A_i = (n-1)!$  sowie, für  $i \neq j$ ,  $\#A_i \cap A_j = (n-2)!$  u.s.w., woraus

$$P(A_i) = \frac{\#A_i}{\#\Omega} = \frac{(n-1)!}{n!}, \quad P(A_i \cap A_j) = \frac{\#A_i \cap A_j}{\#\Omega} = \frac{(n-2)!}{n!} \quad (i \neq j) \quad \text{u.s.w.}$$

folgen. Nun stellt sich die Frage, inwieweit diese Überlegung bei der gewünschten Bestimmung von  $P(A_1 \cup \dots \cup A_n)$  hilft. Wir betrachten dazu zunächst den Fall  $n = 2$ , welcher auch Gegenstand des Schulunterrichts ist. Das Ereignis  $A_1 \cup A_2$  lässt sich folgendermaßen als Vereinigung disjunkter Ereignisse darstellen:

$$A_1 \cup A_2 = (A_1 \cap A_2) \cup (A_1 \setminus (A_1 \cap A_2)) \cup (A_2 \setminus (A_1 \cap A_2)).$$

(Auf das obligatorische Bildchen sei an dieser Stelle verzichtet, da dies auch so offensichtlich ist.)

Damit ergibt sich unter Ausnutzung der Additivität und Subtraktivität von  $P$ , dass

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1 \cap A_2) + P(A_1 \setminus (A_1 \cap A_2)) + P(A_2 \setminus (A_1 \cap A_2)) \\ &= P(A_1 \cap A_2) + P(A_1) - P(A_1 \cap A_2) + P(A_2) - P(A_1 \cap A_2) \\ &= P(A_1) + P(A_2) - P(A_1 \cap A_2). \end{aligned}$$

Dieses Ergebnis kann nun leicht auf den Fall von  $n$  Ereignissen verallgemeinert werden.

**Satz 2.5.** (Formel von Poincaré und Sylvester, Siebformel)

$(\Omega, \mathcal{A}, P)$  sei ein Wahrscheinlichkeitsraum und  $A_1, \dots, A_n \in \mathcal{A}$  seien Ereignisse. Dann gilt:

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{k=1}^n (-1)^{k+1} \sum_{(i_1, \dots, i_k): 1 \leq i_1 < \dots < i_k \leq n} P\left(A_{i_1} \cap \dots \cap A_{i_k}\right).$$

*Beweis.* Übungsaufgabe 7, Serie 2 □

Mit diesem Resultat lässt sich nun die „Geschenkeaufgabe“ einfach lösen. Es gelten  $\#\{(i_1, \dots, i_k): 1 \leq i_1 < \dots < i_k \leq n\} = \binom{n}{k}$  und, für  $1 \leq i_1 < \dots < i_k \leq n$ ,  $\#(A_{i_1} \cap \dots \cap A_{i_k}) = (n-k)!$ . Damit folgt

$$\begin{aligned} & P(A_1 \cup \dots \cup A_n) \\ &= \sum_{i_1: 1 \leq i_1 \leq n} \underbrace{P(A_{i_1})}_{=(n-1)!/n!} - \sum_{(i_1, i_2): 1 \leq i_1 < i_2 \leq n} \underbrace{P(A_{i_1} \cap A_{i_2})}_{=(n-2)!/n!} + \dots \\ & \quad \dots + (-1)^{n+1} \sum_{(i_1, \dots, i_n): 1 \leq i_1 < \dots < i_n \leq n} \underbrace{P(A_{i_1} \cap \dots \cap A_{i_n})}_{=(n-n)!/n!=1/n!} \\ &= \binom{n}{1} \frac{(n-1)!}{n!} - \binom{n}{2} \frac{(n-2)!}{n!} + \dots + (-1)^{n+1} \binom{n}{n} \frac{(n-n)!}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \dots + (-1)^{n+1} \frac{1}{n!} \\ &= 1 - \sum_{k=2}^n \frac{1}{k!} (-1)^k = 1 - \sum_{k=0}^n \frac{1}{k!} (-1)^k. \end{aligned}$$

Wenn wir die Potenzreihenentwicklung  $e^{-\lambda} = \sum_{k=0}^{\infty} (-\lambda)^k / k!$  benutzen, so sehen wir auch, dass selbst für sehr großes  $n$  die Wahrscheinlichkeit, dass ein Gast sein eigenes Geschenk bekommt weder nahe 0 noch nahe 1 ist. Es gilt stattdessen

$$\lim_{n \rightarrow \infty} P(A_1 \cup \dots \cup A_n) = 1 - \sum_{k=0}^{\infty} \frac{1}{k!} (-1)^k = 1 - e^{-1}.$$

Die Verwendung des Konzeptes einer  $\sigma$ -Algebra mag auf den ersten Blick abgehoben und vielleicht auch unnötig wirken. Die nächste Aussage wird zeigen, dass man in einfachen Situationen ein Wahrscheinlichkeitsmaß auf **allen** Teilmengen von  $\Omega$  definieren kann. Das System aller Teilmengen bildet natürlich auch eine solche  $\sigma$ -Algebra, jedoch ist in diesem Fall die abstrakte Definition eines solchen Mengensystems nicht erforderlich. Vermutlich wird dieses Empfinden durch Erfahrungen aus dem Schulunterricht unterstützt, wo man üblicherweise nicht mit dem Begriff der  $\sigma$ -Algebra belästigt wird. Andererseits gibt es durchaus relevante Fälle, in denen die widerspruchsfreie Festlegung eines Wahrscheinlichkeitsmaßes auf allen Teilmengen eines geeigneten Grundraumes nicht möglich

ist. Ein Beispiel dafür wird am Ende dieses Kapitels vorgestellt werden. Auch bei der Festlegung von Wahrscheinlichkeitsmaßen kann man zwangsläufig wieder beim Konzept der  $\sigma$ -Algebra landen; siehe ebenfalls das Beispiel am Ende dieses Kapitels. Die dann vorausgesetzten Struktureigenschaften einer  $\sigma$ -Algebra sichern, dass für die üblicherweise interessierenden Ereignisse tatsächlich die jeweiligen Wahrscheinlichkeiten wohldefiniert sind.

**Lemma 2.6.**  $\Omega$  sei eine nichtleere, höchstens abzählbare Menge, d.h.,  $\Omega = \{\omega_1, \dots, \omega_N\}$  bzw.  $\Omega = \{\omega_1, \omega_2, \dots\}$ . Weiter seien  $p_i \geq 0$  mit  $\sum_{i=1}^N p_i = 1$  bzw.  $\sum_{i=1}^{\infty} p_i = 1$  gegeben. Dann existiert ein eindeutig bestimmtes Wahrscheinlichkeitsmaß  $P: 2^\Omega \rightarrow [0, 1]$  mit  $P(\{\omega_i\}) = p_i \forall i$ .

*Beweis.* Wegen der zu beachtenden  $\sigma$ -Additivität des zu definierenden Wahrscheinlichkeitsmaßes ist die einzig mögliche Definition gegeben durch

$$P(A) := \sum_{i: \omega_i \in A} p_i \quad \forall A \subseteq \Omega.$$

Damit gilt insbesondere  $P(\{\omega_i\}) = p_i \forall i$ . Außerdem erfüllt die so definierte Funktion  $P: 2^\Omega \rightarrow [0, 1]$  die Axiome eines Wahrscheinlichkeitsmaßes. Es gelten nämlich

$$P(\Omega) = \sum_i p_i = 1$$

sowie, für beliebige disjunkte Mengen  $A_1, A_2, \dots \in 2^\Omega$ ,

$$P\left(\bigcup_i A_i\right) = \sum_{j: \omega_j \in \bigcup_i A_i} p_j = \sum_i \sum_{j: \omega_j \in A_i} p_j = \sum_i P(A_i).$$

□

Es kann also konstatiert werden, dass im Falle eines höchstens abzählbaren Grundraumes  $\Omega$  die Festlegung der Wahrscheinlichkeiten auf allen einelementigen Mengen  $\{\omega\}$  für  $\omega \in \Omega$  hinreichend ist. Als  $\sigma$ -Algebra kann dann stets die Potenzmenge von  $\Omega$  gewählt werden. Falls jedoch der Grundraum  $\Omega$  überabzählbar ist, so kann die Definition eines Wahrscheinlichkeitsmaßes durchaus komplizierter werden. Wir betrachten dazu noch das folgende Beispiel, welches wohl über den Unterrichtsstoff an Gymnasien hinausgeht, jedoch eine Illustration für die Notwendigkeit der verwendeten Konzepte liefert.

Wir stellen uns vor, dass eine unverfälschte Münze (d.h., beim Werfen erscheinen „Wappen“ oder „Zahl“ jeweils mit der Wahrscheinlichkeit  $1/2$ ) unendlich oft „rein zufällig“ geworfen wird. (Bevor Protest angemeldet wird: Dies ist in der Praxis vermutlich schwer durchführbar; jedoch sind Modelle mit unendlich vielen Versuchsdurchführungen für theoretische (z.B. asymptotische) Betrachtungen durchaus üblich. Für den Mathematiker ist es hinreichend, dass solch ein Modell widerspruchsfrei konstruierbar ist.)

Als Grundraum zur Beschreibung der möglichen Versuchsausgänge eines solchen Zufallsexperiments drängt sich natürlich

$$\Omega = \left\{ (\omega_1, \omega_2, \dots) : \omega_i \in \{0, 1\} \right\}$$

auf. Bevor wir nun die Festlegung einer geeigneten  $\sigma$ -Algebra diskutieren, überlegen wir uns, welche Eigenschaften das zu konstruierende Wahrscheinlichkeitsmaß  $P$  besitzen sollte. Für beliebige  $\omega'_1, \dots, \omega'_n \in \{0, 1\}$  definieren wir das Ereignis  $A = \{\omega \in \Omega : \omega_1 = \omega'_1, \dots, \omega_n = \omega'_n\}$ . Da die Münze unverfälscht sein soll und die einzelnen Münzwürfe sich nicht beeinflussen sollen (später fassen wir dies unter „stochastischer Unabhängigkeit“), so muss  $P$  so festgelegt werden, dass

$$P(A) = 2^{-n}. \quad (2.4)$$

Somit gilt für beliebiges  $\omega' \in \Omega$ , dass

$$P(\{\omega'\}) \leq P(\{\omega \in \Omega : \omega_1 = \omega'_1, \dots, \omega_n = \omega'_n\}) = 2^{-n} \quad \forall n \in \mathbb{N},$$

woraus

$$P(\{\omega'\}) = 0$$

folgt. An dieser Stelle wird bereits offenbar, dass eine Definition des Wahrscheinlichkeitsmaßes anhand der Festlegung der Wahrscheinlichkeiten einelementiger Teilmengen von  $\Omega$  zu keinem sinnvollen Ergebnis führt. Aus Symmetriegründen ist jedoch klar, dass  $P$  die im Weiteren beschriebene Eigenschaft haben sollte. Ausgehend von Abbildungen  $T_n : \Omega \rightarrow \Omega$  mit

$$T_n(\omega) := (\omega_1, \dots, \omega_{n-1}, 1 - \omega_n, \omega_{n+1}, \omega_{n+2}, \dots)$$

definieren wir zu  $A \subseteq \Omega$  die Menge

$$T_n(A) = \{T_n(\omega) : \omega \in A\}.$$

Falls wir nun  $P$  auf  $2^\Omega$  definieren, so sollte gelten

$$P(T_n(A)) = P(A) \quad \forall A \in 2^\Omega, \forall n \in \mathbb{N}. \quad (2.5)$$

Die nun folgende Aussage zeigt, dass es kein Wahrscheinlichkeitsmaß  $P$  auf  $2^\Omega$  gibt, welches diese Symmetrieeigenschaft besitzt.

**Bemerkung 2.7.** Es gibt **kein** Wahrscheinlichkeitsmaß  $P$  auf  $2^\Omega$ , welches (2.5) erfüllt.

*Beweis.* Die Idee des Beweises besteht darin, dass wir  $\Omega$  in abzählbar viele disjunkte Ereignisse zerlegen, welche wegen (2.5) alle dieselbe Wahrscheinlichkeit besitzen. Dies liefert dann einen Widerspruch.

Zunächst definieren wir mit Hilfe der Abbildungen  $T_n$  ( $n \in \mathbb{N}$ ) eine Äquivalenzrelation auf  $\Omega$ .  $\omega$  und  $\omega'$  seien äquivalent ( $\omega \sim \omega'$ ), falls  $n_1, \dots, n_k \in \mathbb{N}$  und  $k \in \mathbb{N}$  existieren mit

$$\omega = T_{n_1} \circ \dots \circ T_{n_k}(\omega'),$$

d.h.,  $\omega$  unterscheidet sich von  $\omega'$  nur an endlich vielen Stellen. Es ist leicht zu sehen, dass  $\sim$  tatsächlich eine Äquivalenzrelation auf  $\Omega$  definiert, d.h.,  $\sim$  besitzt die Eigenschaften der Reflexivität, Symmetrie und Transitivität. Nun wählen wir aus jeder Äquivalenzklasse genau einen Repräsentanten; dies ergibt die Menge  $A$ . Weiter sei

$$\mathcal{T} := \{T_{n_1} \circ \dots \circ T_{n_k} : n_1, \dots, n_k \in \mathbb{N}, k \in \mathbb{N}\}$$

die Gesamtheit endlicher Verknüpfungen der  $T_n$ . Nun gelten:

- a)  $\mathcal{T}$  ist abzählbar unendlich,
- b)  $\bigcup_{T \in \mathcal{T}} T(A) = \Omega$ ,
- c) falls  $S, T \in \mathcal{T}$  und  $S \neq T$ , so  $S(A) \cap T(A) = \emptyset$ , (Dies sieht man am einfachsten indirekt. Es seien also  $S, T \in \mathcal{T}$  und  $S \neq T$ . Falls nun  $\omega \in S(A) \cap T(A)$ , so existieren  $\omega_S, \omega_T \in A$  mit  $\omega = S(\omega_S) = T(\omega_T)$ . Daraus folgt jedoch, dass  $\omega_T = T \circ S(\omega_S)$ , d.h.  $\omega_S \sim \omega_T$ . Da  $A$  aus jeder Äquivalenzklasse nur ein Element enthält, folgt  $\omega_S = \omega_T$ . Nun müsste aber  $\omega = S(\omega_S) = T(\omega_S)$  gelten, was wegen  $S \neq T$  zu einem Widerspruch führt.)
- d) für  $T = T_{n_1} \circ \dots \circ T_{n_k}$  gilt

$$P(T(A)) = P(T_{n_1} \circ \dots \circ T_{n_k}(A)) = P(T_{n_1} \circ \dots \circ T_{n_{k-1}}(A)) = \dots = P(A).$$

Aus b) bis d) folgt nun, dass

$$1 = P(\Omega) = P\left(\bigcup_{T \in \mathcal{T}} T(A)\right) = \sum_{T \in \mathcal{T}} P(T(A)) = \sum_{T \in \mathcal{T}} P(A).$$

Dies ist jedoch unmöglich, da nach a)  $\mathcal{T}$  abzählbar unendlich ist. □

Trotz der hier dargestellten Problematik ist es möglich, auf widerspruchsfreie Art ein Modell für den abzählbar oft wiederholten Münzwurf zu konstruieren. Dies erfordert allerdings tiefliegende Hilfsmittel aus der sogenannten Maßtheorie, weswegen wir hier auf eine detaillierte Darstellung verzichten. Die Definition eines adäquaten Wahrscheinlichkeitsmaßes  $P$  auf einer geeigneten  $\sigma$ -Algebra  $\mathcal{A}$  erfolgt dann in den folgenden Schritten:

- $P$  wird zunächst auf Mengen einfacher Bauart festgelegt: Für  $A = \{\omega \in \Omega : \omega_1 = \omega'_1, \dots, \omega_n = \omega'_n\}$  für gewisse  $\omega'_1, \dots, \omega'_n \in \{0, 1\}$ ,  $n \in \mathbb{N}$  definieren wir

$$P(A) = 2^{-n}. \tag{2.6}$$

- Wir wählen  $\mathcal{A}$  als die kleinste  $\sigma$ -Algebra in  $\Omega$ , welche alle Mengen der obigen Bauart enthält.

Nun hilft ein tatsächlich tiefliegender und in seiner Bedeutung kaum zu unterschätzender Satz aus der Maßtheorie, welcher besagt, dass es ein eindeutig bestimmtes Wahrscheinlichkeitsmaß  $P$  auf  $\mathcal{A}$  gibt, welches die obige Eigenschaft (2.6) besitzt.

Nach der vorher bewiesenen Bemerkung 2.7 ist klar, dass  $\mathcal{A} \neq 2^\Omega$  gilt. Trotzdem enthält  $\mathcal{A}$  eigentlich alle Mengen, welche normalerweise von Interesse sind. Ein weiterer Vorteil dieses Zuganges ist, dass eine konkrete Festsetzung der Wahrscheinlichkeiten nur für sehr einfach strukturierte Mengen erforderlich ist, was in der Tat elementar ist.

### 3 Elementare bedingte Wahrscheinlichkeiten

In diesem Kapitel wollen wir uns dem Begriff der (elementaren) bedingten Wahrscheinlichkeit nähern. Um die später folgende abstrakte Definition nachvollziehen zu können, betrachten wir zunächst ein einfaches Beispiel.

Wir stellen uns vor, dass sich in einer Urne  $s$  schwarze und  $w$  weiße, ansonsten gleichartige Kugeln befinden ( $s \geq 2$ ,  $w \geq 1$ ). Aus dieser Urne werden „rein zufällig“ nacheinander zwei Kugeln ohne Zurücklegen gezogen. Angenommen, beim ersten Ziehen wurde eine schwarze Kugel gezogen, wie groß ist **dann** die Wahrscheinlichkeit, dass beim zweiten Ziehen wieder eine schwarze Kugel gezogen wird?

**Lösung:** Die gesuchte (bedingte) Wahrscheinlichkeit kann auch ohne komplizierte Rechnungen leicht bestimmt werden. Während die Wahrscheinlichkeit, dass beim ersten Ziehen eine schwarze Kugel gezogen wird, gleich  $s/(s+w)$  ist, finden wir vor den zweiten Ziehen eine veränderte Ausgangssituation vor. Nachdem eine schwarze Kugel aus der Urne entnommen wurde, befinden sich noch  $s-1$  schwarze sowie  $w$  weiße Kugeln in der Urne. Da wiederum davon ausgegangen wird, dass beim zweiten Ziehen jede der noch verfügbaren Kugeln mit der gleichen Wahrscheinlichkeit gezogen wird, ist die gesuchte bedingte Wahrscheinlichkeit gleich  $(s-1)/(s+w-1)$ .

Wir betrachten nun die in Beispiel 1.3 benutzte Modellierung durch ein Laplace-Experiment. Dazu nummerieren wir die schwarzen Kugeln mit  $1, \dots, s$  und die weißen mit  $s+1, \dots, s+w =: N$  durch. Da nur zwei Mal gezogen wird, können wir den Grundraum so wählen:

$$\Omega = \left\{ (\omega_1, \omega_2) : \omega_i \in \{1, \dots, N\}, \omega_1 \neq \omega_2 \right\}.$$

Wie bereits erläutert, handelt es sich hier um ein Laplace-Experiment, d.h.

$$P(\{\omega\}) = \frac{1}{\#\Omega} = \frac{1}{N(N-1)} \quad \forall \omega \in \Omega.$$

Wir definieren die folgenden Ereignisse:

$$\begin{aligned} A_1 &= \text{„erste gezogene Kugel schwarz“} = \{\omega \in \Omega : \omega_1 \leq s\}, \\ A_2 &= \text{„zweite gezogene Kugel schwarz“} = \{\omega \in \Omega : \omega_2 \leq s\}. \end{aligned}$$

Die Wahrscheinlichkeit, dass beim ersten Ziehen eine schwarze Kugel gezogen wird, ergibt sich nun als

$$P(A_1) = \frac{\#A_1}{\#\Omega} = \frac{s(N-1)}{N(N-1)} = \frac{s}{N}.$$

Analog kann man auch die Wahrscheinlichkeit, dass beim zweiten Ziehen eine schwarze Kugel gezogen wird, erhalten:

$$P(A_2) = \frac{\#A_2}{\#\Omega} = \frac{s(N-1)}{N(N-1)} = \frac{s}{N}. \quad (3.1)$$

Wenn nun aber nach der Wahrscheinlichkeit gefragt wird, dass nach Ziehen einer schwarzen Kugel eine weitere schwarze Kugel gezogen werden soll, so ändert sich die Ausgangslage: **Elementarereignisse  $\omega \in A_1^c$  können nun nicht mehr eintreten und**

die Menge  $A_1$  tritt an die Stelle des bisherigen Grundraumes  $\Omega$ . Nach wie vor sind nun die noch verbliebenen Elementarereignisse gleichwahrscheinlich. Jenes Ereignis, dass nach dem Ziehen einer schwarzen Kugel eine weitere schwarze Kugel gezogen wird entspricht dem Ereignis  $A_1 \cap A_2$ . Somit ergibt sich für die (bedingte) Wahrscheinlichkeit  $P(A_2 | A_1)$ , dass nach Ziehen einer schwarzen Kugel eine weitere schwarze Kugel gezogen wird,

$$P(A_2 | A_1) = \frac{\#A_1 \cap A_2}{\#A_1} = \frac{s(s-1)}{s(N-1)} = \frac{s(s-1)}{N(N-1)} \cdot \frac{s}{s} = \frac{P(A_1 \cap A_2)}{P(A_1)}. \quad (3.2)$$

An dieser Stelle kann man erkennen, wie die in (3.1) angegebene Wahrscheinlichkeit für das Eintreten von  $A_2$  angesichts des Bekanntwerdens des Eintretens von  $A_1$  angepasst wurde. Wir kommen nun zu einer formalen Definition der bedingten Wahrscheinlichkeit.

**Definition 3.1.**  $(\Omega, \mathcal{A}, P)$  sei ein Wahrscheinlichkeitsraum und  $A, B \in \mathcal{A}$  seien Ereignisse.

Falls  $P(B) > 0$ , so heißt

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

die (elementare) **bedingte Wahrscheinlichkeit** von  $A$  unter der Bedingung/Hypothese  $B$ .

Falls  $P(B) = 0$ , so setzen wir  $P(A | B) = 0$ .

Falls  $P(B) > 0$ , so heißt

$$P(\cdot | B): \mathcal{A} \rightarrow [0, 1]$$

**bedingte Verteilung** von  $A$  unter der Bedingung/Hypothese  $B$ .

Der hier eingeführte Begriff der bedingten Wahrscheinlichkeit ist im Rahmen dieser Vorlesung und wohl auch im Rahmen des Schulunterrichts ausreichend. In Fällen, wo es wichtig ist, bedingte Wahrscheinlichkeiten anzugeben obwohl die bedingenden Ereignisse Wahrscheinlichkeit 0 besitzen, lässt sich mit Mitteln der Maßtheorie die Definition bedingter Wahrscheinlichkeiten entsprechend erweitern. Zur Abgrenzung von jener in der (fortgeschrittenen) Mathematik üblichen Definition gebrauchen wir hier gelegentlich das Wort „elementar“ zur näheren Charakterisierung der gewählten Reichweite der Definition.

Ein interessanter Aspekt wird bei der obigen Definition der bedingten Wahrscheinlichkeit klar: Nach dieser Definition gilt im Falle des Urnenbeispiels auch

$$P(A_1 | A_2) = \frac{P(A_1 \cap A_2)}{P(A_2)} = \frac{s-1}{N-1}.$$

Das mag zunächst nicht der Intuition entsprechen, da in der zeitlichen Abfolge das Ereignis  $A_1$  nur vor dem Ereignis  $A_2$  eintreten kann. Die Möglichkeit solch einer „Zeitumkehr“ ist aber bei der obigen formalen Definition durchaus gewollt. Und letztlich ist es selbst auf anschaulicher Ebene nachvollziehbar: Nehmen wir an, es wird wie oben beschrieben zwei

Mal hintereinander aus der Urne gezogen. Ein Beobachter des Experimentes verrät Ihnen, dass beim zweiten Ziehen eine schwarze Kugel gezogen wurde. Mit diesem **Zusatzwissen** ausgestattet, wie hoch würden Sie die Wahrscheinlichkeit ansetzen, dass beim ersten Ziehen ebenfalls eine schwarze Kugel gezogen wurde? – Nachdem das Experiment des zweimaligen Ziehens ausgeführt wurde ist das Eintreten von  $A_1$  nicht mehr zufällig.  $A_1$  ist halt eingetreten oder auch nicht. Da wir aber **nicht** über das Eintreten bzw. Nichteintreten von  $A_1$  **informiert** sind, messen wir dem eine gewisse Wahrscheinlichkeit zu. Mit Wahrscheinlichkeiten kann also auch dann sinnvoll gerechnet werden, wenn die Entscheidung über das Eintreten/Nichteintreten des interessierenden Ereignisses bereits gefallen ist, man selbst aber nicht darüber informiert ist. Der Begriff der Wahrscheinlichkeit wird somit auch gebraucht, wenn nicht nur „reiner Zufall“ vorkommt, sondern wenn wir keine oder nur teilweise Kenntnis von Ausgang eines Zufallsexperimentes besitzen. (Auch wenn im Skatenspiel die Karten bereits ausgeteilt sind, so spekuliert ein halbwegs wachsender Spieler dennoch, mit welcher Wahrscheinlichkeit zwei für ihn passende Karten im sogenannten „Skat“ liegen.)

Der folgende Satz fasst einige wichtige Aussagen bzw. Rechenregeln im Zusammenhang mit bedingten Wahrscheinlichkeiten zusammen.

**Satz 3.2.**  $(\Omega, \mathcal{A}, P)$  sei ein Wahrscheinlichkeitsraum.

(i) Falls  $B \in \mathcal{A}$  und  $P(B) > 0$ , so ist

$$P(\cdot | B): \mathcal{A} \rightarrow [0, 1]$$

ein Wahrscheinlichkeitsmaß auf  $\mathcal{A}$ .

(ii) **(Multiplikationssatz)**

$A_1, \dots, A_n \in \mathcal{A}$  seien beliebige Ereignisse. Dann gilt

$$P(A_1 \cap \dots \cap A_n) = P(A_1)P(A_2 | A_1) \cdots P(A_n | A_1 \cap \dots \cap A_{n-1}).$$

(iii) **(Satz von der totalen Wahrscheinlichkeit)**

$A_1, \dots, A_n \in \mathcal{A}$  bzw.  $A_1, A_2, \dots \in \mathcal{A}$  seien disjunkte Ereignisse mit  $\sum_{i \geq 1} P(A_i) = 1$  und  $A \in \mathcal{A}$  sei ein weiteres Ereignis. Dann gilt

$$P(A) = \sum_{i \geq 1} P(A | A_i)P(A_i).$$

Teil (i) des obigen Satzes bedeutet, dass auch für bedingte Verteilungen jene Eigenschaften und Rechenregeln gelten, die für gewöhnliche Wahrscheinlichkeitsverteilungen hergeleitet wurden. Aussagen (ii) und (iii) sind sicher vielen aus dem Schulunterricht bekannt, auch wenn sich diese Rechenregeln dort hinter den berühmten Baumdiagrammen verstecken. Die formelmäßigen Darstellungen verdeutlichen, an welchen Stellen beim Durchgehen eines Baumdiagrammes mit gewöhnlichen bzw. bedingten Wahrscheinlichkeiten gerechnet werden muss. Wir werden im Folgenden einen mathematisch fundierten Beweis dieser Aussagen liefern.

*Beweis von Satz 3.2.* (i) Übungsaufgabe

- (ii) Wir beweisen die Aussage durch vollständige Induktion. Es sei zunächst  $n = 2$ . Falls  $P(A_1) > 0$ , so gilt nach Definition der bedingten Wahrscheinlichkeit  $P(A_2 | A_1) = P(A_1 \cap A_2) / P(A_1)$ , woraus die Behauptung folgt. Falls  $P(A_1) = 0$ , so erhalten wir  $P(A_1 \cap A_2) = 0 = P(A_1)P(A_2 | A_1)$ .

Wir nehmen nun an, dass die Formel für  $n - 1$  Ereignisse gilt. Dann folgt

$$\begin{aligned} & P(\underbrace{A_1 \cap A_2}_{=: B} \cap A_3 \cap \dots \cap A_n) \\ &= \underbrace{P(B)}_{=P(A_1)P(A_2|A_1)} P(A_3 | B) \dots P(A_n | B \cap A_3 \cap \dots \cap A_{n-1}) \\ &= P(A_1)P(A_2 | A_1)P(A_3 | A_1 \cap A_2) \dots P(A_n | A_1 \cap \dots \cap A_{n-1}). \end{aligned}$$

- (iii) Falls  $\bigcup_{i \geq 1} A_i = \Omega$ , so folgt  $A = A \cap \Omega = A \cap (\bigcup_{i \geq 1} A_i) = \bigcup_{i \geq 1} (A \cap A_i)$ . Da aus der Disjunktheit von  $A_1, A_2, \dots$  auch die von  $A \cap A_1, A \cap A_2, \dots$  folgt, erhalten wir aus der  $\sigma$ -Additivität von  $P$  sowie Aussage (ii), dass

$$P(A) = \sum_{i \geq 1} P(A \cap A_i) \underbrace{=}_{(ii)} \sum_{i \geq 1} P(A | A_i)P(A_i).$$

Wenn wir jedoch die Aussage (iii) in voller Allgemeinheit beweisen wollen, müssen wir aufpassen:  $\sum_{i \geq 1} P(A_i) = 1$  bedeutet selbst bei disjunkten  $A_1, A_2, \dots$  nicht notwendigerweise, dass  $\bigcup_{i \geq 1} A_i = \Omega$ , da es ggf. nichtleere Mengen  $B \in \mathcal{A}$  geben kann mit  $P(B) = 0$ . Wir definieren also  $A_0 := \Omega \setminus (\bigcup_{i \geq 1} A_i)$ . Nun gelten  $\Omega = \bigcup_{i \geq 0} A_i$  und

$$P(A_0) = P(\Omega) - P\left(\bigcup_{i \geq 1} A_i\right) = P(\Omega) - \sum_{i \geq 1} P(A_i) = 0.$$

Daraus folgt, dass

$$\begin{aligned} P(A) &= P\left(A \cap \left(\bigcup_{i \geq 0} A_i\right)\right) = P\left(\bigcup_{i \geq 0} (A \cap A_i)\right) \\ &= \sum_{i \geq 0} P(A \cap A_i) = \sum_{i \geq 0} P(A | A_i)P(A_i) \\ &= \sum_{i \geq 1} P(A | A_i)P(A_i). \end{aligned}$$

□

Die folgende Rechenregel dürfte auch den meisten aus dem Schulunterricht bekannt sein.

**Satz 3.3. (Satz von Bayes, Bayessche Formel)**

$(\Omega, \mathcal{A}, P)$  sei ein Wahrscheinlichkeitsraum,  $A_1, \dots, A_n \in \mathcal{A}$  bzw.  $A_1, A_2, \dots \in \mathcal{A}$  seien disjunkte Ereignisse mit  $\sum_{i \geq 1} P(A_i) = 1$ .  $B \in \mathcal{A}$  sei ein weiteres Ereignis mit  $P(B) > 0$ . Dann gilt

$$P(A_k | B) = \frac{P(B | A_k)P(A_k)}{\sum_{i \geq 1} P(B | A_i)P(A_i)}.$$

Die inhaltliche Bedeutung dieser Aussage wird klar, wenn man  $A_1, A_2, \dots$  als mögliche „Ursachen“ und  $B$  als eine mögliche „Wirkung“ ansieht. Falls man nun  $P(A_1), P(A_2), \dots$  und  $P(B | A_1), P(B | A_2), \dots$  kennt und  $B$  eingetreten ist, so kann man mit dieser Formel auf jene Wahrscheinlichkeit schließen, dass  $A_k$  eingetreten ist. (Die abstrakte Definition bedingter Wahrscheinlichkeiten gestattet auch solch eine „Zeitumkehr“.)

*Beweis von Satz 3.3.* Nach (ii) und (iii) aus Satz 3.2 folgt

$$P(A_k | B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(B | A_k)P(A_k)}{\sum_{i \geq 1} P(B | A_i)P(A_i)}.$$

□

**Beispiel 3.4.** Das folgende Beispiel für eine Anwendung der Bayesschen Formel findet sich auf Seite 53 im Buch von Hesse [2].

Die Trisomie 21 ist eine genetische Anomalie, deren Träger das Chromosom 21 dreifach besitzen (Down-Syndrom). Die Ausprägung dieser Anomalie beim Kind kann durch eine Fruchtwasseruntersuchung pränatal festgestellt werden, was jedoch mit einer gewissen Unsicherheit versehen ist. So wird angenommen, dass der Befund in 99% der Fälle positiv ist, falls Trisomie 21 vorliegt. Falls Trisomie 21 nicht vorliegt, so ist der Befund in 99% der Fälle negativ. Die Häufigkeit des Auftretens des Down-Syndroms beim Kind hängt stark vom Alter der werdenden Mutter ab. So kommt es bei einer 25-jährigen Mutter in einem von 1250 Fällen, bei einer 43-jährigen Mutter dagegen in einem von 50 Fällen vor. Fällt nun ein medizinischer Test auf Trisomie 21 positiv aus, so stellt sich für die werdende Mutter die Frage, wie hoch denn nun die Wahrscheinlichkeit ist, dass das Kind das Down-Syndrom besitzt.

**Lösung:** Zur besseren Beschreibung der Überlegungen definieren wir Ereignisse:  $D :=$  „Down-Syndrom liegt vor“,  $A :=$  „Test ist positiv“. Nun gelten im Falle einer 25-jährigen Mutter:

$$\begin{aligned} P(D) &= \frac{1}{1250}, & P(D^c) &= \frac{1249}{1250}, \\ P(A | D) &= \frac{99}{100}, & P(A | D^c) &= \frac{1}{100}. \end{aligned}$$

Damit ergibt sich für die bedingte Wahrscheinlichkeit, dass das Kind einer 25-jährigen bei einem positiven Testergebnis das Down-Syndrom aufweist

$$P(D | A) = \frac{P(A | D)P(D)}{P(A | D)P(D) + P(A | D^c)P(D^c)} = \dots \approx 0,07.$$

Im Falle einer 43-jährigen Mutter ergeben sich dagegen

$$P(D) = \frac{1}{50}, \quad P(D^c) = \frac{49}{50},$$

woraus

$$P(D | A) \approx 0,67$$

folgt.

Sollte jedoch der Test negativ ausfallen, so kann wohl in beiden Fällen Entwarnung gegeben werden:

$$P(D | A^c) \approx \begin{cases} 8,1 \cdot 10^{-6} & \text{bei 25-jähriger Mutter,} \\ 2,1 \cdot 10^{-4} & \text{bei 43-jähriger Mutter.} \end{cases}$$

Bei diesem Beispiel ist übrigens ein typisches Muster erkennbar. Im Falle seltener Krankheiten (oder Ereignisse) muss selbst ein positives Ergebnis eines recht zuverlässigen Tests noch nicht bedeuten, dass diese Krankheit tatsächlich mit einer hohen Wahrscheinlichkeit vorliegt. Umgekehrt liefert ein negatives Testergebnis weitgehend Gewissheit, dass solch eine Krankheit nicht vorliegt.

## 4 Stochastische Unabhängigkeit

Die Definition des Begriffs der stochastischen Unabhängigkeit zweier Ereignisse wird vermutlich auch ohne vorbereitende Betrachtungen als der Intuition nicht vollständig widersprechend erkannt. Andererseits ist vielleicht nicht klar, ob damit genau das abgebildet wird, was man intuitiv unter „Unabhängigkeit“ verstehen würde. Wir werden uns daher mit einer einfachen Überlegung an eine formale Definition herantasten. Wir nehmen an, dass  $A$  und  $B$  zwei Ereignisse auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  sind. Um triviale Fälle auszuschließen setzen wir voraus, dass  $P(A)$  weder 0 noch 1 ist. Wie soll man nun das voneinander „unabhängige“ Eintreten der Ereignisse  $A$  und  $B$  formal beschreiben? Sicher wird jeder zustimmen, dass die Verwendung des Begriffs der Unabhängigkeit gerechtfertigt ist, wenn **das Eintreten/Nichteintreten des Ereignisses  $A$  keinen Einfluss auf die Wahrscheinlichkeit des Eintretens von  $B$  hat**, d.h.,

$$P(B | A) = P(B | A^c).$$

Wegen  $P(B) = P(A)P(B | A) + P(A^c)P(B | A^c) = (P(A) + P(A^c))P(B | A)$  gilt diese Gleichheit genau dann, wenn

$$P(B) = P(B | A).$$

Das jedoch ist äquivalent zu

$$P(A \cap B) = P(A)P(B | A) = P(A)P(B),$$

was nun die folgende formale Definition nahelegt.

**Definition 4.1.**  $(\Omega, \mathcal{A}, P)$  sei ein Wahrscheinlichkeitsraum. Zwei Ereignisse  $A, B \in \mathcal{A}$  heißen **stochastisch unabhängig**, falls

$$P(A \cap B) = P(A)P(B).$$

Die Begriffsbildung der stochastischen Unabhängigkeit wurde hergeleitet unter der Voraussetzung, dass  $P(A)$  weder 0 noch 1 ist. Die gegebene Definition schließt aber den Fall von  $P(A) \in \{0, 1\}$  ein. Es ist leicht zu sehen, dass zwei Ereignisse  $A$  und  $B$  stets stochastisch unabhängig sind, falls  $P(A) \in \{0, 1\}$ .

Wir werden die Definition der stochastischen Unabhängigkeit nun auf den Fall von mehr als zwei Ereignissen erweitern. Dazu betrachten wir das einfache Beispiel des einmaligen Ziehens aus einer Urne mit vier von 1 bis 4 durchnummerierten, ansonsten gleichartigen Kugeln. Dies wird adäquat beschrieben durch ein Modell mit  $\Omega = \{1, 2, 3, 4\}$ ,  $\mathcal{A} = 2^\Omega$  und  $P(\{\omega\}) = 1/4 \forall \omega \in \Omega$ . Wir betrachten die Ereignisse  $A = \{1, 2\}$ ,  $B = \{1, 3\}$  und  $C = \{1, 4\}$ . Nun gelten

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

d.h., jeweils zwei dieser Ereignisse sind stochastisch unabhängig. Andererseits erhalten wir, dass

$$P(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = P(A)P(B)P(C).$$

Dies erklärt die nun folgende Definition.

**Definition 4.2.**  $(\Omega, \mathcal{A}, P)$  sei ein Wahrscheinlichkeitsraum und  $(A_i)_{i \in I}$  sei eine Familie von Ereignissen, wobei  $I$  eine beliebige Indexmenge ist. Dann heißen die Ereignisse

- **paarweise stochastisch unabhängig**, falls

$$P(A_i \cap A_j) = P(A_i)P(A_j) \quad \forall i, j \in I \text{ mit } i \neq j,$$

- **(vollständig) stochastisch unabhängig**, falls für jede endliche und nichtleere Menge  $I_0 \subseteq I$  gilt

$$P\left(\bigcap_{i \in I_0} A_i\right) = \prod_{i \in I_0} P(A_i).$$

**Bemerkung 4.3.** Aus  $P\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i)$  folgt im Allgemeinen **nicht** die (vollständige) stochastische Unabhängigkeit der Ereignisse  $A_1, \dots, A_n$ . Hier ist ein einfaches Gegenbeispiel:

Auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  sei ein Ereignis  $A \in \mathcal{A}$  mit  $P(A) \in (0, 1)$  gegeben. Wir definieren  $A_1 = A_2 := A$  sowie  $A_3 = \emptyset$ . Dann gilt

$$P(A_1 \cap A_2 \cap A_3) = 0 = P(A_1)P(A_2)P(A_3).$$

Andererseits gilt

$$P(A_1 \cap A_2) = P(A) > P(A)^2 = P(A_1)P(A_2),$$

d.h.,  $A_1, A_2, A_3$  sind **nicht** stochastisch unabhängig.

Falls  $\Omega$  eine nichtleere Menge ist und  $A \subseteq \Omega$ , so ist das Mengensystem  $\{\emptyset, A, A^c, \Omega\}$  die kleinste  $\sigma$ -Algebra in  $\Omega$ , welche die Menge  $A$  enthält. Etwas hochgestochen wird  $\{\emptyset, A, A^c, \Omega\}$  auch die von  $A$  erzeugte  $\sigma$ -Algebra genannt. Die folgende Aussage zeigt, dass sich die Eigenschaft der stochastischen Unabhängigkeit von Ereignissen auf die jeweiligen erzeugten  $\sigma$ -Algebren überträgt.

**Lemma 4.4.**  $(\Omega, \mathcal{A}, P)$  sei ein Wahrscheinlichkeitsraum und  $(A_i)_{i \in I}$  sei eine Familie von stochastisch unabhängigen Ereignissen. Weiter seien  $B_i \in \{\emptyset, A_i, A_i^c, \Omega\}$  ( $i \in I$ ) beliebig. Dann ist  $(B_i)_{i \in I}$  ebenfalls eine Familie von stochastisch unabhängigen Ereignissen.

*Beweis.* Wir müssen zeigen, dass für eine beliebige nichtleere und endliche Indexmenge  $I_0 \subseteq I$  und beliebige Ereignisse  $B_i \in \{\emptyset, A_i, A_i^c, \Omega\}$  ( $i \in I_0$ ) die Eigenschaft

$$P\left(\bigcap_{i \in I_0} B_i\right) = \prod_{i \in I_0} P(B_i) \quad (4.1)$$

erfüllt ist. Wir beweisen dies durch eine vollständige Induktion über die Kardinalität der Mengen  $I_0$ .

Der Induktionsanfang erfordert keinerlei Überlegung, da (4.1) im Falle von  $\#I_0 = 1$  trivialerweise erfüllt ist. Wir nehmen nun an, dass die Eigenschaft (4.1) für beliebige Indexmengen  $I_0$  mit  $\#I_0 = n-1 \geq 1$  und beliebige  $B_i \in \{\emptyset, A_i, A_i^c, \Omega\}$  ( $i \in I_0$ ) erfüllt ist. Es sei nun  $I_0$  eine beliebige Indexmenge mit  $\#I_0 = n$ . Zur Vereinfachung der Darstellung nehmen wir an, dass  $I_0 = \{1, \dots, n\}$  sei. Es seien  $B_1 \in \{\emptyset, A_1, A_1^c, \Omega\}, \dots, B_n \in \{\emptyset, A_n, A_n^c, \Omega\}$  beliebig. Es ist also zu zeigen, dass

$$P\left(B_1 \cap \dots \cap B_n\right) = P(B_1) \cdots P(B_n). \quad (4.2)$$

Wegen der vorausgesetzten Unabhängigkeit der  $A_i$  gilt

$$P\left(A_1 \cap \dots \cap A_n\right) = P(A_1) \cdots P(A_n).$$

Wir werden in dieser Gleichung sukzessive die Ereignisse  $A_i$  durch die jeweiligen  $B_i$  ersetzen und dabei zeigen, dass die „Produkteigenschaft“ erhalten bleibt. Wir nehmen daher an, dass für  $k \in \{1, \dots, n\}$  die Gleichheit

$$P\left(B_1 \cap \dots \cap B_{k-1} \cap A_k \cap \dots \cap A_n\right) = P(B_1) \cdots P(B_{k-1})P(A_k) \cdots P(A_n).$$

bereits gezeigt wurde und werden zeigen, dass daraus

$$P\left(B_1 \cap \dots \cap B_k \cap A_{k+1} \cap \dots \cap A_n\right) = P(B_1) \cdots P(B_k)P(A_{k+1}) \cdots P(A_n). \quad (4.3)$$

folgt. Wir unterscheiden nun zwischen den vier möglichen Fällen für das Ereignis  $B_k$ :

- 1) Falls  $B_k = A_k$ , so ist (4.3) bereits gezeigt.
- 2) Falls  $B_k = \emptyset$ , so sind beide Seiten von (4.3) Null.
- 3) Falls  $B_k = \Omega$ , so können wir die Induktionsvoraussetzung nutzen, welche besagt, dass die gewünschte Eigenschaft für  $\#I_0 = n-1$  gilt. Es gilt nämlich

$$\begin{aligned} P\left(B_1 \cap \dots \cap B_k \cap A_{k+1} \cap \dots \cap A_n\right) &= P\left(B_1 \cap \dots \cap B_{k-1} \cap A_{k+1} \cap \dots \cap A_n\right) \\ &\stackrel{IV}{=} P(B_1) \cdots P(B_{k-1})P(A_{k+1}) \cdots P(A_n) \\ &= P(B_1) \cdots P(B_k)P(A_{k+1}) \cdots P(A_n). \end{aligned}$$

4) Falls  $B_k = A_k^c$ , so folgt

$$\begin{aligned}
& P\left(B_1 \cap \cdots \cap B_k \cap A_{k+1} \cap \cdots \cap A_n\right) \\
&= P\left(B_1 \cap \cdots \cap B_{k-1} \cap (\Omega \setminus A_k) \cap A_{k+1} \cap \cdots \cap A_n\right) \\
&= P\left(B_1 \cap \cdots \cap B_{k-1} \cap \Omega \cap A_{k+1} \cap \cdots \cap A_n\right) \\
&\quad - P\left(B_1 \cap \cdots \cap B_{k-1} \cap A_k \cap A_{k+1} \cap \cdots \cap A_n\right) \\
&= P(B_1) \cdots P(B_{k-1}) P(A_{k+1}) \cdots P(A_n) \\
&\quad - P(B_1) \cdots P(B_{k-1}) P(A_k) P(A_{k+1}) \cdots P(A_n) \\
&= P(B_1) \cdots P(B_{k-1}) \underbrace{(1 - P(A_k))}_{P(B_k)} P(A_{k+1}) \cdots P(A_n).
\end{aligned}$$

In allen vier Fällen gilt also (4.3), woraus schließlich (4.2) folgt. Damit ist der Induktionsbeweis beendet und (4.1) ist bewiesen.  $\square$

Mit dem Resultat von Lemma 4.4 sind wir jetzt in der Lage, das in Beispiel 1.5 eingeführte Verteilungsgesetz fundiert zu begründen. Wir setzen also voraus, dass ein Zufallsexperiment  $n$  Mal hintereinander durchgeführt wird und dass die Erfolgswahrscheinlichkeit jeweils  $p \in [0, 1]$  beträgt. Wir nehmen an, dass jene Ereignisse, dass in den einzelnen Versuchen Erfolge eintreten, stochastisch unabhängig sind. Wir erarbeiten uns jetzt eine fundierte Begründung für die bereits in Beispiel 1.5 angegebene Wahrscheinlichkeit, dass insgesamt genau  $k$  Erfolge ( $k \in \{0, 1, \dots, n\}$ ) eintreten.

Wir beschreiben die Gesamtheit der möglichen Versuchsausgänge durch

$$\Omega = \{0, 1\}^n = \left\{ (\omega_1, \dots, \omega_n) : \omega_i \in \{0, 1\} \right\}.$$

Wir werden das entsprechende Wahrscheinlichkeitsmaß auf der Potenzmenge  $\mathcal{A} = 2^\Omega$  vom Grundraum  $\Omega$  festlegen und definieren zunächst noch die folgenden Ereignisse:

$$A_{i,1} = \text{„Erfolg im } i\text{-ten Versuch“} = \{\omega \in \Omega : \omega_i = 1\},$$

$$A_{i,0} = \text{„Misserfolg im } i\text{-ten Versuch“} = A_{i,1}^c = \{\omega \in \Omega : \omega_i = 0\}.$$

Nun ist klar, dass

$$P(A_{i,j}) = \begin{cases} p, & \text{falls } j = 1, \\ 1 - p, & \text{falls } j = 0 \end{cases} = p^j (1 - p)^{1-j}.$$

Es sei nun  $\omega = (\omega_1, \dots, \omega_n) \in \Omega$  ein beliebiges Elementarereignis. Da nach Voraussetzung die Ereignisse  $A_{1,1}, \dots, A_{n,1}$  stochastisch unabhängig sind, so folgt nach Lemma 4.4, dass auch die Ereignisse  $A_{1,\omega_1}, \dots, A_{n,\omega_n}$  diese Eigenschaft besitzen. Daher folgt

$$\begin{aligned}
P(\{\omega\}) &= P\left(A_{1,\omega_1} \cap \cdots \cap A_{n,\omega_n}\right) = \prod_{i=1}^n P\left(A_{i,\omega_i}\right) \\
&= \prod_{i=1}^n p^{\omega_i} (1 - p)^{1-\omega_i} = p^{\sum_{i=1}^n \omega_i} (1 - p)^{n - \sum_{i=1}^n \omega_i}.
\end{aligned}$$

Da  $\#\{\omega \in \Omega: \sum_{i=1}^n \omega_i = k\} = \binom{n}{k}$  gilt, so folgt für die Wahrscheinlichkeit, dass genau  $k$  Erfolge eintreten:

$$P\left(\{\omega \in \Omega: \sum_{i=1}^n \omega_i = k\}\right) = \sum_{\omega: \sum_{i=1}^n \omega_i = k} p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}.$$

## 5 Zufallsvariable

Zum Einstieg in dieses Kapitel beginnen wir wieder mit einem einfachen Beispiel, mit dem Würfeln beim Spiel „Die Siedler von Catan“. Wie bereits geschildert, wird dabei gleichzeitig mit zwei (unterscheidbaren) Zahlenwürfeln sowie einem weiteren „Ereigniswürfel“ gespielt. Hierbei könnten verschiedene Aspekte von Interesse sein:

- Augensumme beider Zahlenwürfel
- Augenzahl des roten Zahlenwürfels
- Symbol des „Ereigniswürfels“

Bei der Modellierung ergeben sich somit die folgenden Erfordernisse:

- den interessierenden Ereignissen sollen Wahrscheinlichkeiten zugeordnet werden
- das „Zusammenwirken“ dieser Ereignisse muss definiert werden (Beispielsweise kann die Augensumme nicht 10 sein, wenn der rote Zahlenwürfel „1“ zeigt.)

Um diesen Erfordernissen gleichzeitig gerecht zu werden, gehen wir folgendermaßen vor: Wir wählen einen geeigneten Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ , welcher das **gesamte** Zufallsexperiment beschreibt. Wenn wir der Einfachheit halber annehmen, dass die Seiten des „Ereigniswürfels“ zusätzlich mit den Zahlen von 1 bis 6 beschriftet sind, so bietet sich folgendes Modell an.

$$\begin{aligned} \Omega &= \left\{ (\omega_1, \omega_2, \omega_3) : \omega_i \in \{1, \dots, 6\} \right\}, \\ P(\{\omega\}) &= 1/6^3 \quad \forall \omega \in \Omega, \quad \mathcal{A} = 2^\Omega. \end{aligned}$$

Die uns interessierenden Aspekte können nun durch **Abbildungen**  $X_i: \Omega \rightarrow \mathbb{R}$  beschrieben werden:

$$\begin{aligned} X_1((\omega_1, \omega_2, \omega_3)) &= \omega_1 + \omega_2, \\ X_2((\omega_1, \omega_2, \omega_3)) &= \omega_1, \\ X_3((\omega_1, \omega_2, \omega_3)) &= \begin{cases} 1, & \text{falls } \omega_3 \leq 3, \\ 0, & \text{falls } \omega_3 > 3. \end{cases} \end{aligned}$$

Was bringt nun diese etwas sperrig wirkende Herangehensweise? Das Ereignis „Augensumme beider Zahlenwürfel ist 10“ kann als eine Teilmenge von  $\Omega$  dargestellt werden, nämlich  $\{\omega \in \Omega: X_1(\omega) = 10\} = X_1^{-1}(\{10\})$ . Hierbei bezeichnet  $X_1^{-1}$  **nicht** die Umkehrfunktion von  $X_1$  (welche in diesem Fall auch nicht existiert), sondern  $X_1^{-1}(\{10\})$  bezeichnet das sogenannte **vollständige  $X_1$ -Urbild** der Menge  $\{10\}$ . An dieser Stelle muss man nun, zumindest bei komplexeren Modellen, vorsichtig sein. Wir sind an Wahrscheinlichkeiten interessiert, dass eine Zufallsvariable  $X$  Werte in einer gewissen Menge  $B$  annimmt. Das zugrundeliegende Wahrscheinlichkeitsmaß  $P$  ist auf dem System  $\mathcal{A}$  von Teilmengen von  $\Omega$  definiert. Die gesuchte Wahrscheinlichkeit ist somit wohldefiniert, falls die Menge  $X^{-1}(B) = \{\omega \in \Omega: X(\omega) \in B\}$  in der  $\sigma$ -Algebra  $\mathcal{A}$  tatsächlich enthalten ist. Aus diesem Grunde sind in der Wahrscheinlichkeitstheorie die folgenden Begriffe üblich.

**Definition 5.1.**  $(\Omega, \mathcal{A}, P)$  sei ein Wahrscheinlichkeitsraum.

Eine Abbildung  $X: \Omega \rightarrow \mathbb{R}$  heißt **messbar**, falls

$$X^{-1}((-\infty, x]) = \{\omega \in \Omega: X(\omega) \leq x\} \in \mathcal{A} \quad \forall x \in \mathbb{R}$$

gilt. ( $X^{-1}(B)$  heißt **vollständiges  $X$ -Urbild** der Menge  $B$ .)

Eine **Zufallsvariable**  $X$  auf  $(\Omega, \mathcal{A}, P)$  ist eine messbare Funktion  $X: \Omega \rightarrow \mathbb{R}$ .

Eine Zufallsvariable  $X$  heißt **diskrete Zufallsvariable**, falls  $\{X(\omega): \omega \in \Omega\}$  eine endliche oder abzählbar unendliche Menge in  $\mathbb{R}$  ist.

Für die Schulpraxis relativiert sich wohl die Wichtigkeit dieser Begriffe, da dort in der Regel mit einfachen Zufallsmodellen gearbeitet wird. Falls beispielsweise  $\mathcal{A} = 2^\Omega$  ist, so ist offensichtlich jede Abbildung  $X: \Omega \rightarrow \mathbb{R}$  messbar. Dennoch sollen diese in der Wahrscheinlichkeitstheorie üblichen Details an dieser Stelle nicht verschwiegen werden, da sie beispielsweise in jedem (ernst zu nehmenden) Lehrbuch über Wahrscheinlichkeitstheorie vorkommen. Das folgende Lemma vermittelt einen ersten Eindruck, Mengen welchen Typs Wahrscheinlichkeiten zugeordnet werden können.

**Lemma 5.2.**  $(\Omega, \mathcal{A}, P)$  sei ein Wahrscheinlichkeitsraum und  $X: \Omega \rightarrow \mathbb{R}$  sei eine Zufallsvariable. Dann gelten:

$$(i) \quad X^{-1}((a, b]) \in \mathcal{A}, \quad X^{-1}(\{a\}) \in \mathcal{A}, \quad X^{-1}([a, b]) \in \mathcal{A} \quad \forall a, b \in \mathbb{R}.$$

$$(ii) \quad \text{Falls zusätzlich } \Omega_X := \{X(\omega): \omega \in \Omega\} \text{ höchstens abzählbar ist, so gilt } X^{-1}(B) \in \mathcal{A} \quad \forall B \subseteq \mathbb{R}.$$

*Beweis.* (i) Es seien  $a, b \in \mathbb{R}$  mit  $a \leq b$  beliebig. Indem wir die Struktureigenschaften

der  $\sigma$ -Algebra  $\mathcal{A}$  ausnutzen sehen wir, dass

$$\begin{aligned} X^{-1}((a, b]) &= \{\omega \in \Omega: X(\omega) \in (a, b]\} \\ &= \{\omega \in \Omega: X(\omega) \leq b\} \setminus \{\omega \in \Omega: X(\omega) \leq a\} \\ &= \underbrace{X^{-1}((-\infty, b])}_{\in \mathcal{A}} \setminus \underbrace{X^{-1}((-\infty, a])}_{\in \mathcal{A}} \in \mathcal{A}, \\ X^{-1}(\{a\}) &= \bigcap_{n=1}^{\infty} X^{-1}((a - 1/n, a]) \in \mathcal{A}, \\ X^{-1}([a, b]) &= \bigcap_{n=1}^{\infty} X^{-1}((a - 1/n, b]) \in \mathcal{A}. \end{aligned}$$

(Hier wurde benutzt, dass Durchschnitte und (Mengen)differenzen von Mengen aus  $\mathcal{A}$  selbst in  $\mathcal{A}$  liegen.)

(ii) Es sei  $B \subseteq \mathbb{R}$  beliebig. Dann gilt

$$X^{-1}(B) = \bigcup_{x \in B} X^{-1}(\{x\}) = \bigcup_{x \in B \cap \Omega_X} \underbrace{X^{-1}(\{x\})}_{\in \mathcal{A}} \in \mathcal{A}.$$

(Hier wurde benutzt, dass die Menge  $B \cap \Omega_X$  höchstens abzählbar ist und eine abzählbare Vereinigung von Mengen aus  $\mathcal{A}$  ebenfalls in  $\mathcal{A}$  liegt.)

□

Wir betrachten kurz zwei einfache Beispiele welche zeigen, dass die Forderung der Messbarkeit keine störende Einschränkung darstellen muss.

(i) (Augensumme beim zweimaligen Würfeln)

Als zugrundeliegender Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  bietet sich an:

$$\Omega = \{1, \dots, 6\} \times \{1, \dots, 6\}, \quad \mathcal{A} = 2^\Omega, \quad P(\{\omega\}) = \frac{1}{36} \quad \forall \omega \in \Omega.$$

Eine geeignete Zufallsvariable  $X$  ist dann gegeben durch

$$X(\omega) = \omega_1 + \omega_2 \quad \forall \omega \in \Omega.$$

Da in diesem Fall  $\mathcal{A} = 2^\Omega$  gilt, ist die Abbildung  $X: \Omega \rightarrow \mathbb{R}$  offensichtlich auch messbar.

(ii) (Indikatorvariable)

$(\Omega, \mathcal{A}, P)$  sei ein beliebiger Wahrscheinlichkeitsraum und  $A \in \mathcal{A}$  sei ein Ereignis. Dann ist  $\mathbb{1}_A: \Omega \rightarrow \mathbb{R}$  mit

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

eine Zufallsvariable auf  $(\Omega, \mathcal{A}, P)$ , welche das Eintreten bzw. Nichteintreten des Ereignisses  $A$  anzeigt. Die erforderliche Messbarkeit lässt sich leicht durch Betrachten der möglichen vollständigen Urbilder überprüfen. Es gilt

$$\mathbb{1}_A^{-1}((-\infty, x]) = \begin{cases} \emptyset, & \text{falls } x < 0, \\ A^c, & \text{falls } x \in [0, 1), \\ \Omega, & \text{falls } x \geq 1. \end{cases}$$

Da mit  $A$  auch die Mengen  $\emptyset$ ,  $A^c$  und  $\Omega$  in  $\mathcal{A}$  liegen, folgt die Messbarkeit der Abbildung  $\mathbb{1}_A$ .

Wir wollen den mit Zufallsvariablen verbundenen Ereignissen auch Wahrscheinlichkeiten zuordnen:

**Definition 5.3.**  $(\Omega, \mathcal{A}, P)$  sei ein Wahrscheinlichkeitsraum und  $X: \Omega \rightarrow \mathbb{R}$  sei eine Zufallsvariable. Dann ist

$$\mathcal{A}_X := \{B \subseteq \mathbb{R}: X^{-1}(B) \in \mathcal{A}\}$$

eine  $\sigma$ -Algebra in  $\mathbb{R}$ . (siehe auch Übungsaufgabe 14)

Die Abbildung  $P^X: \mathcal{A}_X \rightarrow [0, 1]$  mit

$$P^X(B) := P(\{\omega \in \Omega: X(\omega) \in B\}) = P(X^{-1}(B)) \quad \forall B \in \mathcal{A}_X$$

heißt **Verteilung von  $X$  unter  $P$** .

**Lemma 5.4.**  $(\Omega, \mathcal{A}, P)$  sei ein Wahrscheinlichkeitsraum und  $X: \Omega \rightarrow \mathbb{R}$  sei eine Zufallsvariable. Dann ist  $P^X: \mathcal{A}_X \rightarrow [0, 1]$  ein Wahrscheinlichkeitsmaß auf  $\mathcal{A}_X$ .

*Beweis.* Wir überprüfen die Axiome eines Wahrscheinlichkeitsmaßes:

(i) Für beliebiges  $B \in \mathcal{A}_X$  gilt:

$$P^X(B) = P(X^{-1}(B)) \geq 0.$$

(ii)  $P^X(\mathbb{R}) = P(\underbrace{X^{-1}(\mathbb{R})}_{=\Omega}) = 1.$

(iii)  $B_1, B_2, \dots \in \mathcal{A}_X$  seien disjunkt. Dann sind  $X^{-1}(B_1), X^{-1}(B_2), \dots \in \mathcal{A}$  ebenfalls disjunkt und es folgt

$$\begin{aligned} P^X\left(\bigcup_{i=1}^{\infty} B_i\right) &= P\left(X^{-1}\left(\bigcup_{i=1}^{\infty} B_i\right)\right) = P\left(\bigcup_{i=1}^{\infty} X^{-1}(B_i)\right) \\ &= \sum_{i=1}^{\infty} P\left(X^{-1}(B_i)\right) = \sum_{i=1}^{\infty} P^X(B_i). \end{aligned}$$

□

### Beispiele (Fortsetzung)

Da es sich bei den hier betrachteten Beispielen um Zufallsvariable mit einem höchstens endlichen Wertebereich handelt, genügt es, die Verteilung  $P^X$  für einelementige Ereignisse zu bestimmen. Der Rest folgt dann aus der  $\sigma$ -Additivität.

(i) Augensumme

In diesem Fall gilt  $\{X(\omega) : \omega \in \Omega\} = \{2, 3, \dots, 12\}$ . Für  $k \in \{2, 3, \dots, 12\}$  erhalten wir

$$\begin{aligned} P^X(\{k\}) &= P(X^{-1}(\{k\})) = P(\{\omega \in \Omega : \omega_1 + \omega_2 = k\}) \\ &= \frac{\#\{\omega \in \Omega : \omega_1 + \omega_2 = k\}}{36} = \frac{6 - |7 - k|}{36}. \end{aligned}$$

(ii) Indikatorvariable

Hier ist  $\{X(\omega) : \omega \in \Omega\} = \{0, 1\}$  und es gelten

$$\begin{aligned} P^{1_A}(\{1\}) &= P(\mathbb{1}_A^{-1}(\{1\})) = P(A), \\ P^{1_A}(\{0\}) &= P(\mathbb{1}_A^{-1}(\{0\})) = P(A^c). \end{aligned}$$

## Der Erwartungswert einer Zufallsvariable

Der Erwartungswert einer Zufallsvariable soll einen „mittleren“ Wert ausdrücken, um welchen eine Zufallsgröße schwankt. Dabei erweist sich ein mit den entsprechenden Wahrscheinlichkeiten gewichtetes arithmetisches Mittel als in vielerlei Hinsicht gut geeignet. Falls  $X : \Omega \rightarrow \mathbb{R}$  eine diskrete Zufallsvariable auf  $(\Omega, \mathcal{A}, P)$  ist mit  $\{X(\omega) : \omega \in \Omega\} = \{x_1, \dots, x_N\}$  bzw.  $\{X(\omega) : \omega \in \Omega\} = \{x_1, x_2, \dots\}$  ( $x_i \neq x_j$  für  $i \neq j$ ), so liegt folgende Definition nahe:

$$EX := \sum_{i \geq 1} x_i P(\underbrace{\{\omega \in \Omega : X(\omega) = x_i\}}_{\in \mathcal{A} \text{ da } X \text{ messbar}}).$$

Mit diesem etwas vorschnellem Vorpreschen wird jedoch ein Detail außer Acht gelassen: Es ist nicht ausgeschlossen, dass  $X$  abzählbar unendlich viele, sowohl positive als auch negative Werte annimmt. In diesem Fall kann es sein, dass der Wert der obigen unendlichen Reihe von der gewählten Durchnummerierung der Werte von  $X$  und daher von der entstehenden Summationsreihenfolge abhängt. Dies ist jedoch problematisch, da es in solchen Fällen oftmals keine kanonische Summationsreihenfolge gibt. Außerdem wird das Rechnen mit Erwartungswerten öfters eine Änderung der Summationsreihenfolge erfordern. Mit der folgenden Definition wird dieses Problem berücksichtigt.

**Definition 5.5.**  $X$  sei eine diskrete Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ , wobei  $\{X(\omega): \omega \in \Omega\} = \{x_1, \dots, x_N\}$  bzw.  $\{X(\omega): \omega \in \Omega\} = \{x_1, x_2, \dots\}$  ( $x_i \neq x_j$  für  $i \neq j$ ).

(i) Falls  $\{X(\omega): \omega \in \Omega\} = \{x_1, \dots, x_N\}$ , so

$$EX := \sum_{i=1}^N x_i P^X(\{x_i\}).$$

(ii) Falls  $\{X(\omega): \omega \in \Omega\} = \{x_1, x_2, \dots\}$  und  $X(\omega) \geq 0 \forall \omega \in \Omega$ , so

$$EX := \sum_{i=1}^{\infty} x_i P^X(\{x_i\}).$$

(iii) Im allgemeinen Fall zerlegt man  $X$  in seinen nichtnegativen Anteil  $X^+$  und seinen nichtpositiven Anteil  $X^-$ , d.h.,  $X^+(\omega) = \max\{X(\omega), 0\}$ ,  $X^-(\omega) = X^+(\omega) - X(\omega)$ . Falls nun  $EX^+ < \infty$  oder  $EX^- < \infty$ , so

$$EX := EX^+ - EX^- = \sum_{i: x_i > 0} x_i P^X(\{x_i\}) - \sum_{i: x_i < 0} (-x_i) P^X(\{x_i\}).$$

In diesem Fall gilt auch

$$EX := \sum_{i=1}^{\infty} x_i P^X(\{x_i\}),$$

wobei die unendliche Reihe auf der rechten Seite unabhängig von der gewählten Summationsreihenfolge ist.

Falls  $EX^+ = \infty$  und  $EX^- = \infty$ , so ist der Erwartungswert von  $X$  **nicht** definiert.

$EX$  heißt **Erwartungswert** von  $X$ .

Im Hinblick auf den Schulunterricht sollte man sich nicht von dieser etwas länglichen Definition abschrecken lassen. Solange Zufallsvariable mit endlich vielen möglichen Werten oder nichtnegative diskrete Zufallsvariable behandelt werden, ist die Existenz des Erwartungswertes stets gesichert und dessen Definition kann auf direktem Wege wie in (i) bzw. (ii) erfolgen.

**Lemma 5.6.**  $X$  sei eine diskrete Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ . Falls  $\Omega$  höchstens abzählbar ist,  $\mathcal{A} = 2^\Omega$  und der Erwartungswert von  $X$  existiert, so gilt

$$EX = \sum_{\omega \in \Omega} X(\omega) P(\{\omega\}).$$

*Beweis.* Wir betrachten zunächst die Erwartungswerte von  $X^+$  und  $X^-$ , da diese in jedem Fall existieren. Es gelten

$$\begin{aligned} EX^+ &= \sum_{i: x_i > 0} x_i P(\{\omega: X(\omega) = x_i\}) \\ &= \sum_{i: x_i > 0} x_i \sum_{\omega: X(\omega) = x_i} P(\{\omega\}) = \sum_{\omega \in \Omega} X^+(\omega) P(\{\omega\}) \end{aligned}$$

sowie, mit analoger Rechnung,

$$EX^- = \sum_{\omega \in \Omega} X^-(\omega) P(\{\omega\}).$$

Falls nun der Erwartungswert von  $X$  existiert, so ist mindestens einer der Erwartungswerte  $EX^+$  bzw.  $EX^-$  endlich. Dann folgt

$$EX = EX^+ - EX^- = \sum_{\omega: X(\omega) > 0} X(\omega) P(\{\omega\}) - \sum_{\omega: X(\omega) < 0} (-X(\omega)) P(\{\omega\}) = \sum_{\omega \in \Omega} X(\omega) P(\{\omega\}).$$

(Die Summe auf der rechten Seite ist unabhängig von der Summationsreihenfolge.)  $\square$

**Lemma 5.7.**  $X$  sei eine diskrete Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ , wobei  $\{X(\omega): \omega \in \Omega\} = \{x_1, \dots, x_N\}$  bzw.  $\{X(\omega): \omega \in \Omega\} = \{x_1, x_2, \dots\}$  ( $x_i \neq x_j$  für  $i \neq j$ ).  $f: \mathbb{R} \rightarrow \mathbb{R}$  sei eine beliebige Funktion. Dann gelten:

(i)  $Y := f(X): \Omega \rightarrow \mathbb{R}$  mit  $Y(\omega) := f(X(\omega)) \forall \omega \in \Omega$  ist ebenfalls eine diskrete Zufallsvariable auf  $(\Omega, \mathcal{A}, P)$ .

(ii) Falls der Erwartungswert von  $Y$  existiert, so

$$EY = \sum_{j \geq 1} f(x_j) P^X(\{x_j\}).$$

*Beweis.* (i) Da  $\{Y(\omega): \omega \in \Omega\}$  offensichtlich höchstens abzählbar ist, muss nur noch die Messbarkeit der Abbildung  $Y: \Omega \rightarrow \mathbb{R}$  gezeigt werden. Für beliebiges  $y \in \mathbb{R}$  gilt

$$\begin{aligned} Y^{-1}((-\infty, y]) &= \{\omega \in \Omega: f(X(\omega)) \leq y\} \\ &= \bigcup_{i: f(x_i) \leq y} \underbrace{\{\omega \in \Omega: X(\omega) = x_i\}}_{\in \mathcal{A}} \in \mathcal{A}. \end{aligned}$$

(In der letzten Zeile handelt es sich um eine Vereinigung von **höchstens abzählbar** vielen Mengen aus  $\mathcal{A}$ .)

- (ii) Die Zufallsvariable  $Y$  nehme Werte  $y_1, \dots, y_M$  bzw.  $y_1, y_2, \dots$  an ( $y_i \neq y_j$  für  $i \neq j$ ). Falls  $Y(\omega) \geq 0 \forall \omega \in \Omega$ , so ist die Existenz des Erwartungswertes von  $Y$  gesichert und es gilt:

$$\begin{aligned}
 EY &= \sum_i y_i P(\{\omega: f(X(\omega)) = y_i\}) \\
 &= \sum_i y_i P\left(\bigcup_{j: f(x_j)=y_i} \{\omega: X(\omega) = x_j\}\right) \\
 &= \sum_i \sum_{j: f(x_j)=y_i} f(x_j) P(\{\omega: X(\omega) = x_j\}) \\
 &= \sum_j f(x_j) P^X(\{x_j\}).
 \end{aligned}$$

Falls  $Y$  sowohl positive als auch negative Werte annimmt, so zerlegen wir  $Y = Y^+ - Y^-$ . Nach den vorherigen Betrachtungen gelten

$$EY^+ = \sum_{j: f(x_j)>0} f(x_j) P^X(\{x_j\})$$

und

$$EY^- = \sum_{j: f(x_j)<0} (-f(x_j)) P^X(\{x_j\}).$$

Die vorausgesetzte Existenz des Erwartungswertes von  $Y$  bedeutet, dass mindestens einer der beiden Erwartungswerte  $EY^+$  oder  $EY^-$  endlich ist. Somit folgt

$$EY = EY^+ - EY^- = \sum_{j \geq 1} f(x_j) P^X(\{x_j\}).$$

(Die rechte Seite ist hierbei unabhängig von der gewählten Durchnummerierung der Werte von  $Y$ .)

□

**Lemma 5.8.**  $(\Omega, \mathcal{A}, P)$  sei ein Wahrscheinlichkeitsraum und  $X: \Omega \rightarrow \mathbb{R}$  sowie  $Y: \Omega \rightarrow \mathbb{R}$  seien diskrete Zufallsvariable. Dann gelten:

- (i) Falls  $0 \leq X(\omega) \leq Y(\omega) \forall \omega \in \Omega$ , so folgt  $0 \leq EX \leq EY$ .
- (ii) Falls  $E|X| < \infty$ , so folgen  $EX^+ < \infty$  und  $EX^- < \infty$ . Somit existiert auch der Erwartungswert von  $X$  und ist endlich.
- (iii) Falls  $\alpha, \beta \in \mathbb{R}$ , so ist  $Z: \Omega \rightarrow \mathbb{R}$  mit  $Z(\omega) = \alpha X(\omega) + \beta Y(\omega)$  ebenfalls eine diskrete Zufallsvariable. Falls zusätzlich  $EX$  und  $EY$  endlich sind, so gilt

$$EZ = \alpha EX + \beta EY.$$

*Beweis.* Wir bezeichnen mit  $x_1, \dots, x_N$  bzw.  $x_1, x_2, \dots$  sowie  $y_1, \dots, y_M$  bzw.  $y_1, y_2, \dots$  ( $x_i \neq x_j, y_i \neq y_j$  für  $i \neq j$ ) die möglichen Werte von  $X$  bzw.  $Y$ .

- (i) Da  $X$  und  $Y$  nichtnegative Zufallsvariable sind, so ist die Existenz ihrer Erwartungswerte gesichert.  $EX \geq 0$  ist angesichts der Definition des Erwartungswertes als ein gewichtetes arithmetisches Mittel offensichtlich.

Nun gilt

$$\begin{aligned}
 EX &= \sum_i x_i P(\{\omega: X(\omega) = x_i\}) \\
 &= \sum_i x_i \sum_j P(\{\omega: X(\omega) = x_i\} \cap \{\omega: Y(\omega) = y_j\}) \\
 &= \sum_{i,j} x_i P(\{\omega: X(\omega) = x_i, Y(\omega) = y_j\}) \\
 &= \sum_{(i,j): x_i \leq y_j} x_i P(\{\omega: X(\omega) = x_i, Y(\omega) = y_j\}) \\
 &\leq \sum_{(i,j): x_i \leq y_j} y_j P(\{\omega: X(\omega) = x_i, Y(\omega) = y_j\}) \\
 &= \dots = EY.
 \end{aligned}$$

Hierbei gilt das Gleichheitszeichen in der drittletzten Zeile der Formel, da  $P(\{\omega: X(\omega) > Y(\omega)\}) = 0$ .

- (ii) Da  $X^+(\omega), X^-(\omega) \leq |X(\omega)|$  für alle  $\omega \in \Omega$  gilt, so folgt nach (i), dass

$$EX^+, EX^- \leq E|X| < \infty.$$

Somit existiert der Erwartungswert von  $X$  und ist endlich.

- (iii) Offensichtlich ist  $\{Z(\omega): \omega \in \Omega\} = \{\alpha x_i + \beta y_j: i, j \geq 1\}$  höchstens abzählbar. Außerdem gilt für beliebiges  $z \in \mathbb{R}$

$$\{\omega: Z(\omega) \leq z\} = \bigcup_{(i,j): \alpha x_i + \beta y_j \leq z} \underbrace{\{\omega: X(\omega) = x_i\} \cap \{\omega: Y(\omega) = y_j\}}_{\in \mathcal{A}} \in \mathcal{A},$$

d.h.,  $Z$  ist eine diskrete Zufallsvariable.

Wir beweisen nun:

- a)  $E[\alpha X] = \alpha EX$ ,  
 b)  $E[X + Y] = EX + EY$ .

zu a) Es gilt

$$E|\alpha X| = \sum_i |\alpha x_i| P^X(\{x_i\}) = |\alpha| E|X| < \infty.$$

Daher existiert auch der Erwartungswert von  $\alpha X$  und ist endlich. Es gilt nach Aussage (ii) von Lemma 5.7, dass

$$E[\alpha X] = \sum_i \alpha x_i P^X(\{x_i\}) = \alpha EX.$$

zu b) Es seien zunächst  $X$  und  $Y$  nichtnegative Zufallsvariable. Dann existiert auch der Erwartungswert von  $X + Y$  und bei den nachfolgenden Rechnungen kann die Summationsreihenfolge beliebig vertauscht werden. Wir bezeichnen mit  $z_1, \dots, z_K$  bzw.  $z_1, z_2, \dots$  ( $z_i \neq z_j$  für  $i \neq j$ ) die möglichen Werte der Zufallsvariable  $Z := X + Y$ . Es gilt

$$\begin{aligned}
 E[Z] &= \sum_i z_i P(\{\omega: Z(\omega) = z_i\}) \\
 &= \sum_i z_i P\left(\bigcup_{(j,k): x_j+y_k=z_i} \{\omega: X(\omega) = x_j, Y(\omega) = y_k\}\right) \\
 &= \sum_i \sum_{(j,k): x_j+y_k=z_i} (x_j + y_k) P(\{\omega: X(\omega) = x_j, Y(\omega) = y_k\}) \\
 &= \sum_{j,k} (x_j + y_k) P(\{\omega: X(\omega) = x_j, Y(\omega) = y_k\}) \\
 &= \sum_j x_j \underbrace{\sum_k P(\{\omega: X(\omega) = x_j, Y(\omega) = y_k\})}_{P(\{\omega: X(\omega)=x_j\})} \\
 &\quad + \sum_k y_k \underbrace{\sum_j P(\{\omega: X(\omega) = x_j, Y(\omega) = y_k\})}_{P(\{\omega: Y(\omega)=y_k\})} \\
 &= EX + EY.
 \end{aligned}$$

Falls nun  $X$  und/oder  $Y$  nicht notwendigerweise nichtnegativ sind, so können wir folgendermaßen argumentieren: Nach Voraussetzung sind  $EX$  und  $EY$  endlich, was insbesondere  $E|X| < \infty$  und  $E|Y| < \infty$  impliziert. Wegen  $|Z(\omega)| \leq |X(\omega)| + |Y(\omega)| \forall \omega \in \Omega$  folgt

$$E|Z| \leq E[|X| + |Y|] = E|X| + E|Y| < \infty.$$

Somit sind alle oben auftretenden Reihen **absolut summierbar**. Daraus folgt, dass auch in diesem Fall bei den obigen Rechnungen die Summationsreihenfolge beliebig vertauscht werden kann und die obige Gleichungskette bleibt somit gültig. □

**Definition 5.9.**  $X$  sei eine diskrete Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ . Dann heißt

$$E[X^k] \quad \text{das } k\text{-te Moment von } X.$$

Falls zusätzlich  $EX$  endlich, so heißt

$$E[(X - EX)^k] \quad \text{das } k\text{-te zentrale Moment von } X.$$

(In beiden Fällen wird vorausgesetzt, dass diese Erwartungswerte existieren.)

Insbesondere heißen

$$\text{var}(X) = E[(X - EX)^2] \quad \text{die \textbf{Varianz} von } X$$

und

$$\sqrt{\text{var}(X)} \quad \text{die \textbf{Standardabweichung} von } X.$$

**Bemerkung 5.10.**  $X$  sei eine diskrete Zufallsvariable mit  $E[X^2] < \infty$ . Dann gelten

- (i)  $\text{var}(X) = E[X^2] - (EX)^2$ ,
- (ii) falls  $\alpha, \beta \in \mathbb{R}$ ,
- $$\text{var}(\alpha X + \beta) = \alpha^2 \text{var}(X).$$

*Beweis.* (i) Wegen  $|X(\omega)| \leq X^2(\omega) + 1$  folgt  $E|X| \leq E[X^2] + 1 < \infty$ . Somit ist  $EX$  endlich und  $\text{var}(X)$  ist wohldefiniert. Nun gilt:

$$\begin{aligned} E[(X - EX)^2] &= E[X^2 - 2X \cdot EX + (EX)^2] \\ &= E[X^2] - 2 \underbrace{E[X \cdot EX]}_{=(EX)^2} + (EX)^2 \\ &= E[X^2] - (EX)^2. \end{aligned}$$

- (ii) Nach Aussage (iii) von Lemma 5.8 gilt  $E[\alpha X + \beta] = \alpha EX + \beta$ , d.h., der Erwartungswert von  $\alpha X + \beta$  ist endlich und die Varianz ist wohldefiniert. Nun gilt

$$\begin{aligned} \text{var}(\alpha X + \beta) &= E[(\alpha X + \beta - E[\alpha X + \beta])^2] \\ &= E[(\alpha X + \beta - \alpha EX - \beta)^2] \\ &= E[\alpha^2 (X - EX)^2] = \alpha^2 \text{var}(X). \end{aligned}$$

□

Als Nächstes führen wir Maßzahlen ein, welche den Grad der Assoziation zweier Zufallsvariablen beschreiben.

**Definition 5.11.**  $X$  und  $Y$  seien diskrete Zufallsvariable auf  $(\Omega, \mathcal{A}, P)$  mit  $E[X^2] < \infty$  und  $E[Y^2] < \infty$ . Dann heißt

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

die **Kovarianz** von  $X$  und  $Y$ . Falls zusätzlich  $\text{var}(X) > 0$  und  $\text{var}(Y) > 0$  sind, so heißt

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

die **Korrelation** von  $X$  und  $Y$ .

**Bemerkung 5.12.** (i) Die in der vorangegangenen Definition auftretenden Erwartungswerte sind endlich. Neben der bereits bekannten Endlichkeit von  $EX$  und  $EY$  folgt aus der Ungleichung

$$|(X(\omega) - EX)(Y(\omega) - EY)| \leq ((X(\omega) - EX)^2 + (Y(\omega) - EY)^2)/2$$

dass

$$E|(X - EX)(Y - EY)| \leq (\text{var}(X) + \text{var}(Y))/2 < \infty.$$

Somit existiert  $E[(X - EX)(Y - EY)]$  und ist endlich.

(ii) Es gilt

$$\begin{aligned} \text{cov}(X, Y) &= E[X \cdot Y] - \underbrace{E[X \cdot EY]}_{=EX \cdot EY} - \underbrace{E[Y \cdot EX]}_{=EX \cdot EY} + \underbrace{E[EX \cdot EY]}_{=EX \cdot EY} \\ &= E[X \cdot Y] - EX \cdot EY. \end{aligned}$$

In Kapitel 4 wurde der Begriff der stochastischen Unabhängigkeit von Ereignissen eingeführt. Es sei daran erinnert, dass Ereignisse  $(A_i)_{i \in I}$  auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  stochastisch unabhängig sind, falls für alle endlichen und nichtleeren Teilmengen  $I_0$  von  $I$  die folgende Eigenschaft gilt:

$$P\left(\bigcap_{i \in I_0} A_i\right) = \prod_{i \in I_0} P(A_i).$$

Die Definition der stochastischen Unabhängigkeit von Zufallsvariablen baut auf dieser Definition auf: Wir werden Zufallsvariable stochastisch unabhängig nennen, wenn alle damit verbundenen Ereignisse stochastisch unabhängig sind. Hier ist eine formale Definition:

**Definition 5.13.**  $(X_i)_{i \in I}$  sei eine Familie von diskreten Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ . Diese Zufallsvariablen heißen **stochastisch unabhängig**, falls für alle endlichen und nichtleeren Teilmengen  $I_0$  von  $I$  die folgende Eigenschaft gilt:

$$P\left(\bigcap_{i \in I_0} \{\omega \in \Omega: X_i(\omega) \in Q_i\}\right) = \prod_{i \in I_0} P(\{\omega \in \Omega: X_i(\omega) \in Q_i\}) \quad \forall Q_i \subseteq \mathbb{R}.$$

**Bemerkung 5.14.** Zum Nachweis der stochastischen Unabhängigkeit der Zufallsvariablen  $(X_i)_{i \in I}$  genügt es, die obige „Produkteigenschaft“ für einelementige Mengen nachzuweisen. Es seien  $\Omega_i := \{X_i(\omega) : \omega \in \Omega\} = \{x_{i1}, x_{i2}, \dots\}$  bzw.  $\Omega_i := \{X_i(\omega) : \omega \in \Omega\} = \{x_{i1}, \dots, x_{iM_i}\}$  ( $i \in I$ ). Wir nehmen an, dass für jede endliche und nichtleere Teilmenge  $I_0$  von  $I$  die Eigenschaft

$$P\left(\bigcap_{i \in I_0} \{\omega \in \Omega : X_i(\omega) = x_{ij_i}\}\right) = \prod_{i \in I_0} P\left(\{\omega \in \Omega : X_i(\omega) = x_{ij_i}\}\right) \quad \forall j_i \geq 1 \quad (I \in I_0)$$

gilt. Es sei nun zur Vereinfachung der Schreibweise  $I_0 = \{1, \dots, n\}$ . Für beliebige Mengen  $Q_1, \dots, Q_n \subseteq \mathbb{R}$  erhalten wir

$$\begin{aligned} & P\left(\bigcap_{i=1}^n \{\omega : X_i(\omega) \in Q_i\}\right) \\ &= P\left(\bigcap_{i=1}^n \bigcup_{j: x_{ij} \in Q_i} \{\omega : X_i(\omega) = x_{ij}\}\right) \\ &= P\left(\bigcup_{(j_1, \dots, j_n) : x_{1j_1} \in Q_1, \dots, x_{nj_n} \in Q_n} \{\omega : X_1(\omega) = x_{1j_1}, \dots, X_n(\omega) = x_{nj_n}\}\right) \\ &= \sum_{(j_1, \dots, j_n) : x_{1j_1} \in Q_1, \dots, x_{nj_n} \in Q_n} P\left(\{\omega : X_1(\omega) = x_{1j_1}, \dots, X_n(\omega) = x_{nj_n}\}\right) \\ &= \sum_{(j_1, \dots, j_n) : x_{1j_1} \in Q_1, \dots, x_{nj_n} \in Q_n} \prod_{i=1}^n P\left(\{\omega : X_i(\omega) = x_{ij_i}\}\right) \\ &= \prod_{i=1}^n \sum_{j: x_{ij} \in Q_i} P\left(\{\omega : X_i(\omega) = x_{ij}\}\right) \\ &= \prod_{i=1}^n P\left(\{\omega : X_i(\omega) \in Q_i\}\right). \end{aligned}$$

Der folgende Satz vermittelt eine einfache aber wichtige Rechenregel für ganzzahlige unabhängige Zufallsvariable.

**Satz 5.15.**  $X_1: \Omega \rightarrow \mathbb{Z}$  und  $X_2: \Omega \rightarrow \mathbb{Z}$  seien stochastisch unabhängige ganzzahlige Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ . Dann gilt

$$P(X_1 + X_2 = k) = \sum_{l=-\infty}^{\infty} P(X_1 = l)P(X_2 = k - l) \quad \forall k \in \mathbb{Z}.$$

*Beweis.* Es gilt für beliebige  $k \in \mathbb{Z}$

$$\begin{aligned} P(X_1 + X_2 = k) &= P\left(\bigcup_{l=-\infty}^{\infty} \{\omega: X_1(\omega) = l, X_2(\omega) = k - l\}\right) \\ &= \sum_{l=-\infty}^{\infty} P(\{\omega: X_1(\omega) = l, X_2(\omega) = k - l\}) \\ &= \sum_{l=-\infty}^{\infty} P(X_1 = l)P(X_2 = k - l). \end{aligned} \quad \square$$

### Beispiel

$X_1 \sim \text{Bin}(n_1, p)$  und  $X_2 \sim \text{Bin}(n_2, p)$  seien stochastisch unabhängig. ( $P(X_i = k) = \binom{n_i}{k} p^k (1-p)^{n_i-k}$  für  $k = 0, 1, \dots, n_i$ .) Dann gilt

$$X_1 + X_2 \sim \text{Bin}(n_1 + n_2, p).$$

Natürlich ist diese Aussage leicht anschaulich nachvollziehbar: Wenn ein Zufallsexperiment  $n_i$ -mal „voneinander unabhängig“ wiederholt wird und die Wahrscheinlichkeit eines „Erfolges“ jeweils  $p$  ist, so ist die Gesamtzahl der Erfolge binomialverteilt mit Parametern  $n_i$  und  $p$ . Führt man nun nacheinander zwei voneinander unabhängige Blöcke solcher Zufallsexperimente mit jeweils  $n_1$  bzw.  $n_2$  Teilversuchen durch, so entspricht das der  $(n_1 + n_2)$ -fachen voneinander unabhängigen Wiederholung des Zufallsexperimentes und die Gesamtzahl der Erfolge ist dementsprechend binomialverteilt mit Parametern  $n_1 + n_2$  und  $p$ .

Die Korrektheit dieses Schlusses kann man mit der in Satz 5.15 angegebenen Formel leicht nachprüfen. Für  $k \in \{0, 1, \dots, n_1 + n_2\}$  (Die Zufallsvariable  $X_1 + X_2$  kann nur diese Werte annehmen.) gilt:

$$\begin{aligned} P(X_1 + X_2 = k) &= \sum_{l=-\infty}^{\infty} P(X_1 = l)P(X_2 = k - l) \\ &= \sum_{l=0}^k \binom{n_1}{l} p^l (1-p)^{n_1-l} \binom{n_2}{k-l} p^{k-l} (1-p)^{n_2-(k-l)} \\ &= p^k (1-p)^{n_1+n_2-k} \underbrace{\sum_{l=0}^k \binom{n_1}{l} \binom{n_2}{k-l}}_{=\binom{n_1+n_2}{k}}. \end{aligned}$$

**Lemma 5.16.**  $(X_i)_{i \in I}$  sei eine Familie von unabhängigen diskreten Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  und  $f_i: \mathbb{R} \rightarrow \mathbb{R}$  ( $i \in I$ ) seien beliebige Abbildungen.

Dann ist  $(f_i(X_i))_{i \in I}$  ebenfalls eine Familie von unabhängigen diskreten Zufallsvariablen.

*Beweis.*  $I_0$  sei eine beliebige endliche und nichtleere Teilmenge von  $I$ ;  $Q_i \subseteq \mathbb{R}$  ( $i \in I_0$ ) seien ebenfalls beliebig. Dann gilt

$$\begin{aligned} P\left(\bigcap_{i \in I_0} \{\omega: f_i(X_i(\omega)) \in Q_i\}\right) &= P\left(\bigcap_{i \in I_0} \{\omega: X_i(\omega) \in \underbrace{f_i^{-1}(Q_i)}_{=: R_i}\}\right) \\ &= \prod_{i \in I_0} P\left(\{\omega: X_i(\omega) \in \underbrace{f_i^{-1}(Q_i)}_{=: R_i}\}\right) \\ &= \prod_{i \in I_0} P\left(\{\omega: f_i(X_i(\omega)) \in Q_i\}\right). \end{aligned}$$

□

**Lemma 5.17.**  $X$  und  $Y$  seien stochastisch unabhängige diskrete Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ . Falls

$$(i) \quad X(\omega) \geq 0, Y(\omega) \geq 0 \quad \forall \omega \in \Omega$$

oder

$$(ii) \quad E|X| < \infty, E|Y| < \infty,$$

so folgt

$$E[X \cdot Y] = EX \cdot EY.$$

*Beweis.* (i) Es sei  $Z(\omega) := X(\omega) \cdot Y(\omega)$ . Wie üblich seien  $\{X(\omega): \omega \in \Omega\} = \{x_1, x_2, \dots\}$ ,  $\{Y(\omega): \omega \in \Omega\} = \{y_1, y_2, \dots\}$  und  $\{Z(\omega): \omega \in \Omega\} = \{z_1, z_2, \dots\}$ , wobei  $x_i \neq x_j$ ,  $y_i \neq y_j$  und  $z_i \neq z_j$  für  $i \neq j$  gelten. Nun gilt

$$\begin{aligned} E[X \cdot Y] &= EZ = \sum_k z_k P(Z = z_k) \\ &= \sum_k z_k P\left(\bigcup_{(i,j): x_i y_j = z_k} \{\omega: X(\omega) = x_i, Y(\omega) = y_j\}\right) \\ &= \sum_k z_k \sum_{(i,j): x_i y_j = z_k} P(X = x_i, Y = y_j) \\ &= \sum_{i,j} x_i y_j P(X = x_i) P(Y = y_j) = EX \cdot EY. \end{aligned}$$

- (ii) Falls  $E|X| < \infty$  und  $E|Y| < \infty$ , so folgen  $EX^+ < \infty$ ,  $EX^- < \infty$ ,  $EY^+ < \infty$  sowie  $EY^- < \infty$ . Da aus der Unabhängigkeit von  $X$  und  $Y$  auch die von  $X^\pm$  und  $Y^\pm$  folgt, so gilt nach (i)

$$E[X^\pm Y^\pm] = EX^\pm \cdot EY^\pm.$$

Wegen  $(X \cdot Y)^+(\omega) = X^+(\omega) \cdot Y^+(\omega) + X^-(\omega) \cdot Y^-(\omega)$  folgt

$$\begin{aligned} E[(X \cdot Y)^+] &= E[X^+ \cdot Y^+ + X^- \cdot Y^-] \\ &= E[X^+ \cdot Y^+] + E[X^- \cdot Y^-] \\ &= EX^+ \cdot EY^+ + EX^- \cdot EY^-. \end{aligned}$$

Analog erhalten wir wegen  $(X \cdot Y)^-(\omega) = X^+(\omega) \cdot Y^-(\omega) + X^-(\omega) \cdot Y^+(\omega)$ , dass

$$E[(X \cdot Y)^-] = EX^+ \cdot EY^- + EX^- \cdot EY^+.$$

Somit erhalten wir

$$\begin{aligned} E[X \cdot Y] &= E[(X \cdot Y)^+] - E[(X \cdot Y)^-] \\ &= EX^+ \cdot EY^+ + EX^- \cdot EY^- - EX^+ \cdot EY^- - EX^- \cdot EY^+ \\ &= (EX^+ - EX^-) \cdot (EY^+ - EY^-) = EX \cdot EY. \end{aligned}$$

□

**Lemma 5.18.**  $X_1, \dots, X_n$  seien diskrete Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  und es gelte  $E[X_i^2] < \infty \forall i = 1, \dots, n$ .

Dann gelten

(i)  $\text{var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{var}(X_i) + \sum_{(i,j): i \neq j} \text{cov}(X_i, X_j).$

(ii) Falls zusätzlich  $X_1, \dots, X_n$  stochastisch unabhängig sind, so folgen

$$\text{cov}(X_i, X_j) = 0, \quad \text{falls } i \neq j,$$

sowie

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

(iii)  $(E[X_1 \cdot X_2])^2 \leq E[X_1^2] \cdot E[X_2^2]$  (Ungleichung von Cauchy-Schwarz)

*Beweis.* (i) Wegen  $E[X_i^2] < \infty$  folgt  $E|X_i| < \infty \forall i = 1, \dots, n$  und somit ist auch  $E[X_1 + \dots + X_n]$  endlich. Damit ist die Varianz von  $X_1 + \dots + X_n$  wohldefiniert. Es gilt

$$\begin{aligned} \text{var}(X_1 + \dots + X_n) &= E\left[\left(X_1 + \dots + X_n - E[X_1 + \dots + X_n]\right)^2\right] \\ &= E\left[\left((X_1 - EX_1) + \dots + (X_n - EX_n)\right)^2\right] \\ &= \sum_{i,j=1}^n E\left[(X_i - EX_i)(X_j - EX_j)\right] \\ &= \sum_{i=1}^n \text{var}(X_i) + \sum_{(i,j): i \neq j} \text{cov}(X_i, X_j). \end{aligned}$$

(ii) Falls  $X_i$  und  $X_j$  stochastisch unabhängig sind, so sind nach Lemma 5.16 auch  $(X_i - EX_i)$  und  $(X_j - EX_j)$  stochastisch unabhängig. Nach Lemma 5.17 folgt nun für  $i \neq j$ , dass

$$\begin{aligned} \text{cov}(X_i, X_j) &= E\left[(X_i - EX_i)(X_j - EX_j)\right] \\ &= E[X_i - EX_i] \cdot E[X_j - EX_j] = 0. \end{aligned}$$

Die zweite Aussage in (ii) folgt dann aus (i).

(iii) Es gilt

$$\begin{aligned} 0 &\leq E\left[\left(E[X_1^2] \cdot X_2 - E[X_1 \cdot X_2] \cdot X_1\right)^2\right] \\ &= \dots = \left(E[X_1^2]\right)^2 \cdot E[X_2^2] - \left(E[X_1 \cdot X_2]\right)^2 \cdot E[X_1^2] \\ &= E[X_1^2] \cdot \left\{E[X_1^2] \cdot E[X_2^2] - \left(E[X_1 X_2]\right)^2\right\}. \end{aligned}$$

Nun unterscheiden wir zwischen zwei Fällen:

1. *Fall:*  $E[X_1^2] = 0$

Dann folgt jedoch  $P\{\omega: X_1(\omega) \neq 0\} = 0$  und es gilt

$$E[X_1 X_2] = 0 = \sqrt{E[X_1^2]} \cdot \sqrt{E[X_2^2]}.$$

2. *Fall:*  $E[X_1^2] > 0$

Dann ist jedoch die Differenz in geschweiften Klammern nichtnegativ.

In beiden Fällen folgt nun die behauptete Ungleichung. □

Zum Abschluss dieses Kapitels werden noch zwei nützliche Schranken für sogenannte Überschreitungswahrscheinlichkeiten hergeleitet.

**Satz 5.19.** *X sei eine diskrete Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ . Dann gelten*

$$(i) \quad P\left(\{\omega: |X(\omega)| \geq \varepsilon\}\right) \leq \frac{E|X|}{\varepsilon} \quad \forall \varepsilon > 0. \quad \text{(Markoff-Ungleichung)}$$

(ii) *Falls zusätzlich  $E[X^2] < \infty$  gilt, so*

$$P\left(\{\omega: |X(\omega) - EX| \geq \varepsilon\}\right) \leq \frac{\text{var}(X)}{\varepsilon^2} \quad \forall \varepsilon > 0. \quad \text{(Tschebyscheff-Ungleichung)}$$

*Beweis.* (i) Es sei  $\{x_1, \dots, x_N\}$  bzw.  $\{x_1, x_2, \dots\}$  die Menge der möglichen Werte von  $X$ . Dann gilt

$$\begin{aligned} P\left(\{\omega: |X(\omega)| \geq \varepsilon\}\right) &= P\left(\bigcup_{i: |x_i| \geq \varepsilon} \{\omega: X(\omega) = x_i\}\right) \\ &= \sum_{i: |x_i| \geq \varepsilon} P\left(\{\omega: X(\omega) = x_i\}\right) \\ &\leq \sum_{i: |x_i| \geq \varepsilon} \frac{|x_i|}{\varepsilon} P\left(\{\omega: X(\omega) = x_i\}\right) \\ &\leq \frac{1}{\varepsilon} \sum_{i \geq 1} |x_i| P\left(\{\omega: X(\omega) = x_i\}\right) = \frac{E|X|}{\varepsilon}. \end{aligned}$$

(ii) Wir wenden (i) auf die Zufallsvariable  $Y := (X - EX)^2$  an und erhalten

$$\begin{aligned} P\left(\{\omega: |X(\omega) - EX| \geq \varepsilon\}\right) &= P\left(|X - EX|^2 \geq \varepsilon^2\right) \\ &\leq \frac{E[(X - EX)^2]}{\varepsilon^2} = \frac{\text{var}(X)}{\varepsilon^2}. \end{aligned}$$

□

## 6 Gesetze der großen Zahlen

Warum zocken halbwegs gebildete Menschen eher selten an Spielautomaten oder im Casino? Vielleicht spielt hierbei das Gefühl eine Rolle, dass sich über einen langen Zeitraum Glück und Pech irgendwann ausgleichen. Und da der Spielautomatenbetreiber bzw. der Besitzer des Casinos Gewinn machen möchte ist zu erwarten, dass ein jeder Spieler über einen langen Zeitraum Verluste macht. In diesem Kapitel werden wir Gesetze herleiten (und rigoros beweisen) welche diese Heuristik bestätigen.

Betrachten wir das Beispiel des Roulette-Spiels. Der Spieltisch weist 37 Zahlenfelder auf, welche von 0 bis 36 durchnummeriert sind. Ein Spieler kann u.a. auf das Erscheinen einer bestimmten Zahl wetten, indem er einen Jeton (Chip) auf eines dieser Zahlenfelder platziert. Sind die Einsätze getätigt, setzt der Croupier die Roulette-Scheibe in Bewegung und wirft die Kugel entgegen der Drehrichtung in den Zylinder. Jenes Zahlenfeld auf der Roulette-Scheibe, in welches die Kugel schließlich fällt, gibt die gewinnbringende Zahl an. Im Idealfall ist die Wahrscheinlichkeit eines Gewinns pro Spiel gleich  $1/37$ . Die Auszahlungsquote im Gewinnfall wird mit 35:1 angegeben. Da der Spieler im Gewinnfall jedoch zusätzlich seinen Einsatz zurückbekommt, erhält er somit das 36-fache seines Einsatzes. Nehmen wir an, ein Spieler wettet  $n$ -mal auf das Erscheinen der Zahl „7“. Wir bezeichnen mit  $H_n := \text{Anzahl der Gewinne} \cdot (1/n)$  die sogenannte **relative Häufigkeit** der Gewinne. Wenn wir die naheliegenden Annahmen treffen, dass die Gewinnwahrscheinlichkeit pro Spiel jeweils  $1/37$  ist und die Ereignisse, dass in den einzelnen Spielen Gewinne eintreten unabhängig sind, so gilt

$$nH_n \sim \text{Bin}(n, 1/37).$$

Wir werden zeigen, dass

$$H_n \underset{n \rightarrow \infty}{\rightsquigarrow} 1/37$$

gilt, wobei die hinter „ $\underset{n \rightarrow \infty}{\rightsquigarrow}$ “ versteckte Konvergenzart noch präzisiert werden muss. Angenommen, der Spieler setzt nun in jedem seiner Spiele 1 Euro. Wir bezeichnen mit  $X_i$  die Auszahlung, die er nach dem  $i$ -ten Spiel erhält. (Im Gewinnfall sind es 36 Euro, im Verlustfall bekommt er nichts.) Dann ist offensichtlich  $EX_i = 36/37$ . Im Weiteren zeigen wir auch, dass

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \underset{n \rightarrow \infty}{\rightsquigarrow} 36/37$$

gilt. Berücksichtigt man noch, dass jedes Spiel den Einsatz von 1 Euro gekostet hat so wird klar, dass der Spieler tendenziell Geld verliert. Auch hier ist wiederum „ $\underset{n \rightarrow \infty}{\rightsquigarrow}$ “ noch zu präzisieren.

Was genau bedeutet nun „ $\underset{n \rightarrow \infty}{\rightsquigarrow}$ “?

Da z.B.  $nH_n \sim \text{Bin}(n, 1/37)$  ist, so gilt **für alle**  $n \in \mathbb{N}$ , dass

$$P\left(\left\{\omega \in \Omega: \left|H_n(\omega) - \frac{1}{37}\right| \geq \varepsilon\right\}\right) > 0,$$

falls  $\varepsilon \leq \frac{36}{37}$ . Man kann also nicht, wie in der Analysis üblich, einfach sagen, dass für alle  $\varepsilon > 0$  ein (hinreichend großes)  $n_\varepsilon \in \mathbb{N}$  existiert mit

$$\left|H_n(\omega) - \frac{1}{37}\right| \leq \varepsilon \quad \forall n \geq n_\varepsilon.$$

Andererseits folgt mit  $p = 1/37$  aus  $EH_n = p$ ,  $\text{var}(H_n) = p(1-p)/n$  und der Tschebyscheff-Ungleichung, dass

$$P\left(\{\omega: |H_n(\omega) - p| \geq \varepsilon\}\right) \leq \frac{p(1-p)/n}{\varepsilon^2} \quad \forall \varepsilon > 0,$$

Es besteht also eine gewisse Tendenz, dass sich die relative Häufigkeit mit wachsendem  $n$  der Wahrscheinlichkeit  $p$  annähert. Diese Art der Konvergenz wird durch folgende Definition beschrieben.

**Definition 6.1.**  $(X_n)_{n \in \mathbb{N}}$  sei eine Folge von diskreten Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  und  $X$  sei eine weitere diskrete Zufallsvariable. Dann **konvergiert**  $(X_n)_{n \in \mathbb{N}}$   **$P$ -stochastisch** (stochastisch) gegen  $X$ , falls

$$P\left(\{\omega: |X_n(\omega) - X(\omega)| \geq \varepsilon\}\right) \xrightarrow{n \rightarrow \infty} 0 \quad \forall \varepsilon > 0.$$

Im Falle dieser Konvergenz schreiben wir kurz  $X_n \xrightarrow{P} X$ . „ $n \rightarrow \infty$ “ wird hierbei üblicherweise weggelassen um die Notation nicht zu überladen. Falls nicht schon aus dem Zusammenhang hervorgeht, dass dies gemeint ist, so kann man „mit  $n \rightarrow \infty$ “ anfügen.

Diese Definition drückt teilweise das aus, was man sich unter Konvergenz vorstellt. Jede  **feste** Schwelle  $\varepsilon > 0$  wird mit einer Wahrscheinlichkeit überschritten, welche gegen Null konvergiert. Die folgende Aussage zeigt, dass man zusätzlich diese Schwelle gegen Null streben lassen kann, was für eine Anwendung dieses Konzeptes oftmals hilfreich ist.

**Lemma 6.2.**  $(X_n)_{n \in \mathbb{N}}$  und  $X$  seien diskrete Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  und es gelte  $X_n \xrightarrow{P} X$ . Dann existiert eine Nullfolge  $(\varepsilon_n)_{n \in \mathbb{N}}$  (d.h.,  $\varepsilon_n \xrightarrow{n \rightarrow \infty} 0$ ) mit

$$P\left(\{\omega: |X_n(\omega) - X(\omega)| \geq \varepsilon_n\}\right) \leq \varepsilon_n \quad \forall n \in \mathbb{N}.$$

*Beweis.* Nach Definition der stochastischen Konvergenz finden wir für alle  $k \in \mathbb{N}$  ein jeweils hinreichend großes  $n_k \in \mathbb{N}$ , sodass

$$P\left(|X_n - X| \geq \frac{1}{k}\right) \leq \frac{1}{k} \quad \forall n \geq n_k. \quad (6.1)$$

Wir können diese Folge insbesondere so wählen, dass  $n_1 = 1$  ist und dass  $(n_k)_{k \in \mathbb{N}}$  **streng** monoton wachsend ist. Wir wählen nun

$$\varepsilon_n := \frac{1}{k} \quad \text{für } n_k \leq n < n_{k+1}.$$

Mit dieser Wahl gilt offensichtlich

$$P\left(\{\omega: |X_n(\omega) - X(\omega)| \geq \varepsilon_n\}\right) \leq \varepsilon_n \quad \forall n \in \mathbb{N}.$$

Da  $(n_k)_{k \in \mathbb{N}}$  streng monoton wachsend ist, existiert zu jedem  $n \in \mathbb{N}$  ein  $k(n) \in \mathbb{N}$  mit  $n_{k(n)} \leq n < n_{k(n)+1}$ . Da aber  $n_{k(n)} \xrightarrow{n \rightarrow \infty} \infty$ , so folgt

$$\varepsilon_n = \frac{1}{n_{k(n)}} \xrightarrow{n \rightarrow \infty} 0.$$

□

Wir werden nun mit Hilfe der Tschebyscheff-Ungleichung das sogenannte **schwache Gesetz der großen Zahlen** herleiten.

**Satz 6.3.**  $(X_n)_{n \in \mathbb{N}}$  sei eine Folge von diskreten und unkorrelierten (d.h.,  $\text{cov}(X_i, X_j) = 0$  für  $i \neq j$ ) Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  und es gelten

$$EX_i = \mu, \quad \text{var}(X_i) \leq M \quad \forall i \in \mathbb{N},$$

wobei  $M < \infty$ . Dann gilt

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mu.$$

*Beweis.* Es gelten

$$E[\bar{X}_n] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n EX_i = \mu$$

und

$$\text{var}(\bar{X}_n) = \text{var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i) \leq \frac{M}{n}.$$

Daher folgt für beliebiges  $\varepsilon > 0$  aus der Tschebyscheff-Ungleichung, dass

$$P\left(\{\omega: |\bar{X}_n(\omega) - \mu| \geq \varepsilon\}\right) \leq \frac{\text{var}(\bar{X}_n)}{\varepsilon^2} \leq \frac{M}{n \varepsilon^2} \xrightarrow{n \rightarrow \infty} 0.$$

□

Im Weiteren steuern wir jetzt auf das **starke Gesetz der großen Zahlen** zu. Auf dem Weg dahin werden wir eines der sogenannten 0-1-Gesetze beweisen. Im nächsten Satz werden Aussagen darüber getroffen, mit welcher Wahrscheinlichkeit unendlich viele Ereignisse von einer vorgegebenen Folge von Ereignissen eintreten. Wir nehmen an, dass  $(A_n)_{n \in \mathbb{N}}$  solch eine Folge von Ereignissen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  ist. Jenes Ereignis, dass unendlich viele dieser Ereignisse eintreten, kann formal beschrieben werden durch

$$\{\omega: \omega \in A_n \text{ für unendlich viele } n \in \mathbb{N}\} = \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n.$$

Diese formale Darstellung ist vielleicht auf den ersten Blick etwas gewöhnungsbedürftig. Man überzeugt sich am einfachsten von der oben behaupteten Gleichheit der Mengen indem man sich separat die Inklusionen „ $\subseteq$ “ und „ $\supseteq$ “ überlegt:

„ $\subseteq$ “ Falls  $\omega$  in unendlich vielen der Mengen  $A_n$  liegt, so gilt natürlich  $\omega \in \bigcup_{n=k}^{\infty} A_n$  für alle  $k \in \mathbb{N}$  und somit folgt  $\omega \in \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n$ .

„ $\supseteq$ “ Falls  $\omega \in \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n$ , so gilt  $\omega \in \bigcup_{n=k}^{\infty} A_n$  für alle  $k \in \mathbb{N}$ . Nun gibt es  $n_1 \in \mathbb{N}$  mit  $\omega \in A_{n_1}$ . Da aber insbesondere  $\omega \in \bigcup_{n=n_1+1}^{\infty} A_n$ , so existiert  $n_2 \in \mathbb{N}$ ,  $n_2 > n_1$ , mit  $\omega \in A_{n_2}$  u.s.w.. Daher gibt es eine Folge  $(n_k)_{k \in \mathbb{N}}$  mit  $n_k < n_{k+1}$  und  $\omega \in A_{n_k}$  für alle  $k \in \mathbb{N}$ . Folglich liegt  $\omega$  in unendlich vielen Mengen  $A_n$ .

Die Menge  $\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n$  liegt übrigens in der  $\sigma$ -Algebra  $\mathcal{A}$ , da die Mengen  $A_1, A_2, \dots$  in  $\mathcal{A}$  liegen und die Mengenoperationen der Vereinigung und Durchschnittsbildung nicht aus  $\mathcal{A}$  heraus führen.

#### Satz 6.4. (Lemma von Borel-Cantelli)

Es sei  $(A_n)_{n \in \mathbb{N}}$  eine Folge von Ereignissen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ . Dann gelten

(i) Falls  $\sum_{n=1}^{\infty} P(A_n) < \infty$ , so folgt

$$P\left(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n\right) = 0.$$

(ii) Falls  $\sum_{n=1}^{\infty} P(A_n) = \infty$  und zusätzlich die Ereignisse  $A_n$  stochastisch unabhängig sind, so folgt

$$P\left(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n\right) = 1.$$

Für Folgen unabhängiger Ereignisse  $(A_n)_{n \in \mathbb{N}}$  ergibt sich somit die sicherlich nicht vorherzusehende Aussage, dass die Wahrscheinlichkeit des Eintretens von unendlich vielen dieser Ereignisse nur 0 oder 1 sein kann. Das erklärt auch die Bezeichnung „0-1-Gesetz“. Darüber hinaus wird mit  $\sum_{n=1}^{\infty} P(A_n) < \infty$  bzw.  $\sum_{n=1}^{\infty} P(A_n) = \infty$  ein einfaches Kriterium für das Vorliegen eines jeden dieser Fälle angegeben. Dies wird später einen überschaubaren Beweis des starken Gesetzes der großen Zahlen ermöglichen.

*Beweis von Satz 6.4.* (i) Es sei  $\sum_{n=1}^{\infty} P(A_n) < \infty$ . Für beliebiges  $K \in \mathbb{N}$  gilt die Abschätzung

$$P\left(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n\right) \leq P\left(\bigcup_{n=K}^{\infty} A_n\right) \leq \sum_{n=K}^{\infty} P(A_n).$$

Da aus  $\sum_{n=1}^{\infty} P(A_n) < \infty$  jedoch  $\sum_{n=K}^{\infty} P(A_n) \xrightarrow{K \rightarrow \infty} 0$  folgt, erhalten wir Aussage (i).

Der Beweis von (i) ist sicher auch heuristisch nachvollziehbar:  $\sum_{n=1}^{\infty} P(A_n) < \infty$  deutet darauf hin, dass die Summe der Wahrscheinlichkeiten der  $A_n$  nicht ausreicht, dass mit positiver Wahrscheinlichkeit unendlich viele dieser Ereignisse eintreten.

(ii) Der Beweis von Aussage (ii) ist etwas komplizierter und ein „roter Faden“ ist vielleicht nicht auf den ersten Blick erkennbar. Wir betrachten die Wahrscheinlichkeit des komplementären Ereignisses und erhalten unter Ausnutzung der Eigenschaften von Wahrscheinlichkeitsmaßen (siehe insbesondere Satz 2.4, Eigenschaften (v) und (vi)), dass

$$\begin{aligned} P\left(\left(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n\right)^c\right) &= P\left(\bigcup_{k=1}^{\infty} \left(\bigcap_{n=k}^{\infty} A_n\right)^c\right) \\ &= P\left(\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n^c\right) \\ &= \lim_{k \rightarrow \infty} P\left(\bigcap_{n=k}^{\infty} A_n^c\right) && \text{(Stetigkeit von unten)} \\ &= \lim_{k \rightarrow \infty} \lim_{N \rightarrow \infty} P\left(\bigcap_{n=k}^N A_n^c\right) && \text{(Stetigkeit von oben)} \\ &= \lim_{k \rightarrow \infty} \lim_{N \rightarrow \infty} \prod_{n=k}^N (1 - P(A_n)) && \text{(Unabhängigkeit)}. \end{aligned}$$

Wegen  $1 - p \leq e^{-p} \forall p \in [0, 1]$  folgt

$$P\left(\left(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n\right)^c\right) = \lim_{k \rightarrow \infty} \lim_{N \rightarrow \infty} \underbrace{\prod_{n=k}^N (1 - P(A_n))}_{\leq e^{-\sum_{n=k}^N P(A_n)}} \leq \lim_{k \rightarrow \infty} \underbrace{\lim_{N \rightarrow \infty} e^{-\sum_{n=k}^N P(A_n)}}_{=0} = 0.$$

□

Wir nehmen an, dass nacheinander unendlich viele Zufallsexperimente durchgeführt werden, wobei die Ereignisse des Eintretens von „Erfolgen“ stochastisch unabhängig sind und die Wahrscheinlichkeit eines Erfolges beim  $n$ -ten Experiment gleich  $p_n$  sei. (Hier kann man an das Ziehen aus einer Folge von Urnen, aber an das Eintreten bestimmter Ereignisse wie jährlich drohende Naturkatastrophen, Epidemien u.s.w. denken.) Es sei  $A_n$  jenes Ereignis, dass beim  $n$ -ten „Versuch“ ein „Erfolg“ eintritt. Falls nun  $p_n = 1/n$

$\forall n \in \mathbb{N}$ , so gilt  $\sum_{n=1}^{\infty} P(A_n) = \infty$  und es folgt nach dem Lemma von Borel-Cantelli, dass die Wahrscheinlichkeit des Eintretens unendlich vieler  $A_n$  gleich 1 ist. Falls dagegen  $p_n = 1/n^2$  für alle  $n$  ist, so folgt  $\sum_{n=1}^{\infty} P(A_n) < \infty$  und die Wahrscheinlichkeit des Eintretens unendlich vieler  $A_n$  ist gleich 0. Bemerkenswert ist, dass es keine Konstellation für die Wahrscheinlichkeiten  $p_n$  gibt, sodass die Wahrscheinlichkeit des Eintretens unendlich vieler der Ereignisse  $A_n$  echt zwischen 0 und 1 liegt.

Im Hinblick auf die Formulierung des **starken Gesetzes der großen Zahlen** werden wir eine weitere Konvergenzart für Folgen diskreter Zufallsvariablen einführen:

**Definition 6.5.**  $(X_n)_{n \in \mathbb{N}}$  sei eine Folge von diskreten Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  und  $X$  sei eine weitere diskrete Zufallsvariable. Dann **konvergiert**  $(X_n)_{n \in \mathbb{N}}$   **$P$ -fast sicher** (fast sicher, mit Wahrscheinlichkeit 1) gegen  $X$ , falls

$$P\left(\left\{\omega \in \Omega: X_n(\omega) \xrightarrow[n \rightarrow \infty]{} X(\omega)\right\}\right) = 1.$$

Im Falle dieser Konvergenz schreiben wir kurz  $X_n \xrightarrow{P-f.s.} X$  oder auch  $X_n \rightarrow X$   $P$ -f.s. Wie schon bei der  $P$ -stochastischen Konvergenz wird der Zusatz  $n \rightarrow \infty$  üblicherweise weggelassen.

An dieser Stelle sei angemerkt, dass die Menge  $\{\omega \in \Omega: X_n(\omega) \xrightarrow[n \rightarrow \infty]{} X(\omega)\}$  tatsächlich in der  $\sigma$ -Algebra  $\mathcal{A}$  liegt. Ausgehend von der Tatsache, dass eine Folge reeller Zahlen  $(x_n)_{n \in \mathbb{N}}$  gegen  $x \in \mathbb{R}$  konvergiert, falls für alle  $\varepsilon > 0$  ein  $n_\varepsilon \in \mathbb{N}$  existiert mit  $|x_n - x| \leq \varepsilon$   $\forall n \geq n_\varepsilon$ , können wir die obige Menge folgendermaßen darstellen:

$$\left\{\omega \in \Omega: X_n(\omega) \xrightarrow[n \rightarrow \infty]{} X(\omega)\right\} = \bigcap_{r=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \left\{\omega: |X_k(\omega) - X(\omega)| \leq \frac{1}{r}\right\}. \quad (6.2)$$

Da mit  $X_k$  und  $X$  auch  $|X_k - X|$  eine diskrete Zufallsvariable ist, so ist jedoch  $\{\omega: |X_k(\omega) - X(\omega)| \leq 1/r\} \in \mathcal{A}$ . Da bekanntermaßen die Mengenoperationen der Vereinigung und Durchschnittsbildung nicht aus  $\mathcal{A}$  herausführen, folgt die Behauptung.

Das folgende Lemma besagt, dass die  $P$ -fast sichere Konvergenz die  $P$ -stochastische Konvergenz zur Folge hat.

**Lemma 6.6.**  $(X_n)_{n \in \mathbb{N}}$  sei eine Folge von diskreten Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  und  $X$  sei eine weitere diskrete Zufallsvariable.

Falls  $X_n \xrightarrow{P-f.s.} X$ , so folgt  $X_n \xrightarrow{P} X$ .

*Beweis.* Es sei  $\varepsilon > 0$  beliebig. Ausgehend von der Darstellung (6.2) erhalten wir

$$\begin{aligned}
 1 &= P\left(\bigcap_{r=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \left\{\omega: |X_k(\omega) - X(\omega)| \leq \frac{1}{r}\right\}\right) \\
 &\leq P\left(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \left\{\omega: |X_k(\omega) - X(\omega)| < \varepsilon\right\}\right) \\
 &= \lim_{n \rightarrow \infty} P\left(\bigcap_{k=n}^{\infty} \left\{\omega: |X_k(\omega) - X(\omega)| < \varepsilon\right\}\right) \quad (\text{wegen Stetigkeit von unten}) \\
 &\leq \liminf_{n \rightarrow \infty} P\left(\left\{\omega: |X_n(\omega) - X(\omega)| < \varepsilon\right\}\right).
 \end{aligned}$$

(Die Ungleichungen sind natürlich in der Tat Gleichungen.) Daraus folgt

$$P\left(\left\{\omega: |X_n(\omega) - X(\omega)| \geq \varepsilon\right\}\right) \xrightarrow[n \rightarrow \infty]{} 0,$$

also  $X_n \xrightarrow{P} X$ . □

**Bemerkung 6.7.** Die Umkehrung der Aussage von Lemma 6.6 gilt im Allgemeinen nicht. Als ein einfaches (wenn auch nicht gerade realistisches) Gegenbeispiel betrachten wir ein Experiment, wobei abzählbar oft „rein zufällig“ jeweils eine Kugel aus einer Urne gezogen wird. Dabei sei die Urne vor dem  $n$ -ten Ziehen mit einer roten und  $n - 1$  schwarzen, ansonsten gleichartigen Kugeln bestückt. Solch ein Experiment kann durch eine Folge  $(Y_n)_{n \in \mathbb{N}}$  von unabhängigen Zufallsvariablen auf einem geeigneten Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  beschrieben werden, wobei

$$P(Y_n = 1) = 1 - P(Y_n = 0) = \frac{1}{n} \quad \forall n \in \mathbb{N}.$$

Nun gilt für beliebiges  $\varepsilon > 0$

$$P(|Y_n| \geq \varepsilon) \leq P(Y_n \neq 0) \xrightarrow[n \rightarrow \infty]{} 0.$$

Also gilt

$$Y_n \xrightarrow{P} 0.$$

Es sei nun  $\varepsilon \in (0, 1)$ . Wegen

$$\sum_{n=1}^{\infty} P(|Y_n| \geq \varepsilon) = \sum_{n=1}^{\infty} P(Y_n = 1) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

folgt jedoch aus dem Lemma von Borel-Cantelli (Satz 6.4, Teil (ii)), dass

$$1 = P(|Y_n| \geq \varepsilon \text{ unendlich oft}) \leq P(Y_n \not\rightarrow 0).$$

Das zeigt, dass  $Y_n \xrightarrow{P-f.s.} 0$  **nicht** gilt.

Wir zeigen nun, dass sich die Aussage von Satz 6.3 verschärfen lässt.

**Satz 6.8.**  $(X_n)_{n \in \mathbb{N}}$  sei eine Folge von diskreten und unkorrelierten (d.h.,  $\text{cov}(X_i, X_j) = 0$  für  $i \neq j$ ) Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  und es gelten

$$EX_i = \mu, \quad \text{var}(X_i) \leq M \quad \forall i \in \mathbb{N},$$

wobei  $M < \infty$ . Dann gilt

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P\text{-f.s.}} \mu.$$

Zum Beweis dieses Satzes benutzen wir folgendes Hilfsmittel:

**Lemma 6.9.**  $(Y_n)_{n \in \mathbb{N}}$  sei eine Folge von diskreten Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ . Falls

$$\sum_{n=1}^{\infty} P(\{\omega : |Y_n(\omega)| \geq \varepsilon\}) < \infty \quad \forall \varepsilon > 0,$$

so folgt

$$Y_n \xrightarrow{P\text{-f.s.}} 0.$$

*Beweis.* Nach Voraussetzung gilt für alle  $k \in \mathbb{N}$ , dass

$$\sum_{n=1}^{\infty} P(\{\omega : |Y_n(\omega)| \geq 1/k\}) < \infty.$$

Daraus folgt nach dem Lemma von Borel-Cantelli, dass

$$P(|Y_n| \geq 1/k \text{ für unendlich viele } n \in \mathbb{N}) = 0 \quad \forall k \in \mathbb{N}.$$

Wegen  $\{\omega : Y_n \not\rightarrow 0\} = \bigcup_{k \in \mathbb{N}} \{\omega : |Y_n(\omega)| \geq 1/k \text{ für unendlich viele } n \in \mathbb{N}\}$  folgt

$$\begin{aligned} P(Y_n \not\rightarrow 0) &= P\left(\bigcup_{k \in \mathbb{N}} \{\omega : |Y_n(\omega)| \geq 1/k \text{ für unendlich viele } n \in \mathbb{N}\}\right) \\ &\leq \sum_{k=1}^{\infty} P(\{\omega : |Y_n(\omega)| \geq 1/k \text{ für unendlich viele } n \in \mathbb{N}\}) = 0. \end{aligned}$$

Somit gilt

$$Y_n \xrightarrow{P\text{-f.s.}} 0.$$

□

*Beweis von Satz 6.8.* Es sei  $S_n := \sum_{i=1}^n (X_i - \mu)$ . Es ist zu zeigen, dass

$$S_n/n \xrightarrow{P-f.s} 0. \quad (6.3)$$

Der Beweis erfolgt in zwei Schritten, deren Notwendigkeit nicht auf den ersten Blick zu sehen ist. Aus diesem Grund versuchen wir zunächst, die gewünschte Aussage auf direktem Wege zu beweisen: Nach Lemma 6.9 genügt es zu zeigen, dass

$$\sum_{n=1}^{\infty} P(|S_n/n| \geq \varepsilon) < \infty \quad \forall \varepsilon > 0$$

erfüllt ist. Nach der Tschebyscheff-Ungleichung folgt

$$P(|S_n/n| \geq \varepsilon) \leq \frac{\text{var}(S_n/n)}{\varepsilon^2} \leq \frac{M}{n \varepsilon^2}.$$

Wenn wir diese Abschätzung nutzen, erhalten wir jedoch

$$\sum_{n=1}^{\infty} P(|S_n/n| \geq \varepsilon) \leq \frac{M}{\varepsilon} \sum_{n=1}^{\infty} \frac{1}{n} = \infty,$$

woraus wir **nicht** die Gültigkeit der getroffenen Aussage folgern können.

Um die Aussage mit den angedachten Hilfsmitteln zu beweisen, gehen wir in zwei Schritten vor. Zunächst betrachten wir eine **Teilfolge**  $(S_{n^2}/n^2)_{n \in \mathbb{N}}$ . Nun gilt für beliebiges  $\varepsilon > 0$

$$\sum_{n=1}^{\infty} P(|S_{n^2}/n^2| \geq \varepsilon) \leq \frac{M}{\varepsilon} \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty,$$

woraus nach Lemma 6.9

$$S_{n^2}/n^2 \xrightarrow{P-f.s} 0 \quad \text{mit } n \rightarrow \infty \quad (6.4)$$

folgt, d.h., die behauptete fast sichere Konvergenz gilt zumindest für eine Teilfolge.

In einem zweiten Schritt füllen wir noch die „Lücken“ zwischen den Teilfolgengliedern. Um nicht wieder wie beim obigen Versuch zu scheitern, vergleichen wir jetzt  $S_m$  mit  $S_{n_m^2}$ , wobei  $n_m \in \mathbb{N}$  so ist, dass  $n_m^2 \leq m < (n_m + 1)^2$ . Wiederum nach der Tschebyscheff-Ungleichung gilt

$$P(|(S_m - S_{n_m^2})/n_m^2| \geq \varepsilon) \leq \frac{M}{n_m^4 \varepsilon^2} (m - n_m^2) \quad \forall m \in \mathbb{N}.$$

Daraus folgt

$$\begin{aligned} \sum_{m=1}^{\infty} P(|(S_m - S_{n_m^2})/n_m^2| \geq \varepsilon) &= \sum_{n=1}^{\infty} \underbrace{\sum_{m=n^2}^{(n+1)^2-1} P(|(S_m - S_{n^2})/n^2| \geq \varepsilon)}_{\leq \frac{M}{n^4 \varepsilon^2} (0 + 1 + \dots + ((n+1)^2 - 1 - n^2))} \\ &\leq \frac{M}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{2n+1}{n^3} < \infty. \end{aligned}$$

Somit folgt nach Lemma 6.9, dass

$$(S_m - S_{n_m^2})/n_m^2 \xrightarrow{P-f.s.} 0 \quad \text{mit } m \rightarrow \infty. \quad (6.5)$$

Aus (6.4) und (6.5) folgt

$$S_m/n_m^2 = S_{n_m^2}/n_m^2 + (S_m - S_{n_m^2})/n_m^2 \xrightarrow{P-f.s.} 0.$$

Wegen  $|S_m|/m \leq |S_m|/n_m^2$  folgt aber nun (6.3). □

## 7 Wahrscheinlichkeitsmaße mit Dichten

Wir haben bisher ausschließlich diskrete Zufallsvariable  $X$ , welche auf einem geeigneten Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  definiert sind, betrachtet:

- $X$  kann höchstens abzählbar viele Werte  $x_1, \dots, x_N$  bzw.  $x_1, x_2, \dots$  (mit  $x_i \neq x_j$  für  $i \neq j$ ) annehmen.
- Die Beschreibung des Verteilungsgesetzes von  $X$  ist einfach. Es gilt

$$\begin{aligned} P^X(A) &= P(\{\omega \in \Omega: X(\omega) \in A\}) \\ &= \sum_{i: x_i \in A} P(\{\omega \in \Omega: X(\omega) = x_i\}) \quad \forall A \subseteq \mathbb{R}. \end{aligned}$$

Dementsprechend genügt es zur Definition einer Verteilung, die Wahrscheinlichkeiten  $P(X = x_i)$  festzulegen; der Rest wird durch die  $\sigma$ -Additivität „automatisch erledigt“.

- Die Definition des Erwartungswertes ist ebenfalls einfach,

$$EX = \sum_{i \geq 1} x_i P(X = x_i),$$

wobei lediglich darauf zu achten ist, dass  $E[X^+] = \sum_{i: x_i > 0} x_i P(X = x_i)$  oder  $E[X^-] = \sum_{i: x_i < 0} (-x_i) P(X = x_i)$  endlich sind. (Andernfalls würde  $\lim_{N \rightarrow \infty} \sum_{i=1}^N x_i P(X = x_i)$  von der gewählten Summationsreihenfolge abhängen oder sogar überhaupt nicht existieren.)

Wir wollen jetzt Zufallsvariable einführen, deren Wertebereich überabzählbar ist. Dabei werden wir uns auf den Fall beschränken, dass diese Zufallsvariablen eine Riemann-integrierbare Dichte besitzen. Der Wunsch nach solch einer Erweiterung mag auf den ersten Blick übertrieben erscheinen. Zufällige Größen in den verschiedensten Bereichen (fehlerbehaftete Messungen in den Naturwissenschaften, Weiten im Sport, Börsenkurse,...) werden mit endlich vielen Nachkommastellen angegeben. Insofern erscheint eine Modellierung durch diskrete Zufallsvariable sogar natürlicher. Andererseits zeigt sich, dass sich solche Phänomene mit Zufallsvariablen, deren Verteilung durch eine Dichte gegeben ist, wesentlich einfacher beschreiben lassen. Darüber hinaus führen sogenannte Grenzwertsätze der Stochastik (Aus dem Schulunterricht sollte die Approximation der Binomialverteilung durch die Normalverteilung bekannt sein.) zwangsläufig zu solchen Verteilungen.

**Definition 7.1.** *Eine Wahrscheinlichkeitsdichte (Dichte) ist eine nichtnegative und Riemann-integrierbare Funktion  $f: \mathbb{R} \rightarrow [0, \infty)$  mit*

$$\int_{-\infty}^{\infty} f(t) dt = 1.$$

Eine Zufallsvariable  $X$  auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$  besitzt eine Dichte  $f$ , falls

$$P(\{\omega: X(\omega) \leq x\}) = \int_{-\infty}^x f(t) dt \quad \forall x \in \mathbb{R}$$

gilt.

Diese Definition ist korrekt, da nach Definition einer Zufallsvariable  $X$  das Ereignis  $\{\omega: X(\omega) \leq x\}$  für beliebiges  $x \in \mathbb{R}$  in der  $\sigma$ -Algebra  $\mathcal{A}$  liegt und somit  $P(\{\omega: X(\omega) \leq x\})$  wohldefiniert ist. Für beliebige  $a, b \in \mathbb{R}$  mit  $a < b$  gilt dann

$$\begin{aligned} P(\{\omega: X(\omega) \in (a, b]\}) &= P(\{\omega: X(\omega) \in (-\infty, b]\}) - P(\{\omega: X(\omega) \in (-\infty, a]\}) \\ &= \int_{-\infty}^b f(t) dt - \int_{-\infty}^a f(t) dt = \int_a^b f(t) dt. \end{aligned}$$

Wie sieht nun die Verteilung von  $X$  unter  $P$  aus? (Für diskrete Zufallsvariable wurde der Begriff der Verteilung in Definition 5.3 auf Seite 34 dieses Skriptes definiert.) Analog zu dieser Definition lässt sich für beliebige halboffene Intervalle  $(a, b]$  definieren:

$$P^X((a, b]) := P(X^{-1}((a, b])) = P(\{\omega: X(\omega) \in (a, b]\}) = \int_a^b f(t) dt. \quad (7.1)$$

Diese Definition ist für Intervalle sowohl naheliegend als auch einfach. Problematisch erscheint allerdings die Tatsache, dass Verteilungen Wahrscheinlichkeitsmaße sind, welche stets auf geeigneten  $\sigma$ -Algebren definiert sein müssen. Da die Festlegung von  $P^X$  auf der Menge der halboffenen Intervalle problemlos möglich ist, ist es verlockend, diese  $\sigma$ -Algebra so zu wählen, dass zumindest alle halboffenen Intervalle darin enthalten sind. Eine solche  $\sigma$ -Algebra ist allerdings äußerst reichhaltig und enthält Mengen, welche im Rahmen der Riemann-Integrale nicht als Integrationsbereiche geeignet sind. An dieser Stelle helfen nun tiefliegende Sätze aus der Maßtheorie, welche nicht Gegenstand dieser Vorlesung sind und daher nur zu Informationszwecken erwähnt werden. Dazu sei  $\mathcal{B}$  die kleinste  $\sigma$ -Algebra in  $\mathbb{R}$ , welche alle halboffenen Intervalle enthält. Diese  $\sigma$ -Algebra heißt  $\sigma$ -Algebra der Borel-Mengen. Sätze aus der Maßtheorie besagen, dass es ein eindeutig bestimmtes Wahrscheinlichkeitsmaß  $Q: \mathcal{B} \rightarrow [0, 1]$  gibt mit

$$Q((a, b]) = P^X((a, b]) \quad \forall a, b \in \mathbb{R}, a < b.$$

Daher ist es völlig ausreichend, wenn eine explizite Bestimmung von  $P^X$  mittels (7.1) auf der Menge aller halboffenen Intervalle vorgenommen wird. Die Erweiterung dieser Festlegung zu einer Funktion auf  $\mathcal{B}$  wird uns von der Maßtheorie abgenommen und wir können  $P^X$  auf beliebigen Mengen  $B$  aus  $\mathcal{B}$  festlegen durch

$$P^X(B) := Q(B).$$

Eine möglicherweise gefühlte Ungewissheit sollte dabei nicht irritieren. Beim praktischen Rechnen hat man es in der Regel mit Ereignissen (also Mengen aus  $\mathcal{B}$ ) zu tun, deren Wahrscheinlichkeiten man unter Ausnutzung der üblichen Rechenregeln bestimmen oder geeignet abschätzen kann.

Da wir im Weiteren nur den bereits im Schulunterricht vermittelten Begriff des Riemann-Integrals verwenden werden, werden wir anstelle der oben geforderten Riemann-Integrierbarkeit einer Wahrscheinlichkeitsdichte  $f$  eine etwas stärkere Forderung aufstellen. Wir werden fordern, dass eine solche Funktion  $f: \mathbb{R} \rightarrow [0, \infty)$  **höchstens endlich viele Unstetigkeitsstellen** besitzen möge. Dies sichert natürlich deren Riemann-Integrierbarkeit, bietet aber darüber hinaus die Gewähr, dass die im Weiteren folgende Definition des Erwartungswertes entsprechender Zufallsvariablen keinerlei Probleme bereitet.

An dieser Stelle geben wir nun einige Beispiele für Wahrscheinlichkeitsverteilungen  $P$  auf  $\mathcal{B}$  an, welche durch eine Wahrscheinlichkeitsdichte bestimmt werden. Eine Zufallsvariable besitzt eine ...

- 1) **Gleichverteilung über  $[a, b]$**   
... falls sie eine Dichte  $f$  mit

$$f(t) = \begin{cases} \frac{1}{b-a}, & \text{falls } t \in [a, b], \\ 0, & \text{falls } t \notin [a, b] \end{cases}$$

besitzt, wobei  $-\infty < a < b < \infty$ . Die Gleichverteilung wird gelegentlich auch Rechteckverteilung genannt. Man schreibt symbolisch

$$X \sim U[a, b].$$

- 2) **Exponentialverteilung mit Parameter  $\lambda$**   
... falls sie eine Dichte  $f$  mit

$$f(t) = \begin{cases} 0, & \text{falls } t < 0, \\ \lambda e^{-\lambda t}, & \text{falls } t \geq 0 \end{cases}$$

besitzt, wobei  $\lambda > 0$ . Symbolisch:

$$X \sim \text{Exp}(\lambda).$$

- 3) **Normalverteilung mit Parametern  $\mu$  und  $\sigma^2$**   
... falls sie eine Dichte  $f$  mit

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \quad \forall t \in \mathbb{R}$$

besitzt. Die Bedeutung der Parameter  $\mu$  und  $\sigma^2$  wird später klar werden, nachdem der Begriff des Erwartungswertes für Zufallsvariable mit Dichten definiert wurde.  $\mu$  ist der Erwartungswert und  $\sigma^2$  die Varianz einer entsprechenden Zufallsvariable. Symbolisch:

$$X \sim N(\mu, \sigma^2).$$

Falls  $X$  eine Zufallsvariable auf  $(\Omega, \mathcal{A}, P)$  mit einer Dichte  $f$  ist, so folgt für beliebiges  $x \in \mathbb{R}$

$$\begin{aligned} P(\{\omega: X(\omega) = x\}) &= P^X(\{x\}) \\ &= P^X\left(\bigcap_{n=1}^{\infty} (x - 1/n, x]\right) \\ &= \lim_{n \rightarrow \infty} P^X((x - 1/n, x]) \\ &= \lim_{n \rightarrow \infty} \int_{x-1/n}^x f(t) dt = 0. \end{aligned}$$

(Selbst wenn  $f$  im Punkt  $x$  eine Singularität besitzt, so gilt nach der Definition uneigentlicher Riemann-Integrale, dass  $\int_{x-1/n}^x f(t) dt = \lim_{n \rightarrow \infty} \int_{x-1/n}^x f(t) dt$ , woraus wegen  $\int_{x-1/n}^x f(t) dt = \int_{x-1}^x f(t) dt - \int_{x-1}^{x-1/n} f(t) dt$  die letzte Gleichheit folgt.)

Die Verteilung einer Zufallsvariable mit einer Dichte kann also keine „Masse“ in einem einzelnen Punkt  $x \in \mathbb{R}$  besitzen. Damit wird insbesondere klar, dass diskrete Zufallsvariable keine Dichte (in unserem Sinne) besitzen. Eine Brücke zwischen diskreten Zufallsvariablen und solchen mit Dichten wird durch den folgenden Begriff der Verteilungsfunktion geschlagen.

**Definition 7.2.**  $X$  sei eine beliebige Zufallsvariable auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ . Die Funktion  $F: \mathbb{R} \rightarrow [0, 1]$  mit

$$F(x) = P(\{\omega: X(\omega) \leq x\}) \quad \forall x \in \mathbb{R}$$

heißt (**kumulative**) **Verteilungsfunktion** von  $X$  bzw.  $P^X$ .

Gemäß dieser Definition gilt

$$P(\{\omega: X(\omega) \in (a, b]\}) = F(b) - F(a) \quad \forall a, b \in \mathbb{R}, a < b.$$

Falls  $X$  eine diskrete Zufallsvariable mit möglichen Werten  $x_1, x_2, \dots$  ( $x_i \neq x_j$  für  $i \neq j$ ) ist, so gilt

$$F(x) = \sum_{i: x_i \leq x} P(X = x_i).$$

Wir fassen einige Eigenschaften von Verteilungsfunktionen in der folgenden Aussage zusammen.

**Lemma 7.3.**  $X$  sei eine Zufallsvariable auf  $(\Omega, \mathcal{A}, P)$  mit einer Verteilungsfunktion  $F$ . Dann gelten

- (i)  $F$  ist monoton nichtfallend,
- (ii)  $F$  ist rechtsseitig stetig, d.h., für alle  $x \in \mathbb{R}$  und eine beliebige Folge  $(x_n)_{n \in \mathbb{N}}$  mit  $x_n > x \forall n \in \mathbb{N}$  und  $x_n \xrightarrow{n \rightarrow \infty} x$  gilt

$$F(x_n) \xrightarrow{n \rightarrow \infty} F(x),$$

- (iii)  $\lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow \infty} F(x) = 1.$

*Beweis.* (i) Es sei  $x < y$ . Wegen der Isotonie von  $P$  folgt

$$F(x) = P(\{\omega: X(\omega) \leq x\}) \leq P(\{\omega: X(\omega) \leq y\}) = F(y).$$

(ii) Es sei  $\epsilon > 0$  beliebig. Es ist zu zeigen, dass ein  $N(\epsilon) \in \mathbb{N}$  existiert mit  $|F(x) - F(x_n)| \leq \epsilon \forall n \geq N(\epsilon)$ . Da  $P$  die Eigenschaft der Stetigkeit von oben besitzt, erhalten wir

$$F(x+1/k) = P(\{\omega: X(\omega) \leq x+1/k\}) \xrightarrow{k \rightarrow \infty} P\left(\underbrace{\bigcap_{k=1}^{\infty} \{\omega: X(\omega) \leq x+1/k\}}_{\{\omega: X(\omega) \leq x\}}\right) = F(x).$$

Daher existiert ein  $K(\epsilon) \in \mathbb{N}$ , sodass  $F(x + 1/K(\epsilon)) \leq F(x) + \epsilon$ . Da aber  $x < x_n \leq x + 1/K(\epsilon) \forall n \geq N(\epsilon)$  für ein hinreichend großes  $N(\epsilon) \in \mathbb{N}$  gilt, so folgt

$$F(x) \leq F(x_n) \leq F(x + 1/K(\epsilon)) \leq F(x) + \epsilon \quad \forall n \geq N(\epsilon).$$

(iii) Folgt aus Stetigkeit von unten/oben. □

Lemma 7.3 fasst Eigenschaften zusammen, die jede Verteilungsfunktion auf  $\mathbb{R}$  erfüllt. Der nun folgende Satz beinhaltet die bemerkenswerte Aussage, dass jede Funktion, welche Eigenschaften (i)-(iii) aus Lemma 7.3 erfüllt, die Verteilungsfunktion eines Wahrscheinlichkeitsmaßes  $P$  ist.

**Satz 7.4.** *Falls eine Funktion  $F: \mathbb{R} \rightarrow [0, 1]$  die Eigenschaften (i)-(iii) aus Lemma 7.3 erfüllt, so existiert ein eindeutig bestimmtes Wahrscheinlichkeitsmaß  $P$  auf  $\mathcal{B}$  mit*

$$P((-\infty, x]) = F(x) \quad \forall x \in \mathbb{R}.$$

Der Beweis dieses Satzes erfordert Hilfsmittel aus der Maßtheorie und wird an dieser Stelle nicht gebracht. Als eine unmittelbare Folgerung erhalten wir die Aussage, dass jede Funktion, welche die von uns verlangten Eigenschaften einer Dichte erfüllt, tatsächlich die Dichte eines Wahrscheinlichkeitsmaßes ist.

**Folgerung 7.5.**  *$f: \mathbb{R} \rightarrow [0, \infty)$  möge höchstens endlich viele Unstetigkeitsstellen besitzen und es gelte  $\int_{-\infty}^{\infty} f(t) dt = 1$ . Dann existiert ein eindeutig bestimmtes Wahrscheinlichkeitsmaß  $P$  auf  $\mathcal{B}$  mit*

$$P((a, b]) = \int_a^b f(t) dt \quad \forall a, b \in \mathbb{R}, a < b.$$

*Beweis.* Es sei  $f$  eine Funktion, welche die Bedingungen der Folgerung erfüllt. Falls  $P$  das zugehörige Wahrscheinlichkeitsmaß ist, so muss gelten

$$P((-\infty, x]) = \int_{-\infty}^x f(t) dt =: F(x) \quad \forall x \in \mathbb{R}.$$

Die Funktion  $F$  erfüllt die Eigenschaften (i)-(iii) aus Satz 7.4. ( $F$  ist sogar stetig statt nur rechtsseitig stetig.) Nach Satz 7.4 folgt die Behauptung.  $\square$

In Kapitel 5 hatten wir die stochastische Unabhängigkeit diskreter Zufallsvariablen definiert (Definition 5.13). An dieser Stelle wollen wir diese Definition auf allgemeine Zufallsvariable erweitern, was insbesondere den Fall von Zufallsvariablen mit einer Dichte einschließt.

**Definition 7.6.**  $(X_i)_{i \in I}$  sei eine Familie von beliebigen Zufallsvariablen auf einem Wahrscheinlichkeitsraum  $(\Omega, \mathcal{A}, P)$ . Diese Zufallsvariablen heißen **stochastisch unabhängig**, falls für alle endlichen und nichtleeren Teilmengen  $I_0$  von  $I$  die folgende Eigenschaft gilt:

$$P\left(\bigcap_{i \in I_0} \{\omega \in \Omega: X_i(\omega) \leq x_i\}\right) = \prod_{i \in I_0} P\left(\{\omega \in \Omega: X_i(\omega) \leq x_i\}\right) \quad \forall x_i \in \mathbb{R} \quad (i \in I_0).$$

Für den Spezialfall diskreter Zufallsvariable ist die Forderung in dieser Definition schwächer als jene in Definition 5.13. Es lässt sich jedoch leicht zeigen, dass in diesem Fall beide Definitionen äquivalent sind. Wir nehmen an, dass  $(X_i)_{i \in I}$  eine Familie von diskreten Zufallsvariablen ist, welche nach der eben gegebenen Definition stochastisch unabhängig sind. Es sei  $I_0$  eine beliebige endlich und nichtleere Teilmenge von  $I$ . Weiter seien  $x_{i_1}, x_{i_2}, \dots$  die möglichen Werte von  $X_i$ . Dann gilt

$$\begin{aligned} & P\left(\bigcap_{i \in I_0} \{\omega: X_i \in (x_{ij_i} - 1/n, x_{ij_i}]\}\right) \\ &= P\left(\bigcap_{i \in I_0} \{\omega: X_i \in (-\infty, x_{ij_i}]\} \setminus \{\omega: X_i \in (-\infty, x_{ij_i} - 1/n]\}\right) \\ &= \dots = \prod_{i \in I_0} \left(P(\{\omega: X_i \in (-\infty, x_{ij_i}]\}) - P(\{\omega: X_i \in (-\infty, x_{ij_i} - 1/n]\})\right) \\ &= \prod_{i \in I_0} P(\{\omega: X_i \in (x_{ij_i} - 1/n, x_{ij_i}]\}). \end{aligned}$$

Mit  $n \rightarrow \infty$  folgt wegen Stetigkeit von oben, dass

$$P\left(\bigcap_{i \in I_0} \{\omega: X_i = x_{ij_i}\}\right) = \prod_{i \in I_0} P(\{\omega: X_i = x_{ij_i}\}),$$

woraus gemäß Bemerkung 5.14 auch die stochastische Unabhängigkeit entsprechend der im aktuellen Kapitel gegebenen Definition folgt.

An dieser Stelle sei daran erinnert, dass nach Satz 5.15 für stochastisch unabhängige ganzzahlige Zufallsvariable  $X_1$  und  $X_2$  die folgende Rechenregel gilt:

$$P(X_1 + X_2 = k) = \sum_{l=-\infty}^{\infty} P(X_1 = l)P(X_2 = k - l) \quad \forall k \in \mathbb{Z}.$$

Mit Argumenten aus der Maßtheorie, auf die an dieser Stelle wiederum nicht näher eingegangen werden kann, lässt sich nun eine analoge Rechenregel für unabhängige Zufallsvariablen mit Dichten herleiten.

**Satz 7.7.**  *$X$  und  $Y$  seien stochastisch unabhängige Zufallsvariable auf  $(\Omega, \mathcal{A}, P)$  mit jeweiligen Dichten  $f_X$  bzw.  $f_Y$ , welche höchstens endlich viele Unstetigkeitsstellen besitzen. Dann besitzt  $X + Y$  eine stetige Dichte  $f_{X+Y}$  mit*

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(s)f_Y(t - s) ds \quad \forall t \in \mathbb{R}.$$

**Beispiel:**

$X_1 \sim N(\mu_1, \sigma_1^2)$  und  $X_2 \sim N(\mu_2, \sigma_2^2)$  seien stochastisch unabhängig. Dann gilt

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

*Beweis.* Es sei zunächst  $\mu_1 = \mu_2 = 0$ . Dann folgt nach Satz 7.7

$$\begin{aligned} f_{X_1+X_2}(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{s^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(t-s)^2}{2\sigma_2^2}} ds \\ &\propto e^{-\frac{t^2}{2(\sigma_1^2+\sigma_2^2)}} \underbrace{\int_{-\infty}^{\infty} e^{\frac{t^2}{2(\sigma_1^2+\sigma_2^2)}} e^{-\frac{s^2}{2\sigma_1^2}} e^{-\frac{(t-s)^2}{2\sigma_2^2}} ds}_{=: I(t)}. \end{aligned}$$

Wir betrachten nun die Exponenten im obigen Integral  $I(t)$ :

$$\begin{aligned} &\frac{t^2}{2(\sigma_1^2 + \sigma_2^2)} - \frac{s^2}{2\sigma_1^2} - \frac{(t-s)^2}{2\sigma_2^2} \\ &= -t^2 \frac{\sigma_1^2}{2(\sigma_1^2 + \sigma_2^2)\sigma_2^2} \\ &\quad - s^2 \frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1^2\sigma_2^2} \\ &\quad + ts \frac{1}{\sigma_2^2} \\ &= -\frac{1}{2\sigma_2^2} \left( t^2 \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} + s^2 \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2} - 2ts \right) \\ &= -\frac{1}{2\sigma_2^2} \left( s \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2}} - t \sqrt{\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}} \right)^2. \end{aligned}$$

Daher gilt

$$\begin{aligned} I(t) &= \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2\sigma_2^2} \left( s \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2}} - t \sqrt{\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}} \right)^2 \right\} ds \\ &= \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2\sigma_2^2} z^2 \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2} \right\} dz \propto 1, \end{aligned}$$

d.h.,  $I(t)$  hängt nicht von  $t$  ab. Somit gilt

$$f_{X_1+X_2}(t) \propto e^{-\frac{t^2}{2(\sigma_1^2 + \sigma_2^2)}} \quad \forall t \in \mathbb{R}.$$

Wegen  $\int_{-\infty}^{\infty} f_{X_1+X_2}(t) dt = 1$  können wir den noch fehlenden Faktor bestimmen und erhalten

$$f_{X_1+X_2}(t) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} e^{-\frac{t^2}{2(\sigma_1^2 + \sigma_2^2)}} \quad \forall t \in \mathbb{R}.$$

Falls nun  $\mu_1$  und  $\mu_2$  beliebig sind, so erhalten wir aus den obigen Rechnungen

$$\begin{aligned} f_{X_1+X_2}(t) &\propto \int_{-\infty}^{\infty} e^{-\frac{(s-\mu_1)^2}{2\sigma_1^2}} e^{-\frac{((t-s)-\mu_2)^2}{2\sigma_2^2}} ds \\ &= \int_{-\infty}^{\infty} e^{-\frac{s^2}{2\sigma_1^2}} e^{-\frac{((t-\mu_1-\mu_2)-s)^2}{2\sigma_2^2}} ds \\ &\propto e^{-\frac{(t-\mu_1-\mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}}. \end{aligned}$$

Der noch fehlende Proportionalitätsfaktor kann wie oben bestimmt werden und es folgt die Behauptung.  $\square$

Die folgende Aussage beschreibt noch einmal den Zusammenhang zwischen Verteilungsfunktionen und jeweiligen zugehörigen Dichten.

**Lemma 7.8.**  $F: \mathbb{R} \rightarrow [0, 1]$  sei eine **stetige** Verteilungsfunktion, welche bis auf endlich viele Punkte stetig differenzierbar ist. Dann ist  $f$  mit

$$f(t) = \begin{cases} F'(t), & \text{falls } F \text{ in } t \text{ differenzierbar ist,} \\ 0 & \text{sonst} \end{cases}$$

eine Wahrscheinlichkeitsdichte zu  $F$ .

*Beweis.* Es seien  $a, b \in \mathbb{R}$  mit  $a < b$  beliebig. Falls  $F$  im Intervall  $(a, b)$  überall differenzierbar ist, so gilt (da  $F$  in den Punkten  $a$  und  $b$  zumindest stetig ist), dass

$$\begin{aligned} F(b) - F(a) &= \lim_{\delta \rightarrow 0} \{F(b - \delta) - F(a + \delta)\} \\ &= \lim_{\delta \rightarrow 0} \int_{a+\delta}^{b-\delta} \underbrace{F'(t)}_{=f(t)} dt = \int_a^b f(t) dt. \end{aligned}$$

Falls nun  $F$  im Intervall  $(a, b)$  bis auf die Punkte  $x_1, \dots, x_M$  differenzierbar ist, so folgt (mit  $x_0 := a, x_{M+1} := b$ ), dass

$$\begin{aligned} F(b) - F(a) &= \sum_{i=1}^{M+1} F(x_i) - F(x_{i-1}) \\ &= \sum_{i=1}^{M+1} \int_{x_{i-1}}^{x_i} f(t) dt = \int_a^b f(t) dt. \end{aligned}$$

□

**Bemerkung 7.9.**  $F$  sei eine Verteilungsfunktion und  $P$  sei das zugehörige Wahrscheinlichkeitsmaß. Dann gilt

$$P(\{x\}) = \lim_{n \rightarrow \infty} P\left(\left(x - 1/n, x\right]\right) = F(x) - \lim_{n \rightarrow \infty} F(x - 1/n)$$

- (i) Falls  $F$  im Punkt  $x$  stetig ist, so folgt  $P(\{x\}) = 0$ .
- (ii) Falls  $F$  im Punkt  $x$  unstetig ist, so ist der linksseitige Grenzwert  $F(x - 0)$  kleiner als  $F(x)$  und es folgt  $P(\{x\}) > 0$ .

Der folgende Satz erlaubt, die Dichte von abgeleiteten Zufallsvariablen von der Dichte einer zugrundeliegenden Zufallsvariable abzuleiten.

**Satz 7.10.** Die Zufallsvariable  $X$  auf  $(\Omega, \mathcal{A}, P)$  möge eine Dichte  $p_X$  mit höchstens endlich vielen Unstetigkeitsstellen besitzen.  $\varphi: \mathbb{R} \rightarrow G := (c, d)$  sei eine bijektive und stetig differenzierbare Abbildung  $(-\infty \leq c < d \leq \infty)$  und es gelte  $\varphi'(u) \neq 0$  bis auf endlich viele Punkte  $u$  in  $G$ .

Dann besitzt die Zufallsvariable  $Y = \varphi(X)$  eine Dichte  $p_Y$  mit höchstens endlich vielen Unstetigkeitsstellen, wobei

$$p_Y(t) = p_X(\varphi^{-1}(t)) \frac{1}{|\varphi'(\varphi^{-1}(t))|} \mathbb{1}_G(t)$$

bis auf endlich viele Punkte  $t$  aus  $\mathbb{R}$  gilt.

(Die Gleichheit gilt, falls  $t \notin G$  oder falls  $t \in G$  und  $\varphi'(\varphi^{-1}(t)) \neq 0$ .)

*Beweis.*  $F_X$  und  $F_Y$  seien die jeweiligen Verteilungsfunktionen von  $X$  bzw.  $Y$ . Da  $\varphi$  bijektiv und stetig ist, so ist  $\varphi$  entweder streng monoton wachsend oder streng monoton fallend. Wir unterscheiden hier und weiter unten zwischen diesen beiden Fällen.

- a)  $\varphi$  ist streng wachsend  
Dann folgt

$$F_Y(t) = P(\varphi(X) \leq t) = P(X \leq \varphi^{-1}(t)) = F_X(\varphi^{-1}(t)).$$

- b)  $\varphi$  ist streng fallend  
Dann gilt

$$F_Y(t) = P(\varphi(X) \leq t) = P(X \geq \varphi^{-1}(t)) = P(X > \varphi^{-1}(t)) = 1 - F_X(\varphi^{-1}(t)).$$

(Die vorletzte Gleichheit folgt nach Bemerkung 7.9(i).)

Da  $F_X$  eine Dichte  $p_X$  besitzt, so ist  $F_X$  stetig. Da die Abbildung  $t \mapsto \varphi^{-1}(t)$  ebenfalls stetig ist, folgt aus den obigen Formeln die Stetigkeit von  $F_Y$ .

Wir zeigen nun die Differenzierbarkeit von  $F_Y$  und berechnen die Ableitung.

- 1)  $t \in (-\infty, c)$   
Wegen  $F_Y(t) = P(Y \leq t) = 0 \forall t < c$  folgt  $\frac{d}{dt}F_Y(t) = 0 \forall t < c$ .
- 2)  $t \in (d, \infty)$   
Wegen  $F_Y(t) = P(Y \leq t) = 1 \forall t > d$  folgt  $\frac{d}{dt}F_Y(t) = 0 \forall t > d$ .
- 3)  $t \in (c, d)$   
Hier müssen wir nochmals zwischen zwei Fällen unterscheiden:
  - a)  $\varphi$  streng wachsend

$$\begin{aligned} \frac{d}{dt}F_Y(t) &= \frac{d}{dt}F_X(\varphi^{-1}(t)) = \frac{d}{dt} \int_{-\infty}^{\varphi^{-1}(t)} p_X(u) du \\ &= p_X(\varphi^{-1}(t)) \frac{1}{\varphi'(\varphi^{-1}(t))} \\ &= p_X(\varphi^{-1}(t)) \frac{1}{|\varphi'(\varphi^{-1}(t))|}. \end{aligned}$$

Hierbei gilt die vorletzte Gleichung, falls  $p_X$  im Punkte  $\varphi^{-1}(t)$  stetig ist und  $\varphi'(\varphi^{-1}(t)) \neq 0$  ist. Das ist jedoch bis auf endlich viele Punkte  $t$  erfüllt.

- b)  $\varphi$  streng fallend

$$\begin{aligned} \frac{d}{dt}F_Y(t) &= \frac{d}{dt} \left( 1 - F_X(\varphi^{-1}(t)) \right) = \frac{d}{dt} \left( 1 - \int_{-\infty}^{\varphi^{-1}(t)} p_X(u) du \right) \\ &= - p_X(\varphi^{-1}(t)) \frac{1}{\varphi'(\varphi^{-1}(t))} \\ &= p_X(\varphi^{-1}(t)) \frac{1}{|\varphi'(\varphi^{-1}(t))|}. \end{aligned}$$

Hierbei gilt die letzte Gleichung, da  $\varphi$  monoton fallend ist. Wie bei a) gilt die vorletzte Gleichung, falls  $p_X$  im Punkte  $\varphi^{-1}(t)$  stetig ist und  $\varphi'(\varphi^{-1}(t)) \neq 0$  ist. Das ist wiederum bis auf endlich viele Punkte  $t$  erfüllt.

□

Die in Satz 7.10 angegebene Formel kann auf den Fall verallgemeinert werden, dass die Funktion  $\varphi$  lediglich stückweise bijektiv ist. Der folgende Satz liefert diese Verallgemeinerung.

**Satz 7.11.** Die Zufallsvariable  $X$  auf  $(\Omega, \mathcal{A}, P)$  möge eine Dichte  $p_X$  mit höchstens endlich vielen Unstetigkeitsstellen besitzen.  $\varphi: \mathbb{R} \rightarrow \mathbb{R}$  sei eine Abbildung und es mögen disjunkte Intervalle  $(a_i, b_i)$ ,  $i = 1, \dots, k$ , existieren, sodass

$$P^X\left(\bigcup_{i=1}^k (a_i, b_i)\right) = 1$$

gilt und

$$\varphi_i := \varphi|_{(a_i, b_i)}: (a_i, b_i) \rightarrow G_i \quad (i = 1, \dots, k)$$

bijektive Abbildungen sind, wobei  $\varphi_i$  auf  $(a_i, b_i)$  stetig differenzierbar ist und  $\varphi'_i(t) \neq 0$  bis auf endlich viele Punkte aus  $(a_i, b_i)$  gilt.

Dann besitzt die Zufallsvariable  $Y = \varphi(X)$  eine Dichte  $p_Y$  mit höchstens endlich vielen Unstetigkeitsstellen, wobei

$$p_Y(t) = \sum_{i=1}^k p_X(\varphi_i^{-1}(t)) \frac{1}{|\varphi'_i(\varphi_i^{-1}(t))|} \mathbb{1}_{G_i}(t)$$

bis auf endlich viele Punkte  $t$  aus  $\mathbb{R}$  gilt.

(Die Gleichheit gilt, falls  $t \notin \bigcup_{i=1}^k G_i$  oder falls  $\varphi'_i(\varphi_i^{-1}(t)) \neq 0$  für alle  $i$  mit  $t \in G_i$ .)

*Beweis.* Es sei  $F_Y$  die Verteilungsfunktion von  $Y$ . Wir erhalten

$$\begin{aligned} F_Y(t) &= P(\varphi(X) \leq t) \\ &= \sum_{i=1}^k P\left(X \in (a_i, b_i), \varphi_i(X) \leq t\right) \\ &= \sum_{i=1}^k P\left(X \in \varphi_i^{-1}((-\infty, t])\right) \\ &= \sum_{i=1}^k \begin{cases} \int_{-\infty}^{\varphi_i^{-1}(t)} p_X(u) du, & \text{falls } \varphi_i \text{ wachsend,} \\ 1 - \int_{-\infty}^{\varphi_i^{-1}(t)} p_X(u) du, & \text{falls } \varphi_i \text{ fallend,} \end{cases} \end{aligned}$$

wobei die vorletzte Gleichung wegen  $\varphi_i^{-1}((-\infty, t]) \subseteq (a_i, b_i)$  folgt. Da die Umkehrabbildungen  $\varphi_i^{-1}: G_i \rightarrow (a_i, b_i)$  stetig sind folgt, dass die Verteilungsfunktion  $F_Y$  ebenfalls stetig ist.

Wie beim Beweis des vorangegangenen Satzes 7.10 müssen wir nur noch die Ableitung dieser Verteilungsfunktion berechnen und erhalten damit die gesuchte Dichte  $p_Y$ . Falls  $t$  ein innerer Punkt von  $\left(\bigcup_{i=1}^k G_i\right)^c$  ist, so gilt offensichtlich, dass  $\frac{d}{dt} F_Y(t) = 0$ . Falls dagegen

$t \in G_i$  ist, so folgt analog zu den Rechnungen aus dem Beweis von Satz 7.10, dass (falls, o.B.d.A.  $\varphi_i$  wachsend)

$$\frac{d}{dt} \int_{-\infty}^{\varphi_i^{-1}(t)} p_X(u) du = p_X(\varphi_i^{-1}(t)) \frac{1}{|\varphi_i'(\varphi_i^{-1}(t))|},$$

falls  $\varphi_i'(\varphi_i^{-1}(t)) \neq 0$  ist. Dadurch erhält man schließlich, dass die angegebene Formel bis auf endlich viele  $t \in \mathbb{R}$  gilt.  $\square$

Die hergeleiteten Hilfsmittel (Faltungsformel in Satz 7.7, Dichtetransformationsformel in Satz 7.11) können nun benutzt werden um die Dichte einer sogenannten  $\chi^2$ -Verteilung (sprich: chi-Quadrat-Verteilung) herzuleiten. Zunächst definieren wir diese Verteilung auf konstruktive Weise.

**Definition 7.12.**  $X_1, \dots, X_n$  seien stochastisch unabhängig,  $X_i \sim N(0, 1) \forall i = 1, \dots, n$ . Dann heißt die Verteilung von  $\sum_{i=1}^n X_i^2$   $\chi^2$ -Verteilung mit  $n$  Freiheitsgraden. (Kurz:  $\sum_{i=1}^n X_i^2 \sim \chi_n^2$ )

**Lemma 7.13.**  $X_1, \dots, X_n \sim N(0, 1)$  seien stochastisch unabhängig. Dann besitzt  $X_1^2 + \dots + X_n^2$  eine Dichte  $p_{X_1^2 + \dots + X_n^2}$  mit

$$p_{X_1^2 + \dots + X_n^2}(t) = \frac{1}{2^{n/2} \Gamma(n/2)} e^{-t/2} t^{n/2-1} \mathbb{1}_{(0, \infty)}(t),$$

wobei  $\Gamma$  mit  $\Gamma(p) = \int_0^\infty e^{-z} z^{p-1} dz$  die (Eulersche) Gammafunktion ist.

*Beweis.* Wir beweisen die Aussage durch Induktion über  $n$ .

**Induktionsanfang:** ( $n = 1$ )

Es gilt  $p_{X_1}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ . Es seien  $\varphi(x) = x^2$ ,  $\varphi_1 = \varphi|_{(0, \infty)}$  und  $\varphi_2 = \varphi|_{(-\infty, 0)}$ . Diese Funktionen erfüllen die Regularitätsvoraussetzungen von Satz 7.11, sodass die Dichtetransformationsformel angewendet werden kann. Mit  $\varphi_1^{-1}(t) = \sqrt{t}$ ,  $\varphi_2^{-1}(t) = -\sqrt{t}$  und  $\varphi_i'(x) = 2x$  ( $i = 1, 2$ ) erhalten wir

$$\begin{aligned} p_{X_1^2}(t) &= p_{X_1}(\varphi_1^{-1}(t)) \frac{1}{|\varphi_1'(\varphi_1^{-1}(t))|} \mathbb{1}_{(0, \infty)}(t) + p_{X_1}(\varphi_2^{-1}(t)) \frac{1}{|\varphi_2'(\varphi_2^{-1}(t))|} \mathbb{1}_{(0, \infty)}(t) \\ &= \left\{ p_{X_1}(\sqrt{t}) \frac{1}{|2\sqrt{t}|} + p_{X_1}(-\sqrt{t}) \frac{1}{|-2\sqrt{t}|} \right\} \mathbb{1}_{(0, \infty)}(t) \\ &= C_1 e^{-t/2} t^{1/2-1} \mathbb{1}_{(0, \infty)}(t), \end{aligned}$$

wobei wegen  $\int_{-\infty}^\infty p_{X_1^2}(t) dt = 1$

$$\begin{aligned} C_1 &= \left( \int_0^\infty e^{-t/2} t^{1/2-1} dt \right)^{-1} = \left( \int_0^\infty e^{-z} (2z)^{1/2-1} 2dz \right)^{-1} \\ &= \left( \sqrt{2} \int_0^\infty e^{-z} z^{1/2-1} dz \right)^{-1} = \frac{1}{2^{1/2} \Gamma(1/2)} \end{aligned}$$

ist.

**Induktionsschritt** ( $n - 1 \Rightarrow n$ )

Nach der Faltungsformel (Satz 7.7) gilt

$$\begin{aligned}
 p_{X_1^2 + \dots + X_n^2}(t) &= \int_{-\infty}^{\infty} p_{X_1^2 + \dots + X_{n-1}^2}(s) p_{X_n^2}(t-s) ds \\
 &= C_n \int_{-\infty}^{\infty} e^{-s/2} s^{\frac{n-1}{2}-1} \mathbb{1}_{(0,\infty)}(s) e^{-(t-s)/2} (t-s)^{\frac{1}{2}-1} \mathbb{1}_{(0,\infty)}(t-s) ds \\
 &= C_n e^{-t/2} \underbrace{\int_0^t s^{\frac{n-1}{2}-1} (t-s)^{\frac{1}{2}-1} ds}_{= t^{\frac{n}{2}-1} \int_0^1 u^{\frac{n-1}{2}-1} (1-u)^{\frac{1}{2}-1} du} \mathbb{1}_{(0,\infty)}(t) \\
 &= C'_n e^{-t/2} t^{\frac{n}{2}-1} \mathbb{1}_{(0,\infty)}(t),
 \end{aligned}$$

wobei  $C_n$  und  $C'_n$  Konstanten sind, welche nicht von  $t$  abhängen. Wegen  $\int_{-\infty}^{\infty} p_{X_1^2 + \dots + X_n^2}(t) dt = 1$  erhalten wir

$$\begin{aligned}
 C'_n &= \left( \int_0^{\infty} e^{-t/2} t^{\frac{n}{2}-1} dt \right)^{-1} \\
 &= \left( \int_0^{\infty} e^{-z} (2z)^{\frac{n}{2}-1} 2 dz \right)^{-1} = \frac{1}{2^{n/2} \Gamma(n/2)},
 \end{aligned}$$

was den Beweis vollendet.  $\square$

**Der Erwartungswert von Zufallsvariablen mit einer Riemann-Dichte**

Wir erinnern uns, dass im Falle einer diskreten Zufallsvariable  $X$  auf  $(\Omega, \mathcal{A}, P)$  mit möglichen Werten  $x_1, \dots, x_N$  bzw.  $x_1, x_2, \dots$  ( $x_i \neq x_j$  falls  $i \neq j$ ) der Erwartungswert definiert ist durch

$$EX = \sum_{i \geq 1} x_i P(X = x_i),$$

wobei lediglich darauf zu achten ist, dass  $E[X^+] = \sum_{i: x_i > 0} x_i P(X = x_i)$  oder  $E[X^-] = \sum_{i: x_i < 0} (-x_i) P(X = x_i)$  endlich sind. (Andernfalls würde  $\lim_{N \rightarrow \infty} \sum_{i=1}^N x_i P(X = x_i)$  von der gewählten Summationsreihenfolge abhängen oder sogar überhaupt nicht existieren.) Im Fall von Zufallsvariablen mit Dichte gehen wir analog vor.

**Definition 7.14.**  $X$  sei eine Zufallsvariable mit einer Wahrscheinlichkeitsdichte  $f$ , welche höchstens endlich viele Unstetigkeitsstellen besitzt.

(i) Falls  $\int_0^{\infty} x f(x) dx < \infty$  oder  $\int_{-\infty}^0 (-x) f(x) dx < \infty$ , so

$$EX := \int_{-\infty}^{\infty} x f(x) dx = \lim_{N \rightarrow \infty} \int_0^N x f(x) dx + \lim_{N \rightarrow \infty} \int_{-N}^0 x f(x) dx.$$

Andernfalls ist  $EX$  nicht definiert.

- (ii) Falls  $g: \mathbb{R} \rightarrow \mathbb{R}$  höchstens endlich viele Unstetigkeitsstellen besitzt,  $\int_{-\infty}^{\infty} g^+(x)f(x) dx < \infty$  oder  $\int_{-\infty}^{\infty} g^-(x)f(x) dx < \infty$ , so

$$E[g(X)] := \int_{-\infty}^{\infty} g(x)f(x) dx.$$

Falls sowohl  $\int_{-\infty}^{\infty} g^+(x)f(x) dx = \infty$  als auch  $\int_{-\infty}^{\infty} g^-(x)f(x) dx = \infty$  sind, so ist  $E[g(X)]$  nicht definiert.

Falls unter den obigen Bedingungen  $\int_{-\infty}^{\infty} |g(x)|f(x) dx < \infty$  gilt, so folgt wegen  $g^{\pm}(x) \leq |g(x)|$ , dass  $\int_{-\infty}^{\infty} g^{\pm}(x)f(x) dx < \infty$ . Dann existiert der Erwartungswert von  $g(X)$  und ist endlich. Falls  $E[X^2] < \infty$ , so folgt wegen  $|x| \leq x^2 + 1$ , dass  $E[|X|] \leq E[X^2] + 1 < \infty$ . Also ist auch  $EX$  endlich und die Varianz von  $X$  ist definiert durch

$$\text{var}(X) := E[(X - EX)^2] = \int_{-\infty}^{\infty} (x - EX)^2 f(x) dx.$$

### Beispiel: Momente der Normalverteilung

- (i) Falls  $X \sim N(0, 1)$ , so gilt

$$E[X^k] = \begin{cases} (k-1)(k-3)\cdots 3 \cdot 1, & \text{falls } k \text{ gerade,} \\ 0, & \text{falls } k \text{ ungerade.} \end{cases}$$

- (ii) Falls  $X \sim N(\mu, \sigma^2)$ , so gelten

$$EX = \mu, \quad \text{var}(X) = \sigma^2.$$

*Beweis.* (i) Es sei  $\varphi$  mit  $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$  die Dichte der Standardnormalverteilung.

Wegen  $\sup \{|x^k| e^{-x^2/4} : x \in \mathbb{R}\} =: M_k < \infty$  folgt

$$\begin{aligned} E[|X^k|] &= \int_{-\infty}^{\infty} |x^k| \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &\leq \sqrt{2} M_k \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi} 2} e^{-x^2/4} dx}_{=1} < \infty. \end{aligned}$$

Also existiert der Erwartungswert von  $X^k$  und es gilt

$$E[X^k] = \lim_{N \rightarrow \infty} \int_{-N}^N x^k \varphi(x) dx.$$

Falls  $k$  ungerade ist, so ist der Integrand eine ungerade Funktion und es folgt

$$E[X^k] = 0.$$

Falls  $k$  gerade ist, so gilt

$$\begin{aligned} E[X^k] &= \lim_{N \rightarrow \infty} \int_{-N}^N x^k \varphi(x) dx \\ &= \lim_{N \rightarrow \infty} \left\{ - \int_{-N}^N x^{k-1} \underbrace{(-x)\varphi(x)}_{=\varphi'(x)} dx \right\} \\ &= \lim_{N \rightarrow \infty} \left\{ -[x^{k-1}\varphi(x)]_{-N}^N + \int_{-N}^N (k-1)x^{k-2}\varphi(x) dx \right\} \\ &= \underbrace{\lim_{N \rightarrow \infty} \left\{ -[x^{k-1}\varphi(x)]_{-N}^N \right\}}_{=0} + \lim_{N \rightarrow \infty} \int_{-N}^N (k-1)x^{k-2}\varphi(x) dx \\ &= (k-1) \int_{-\infty}^{\infty} x^{k-2}\varphi(x) dx \\ &= \dots = (k-1)(k-3)\dots 3 \cdot 1 \int_{-\infty}^{\infty} \varphi(x) dx. \end{aligned}$$

(ii) Es gelten

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \int_{-\infty}^{\infty} (\sigma u + \mu)\varphi(u) du = \mu \end{aligned}$$

sowie

$$\begin{aligned} \text{var}(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \int_{-\infty}^{\infty} \sigma^2 u^2 \varphi(u) du = \sigma^2. \end{aligned}$$

□

Zum Abschluss dieses Kapitels wollen wir noch ein interessantes Phänomen besprechen, das sogenannte **Benfordsche Gesetz**. Der Astronom und Mathematiker Simon Newcomb publizierte im Jahr 1881 im *American Journal of Mathematics* die Feststellung, dass vordere Seiten von Büchern mit Logarithmentafeln stärker abgegriffen waren als hintere. Das deutet darauf hin, dass Zahlen mit niedrigen Anfangsziffern häufiger vorkommen als solche mit hohen. Nachdem dieses Phänomen in Vergessenheit geraten war, wurde es vom dem Physiker Frank Benford im Jahre 1938 wieder entdeckt und erneut publiziert. Das Phänomen des häufigeren Auftretens niedriger Anfangsziffern lässt sich auch auf vielen anderen Gebieten beobachten: Börsenkurse, Einwohnerzahlen von Städten usw.. Es stellt sich die Frage, ob es ein Zufallsmodell gibt, welches dieses Phänomen erklärt.

Dazu sei  $X$  eine positive Zufallsvariable auf  $(\Omega, \mathcal{A}, P)$ . Da nur die erste signifikante Ziffer von Interesse ist, nicht jedoch die Anzahl der Stellen vor dem Komma, so richten wir unsere Aufmerksamkeit auf Mengen der Bauart

$$M_B := \bigcup_{n=-\infty}^{\infty} 10^n \cdot B,$$

wobei  $B$  eine (geeignete) Teilmenge von  $\mathbb{R}$  ist. Falls nun solch eine Regel für die Häufigkeit des Auftretens der einer ersten signifikanten Ziffer beispielsweise auf Börsenkurse anwendbar wäre, so müsste sie unabhängig von der gewählten Währung (Euro, Dollar,...) gelten. Die **einzige** Forderung, welche wir an die Verteilung der Zufallsvariable  $X$  stellen, ist folglich die **Skaleninvarianz**, d.h.,

$$P(X \in M_B) = P(X \in s \cdot M_B) \quad \forall s > 0. \quad (7.2)$$

Wir betrachten nun spezielle Mengen:

$$\begin{aligned} M_{[1,t)} &= \bigcup_{n=-\infty}^{\infty} [10^n, t \cdot 10^n), \\ s \cdot M_{[1,t)} &= \bigcup_{n=-\infty}^{\infty} [s \cdot 10^n, st \cdot 10^n), \end{aligned}$$

wobei  $1 \leq s \leq st \leq 10$  seien. Nun gelten für beliebige  $s, t$  mit  $1 \leq s \leq st \leq 10$

$$\begin{aligned} P(X \in M_{[1,t)}) &= P\left(\log_{10} X \in \bigcup_{n=-\infty}^{\infty} [\log_{10}(10^n), \log_{10}(t \cdot 10^n))\right) \\ &= P\left(\log_{10} X \in \bigcup_{n=-\infty}^{\infty} [n, n + \log_{10} t)\right) \\ &= P\left(\log_{10} X / [\log_{10} X] \in [1, 1 + \log_{10} t)\right) \end{aligned}$$

sowie

$$\begin{aligned}
 P(X \in s \cdot M_{[1,t)}) &= P\left(\log_{10} X \in \bigcup_{n=-\infty}^{\infty} [\log_{10}(s \cdot 10^n), \log_{10}(st \cdots 10^n))\right) \\
 &= P\left(\log_{10} X \in \bigcup_{n=-\infty}^{\infty} [n + \log_{10} s, n + \log_{10} s + \log_{10} t)\right) \\
 &= P\left(\log_{10} X / [\log_{10} X] \in [1 + \log_{10} s, 1 + \log_{10} s + \log_{10} t)\right).
 \end{aligned}$$

Wegen der vorausgesetzten Skaleninvarianz (7.2) sind die obigen Wahrscheinlichkeiten gleich und es gilt mit  $\alpha := \log_{10} s$ ,  $\beta = \log_{10} t$ , dass

$$P\left(\log_{10} X / [\log_{10} X] \in [1, 1 + \beta)\right) = P\left(\log_{10} X / [\log_{10} X] \in [1 + \alpha, 1 + \alpha + \beta)\right).$$

Diese Gleichheit gilt für alle  $\alpha, \beta$  mit  $0 \leq \alpha \leq \alpha + \beta \leq 1$ , folglich besitzt  $\log_{10} X / [\log_{10} X]$  eine Gleichverteilung auf dem Intervall  $[1, 2]$ . Es folgt somit

$$P(X \in M_{[1,t)}) = P\left(\log_{10} X / [\log_{10} X] \in [1, 1 + \log_{10} t)\right) = \log_{10} t \quad \forall t \in [1, 10].$$

Damit erhalten wir folgendes Resultat:

$$\begin{aligned}
 &P(\text{„erste signifikante Ziffer von } X \text{ ist } i\text{“}) \\
 &= P(X \in M_{[i,i+1)}) = P(X \in M_{[1,i+1)}) - P(X \in M_{[1,i)}) \\
 &= \log_{10}(i+1) - \log_{10}(i).
 \end{aligned} \tag{7.3}$$

Aus (7.3) ergibt sich folgende Tabelle:

$i$	1	2	3	4	5	6	7	8	9
$P(\text{„}i\text{“})$	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046

Tabelle 1: Wahrscheinlichkeiten der Anfangsziffern

## 8 Approximationen der Binomialverteilung

Die Binomialverteilung ist eines der Wahrscheinlichkeitsmaße, auf das im Schulunterricht ausführlich eingegangen wird bzw. werden sollte. Die Wahrscheinlichkeit, dass man beim  $n$ -fachen Würfeln genau  $k$  Sechsen erhält, wird unter den naheliegenden Voraussetzungen der Unverfälschtheit des Würfels und der Unabhängigkeit der Ergebnisse der jeweiligen Würfe durch eine Binomialwahrscheinlichkeit beschrieben und ist gegeben durch

$$b_{n,1/6}(k) = \binom{n}{k} (1/6)^k (1 - 1/6)^{n-k}. \quad (8.1)$$

Die Wahrscheinlichkeit, bei der Lotterie „6 aus 49“ einen Dreier ohne Zusatzzahl zu erreichen, ist beim einmaligen Spiel gleich

$$p := \frac{\binom{6}{3} \binom{43}{3}}{\binom{49}{6}} \approx 0,0176.$$

Dementsprechend ist die Wahrscheinlichkeit, dass man beim Spielen über 1000 Ziehungen mindestens 10 Dreier ohne Zusatzzahl erreicht, gleich

$$\sum_{k=10}^{1000} \binom{1000}{k} p^k (1-p)^{n-k}. \quad (8.2)$$

Die exakte Berechnung der in (8.1) und (8.2) gegebenen Wahrscheinlichkeiten ist wegen des Auftretens von Fakultäten großer Zahlen sehr mühsam und die Möglichkeit einer einfachen Bestimmung ist sicher wünschenswert. Darüber hinaus gibt es einen weiteren Grund, der denjenigen, welchen an „schönen“ Ergebnissen der Mathematik interessiert sind, überzeugen sollte: Die in diesem Kapitel hergeleitete Approximation der Binomialverteilung durch die Normalverteilung ist ein Spezialfall des sogenannten Zentralen Grenzwertsatzes. Dieser besagt, dass die Verteilung von Summen unabhängiger Zufallsvariablen, **egal wie deren jeweilige Verteilungen sind**, durch die Normalverteilung approximiert werden kann.

Bevor wir nun zur angekündigten Normalapproximation der Binomialverteilung kommen, soll zunächst auf eine alternative Approximationsmöglichkeit durch die Poissonverteilung eingegangen werden. Diese Approximation ist angebracht, wenn es sich um eine zu approximierende  $\text{Bin}(n, p)$ -Verteilung mit großem  $n$  und (sehr) kleinem  $p$  handelt. Beispielsweise lässt sich die in (8.2) angegebene Wahrscheinlichkeit auf diese Weise approximieren.

**Lemma 8.1.** (*Poisson-Approximation der Binomialverteilung*)

$(X_n)_{n \in \mathbb{N}}$  sei eine Folge von Zufallsvariablen mit  $X_n \sim \text{Bin}(n, p/n) \forall n \in \mathbb{N}$ , wobei  $p > 0$ . Dann gilt

$$P(X_n = k) \xrightarrow{n \rightarrow \infty} e^{-p} \frac{p^k}{k!} \quad \forall k \in \mathbb{N}_0 := \{0, 1, \dots\}.$$

*Beweis.* Für  $k \in \mathbb{N}_0$  gilt

$$\begin{aligned}
 P(X_n = k) &= \frac{n!}{k!(n-k)!} \left(\frac{p}{n}\right)^k \left(1 - \frac{p}{n}\right)^{n-k} \\
 &= \frac{1}{k!} \underbrace{\frac{n \cdot (n-1) \cdots (n-k+1)}{n^k}}_{\xrightarrow{n \rightarrow \infty} 1} p^k \underbrace{\left(1 - \frac{p}{n}\right)^n}_{\xrightarrow{n \rightarrow \infty} e^{-p}} \underbrace{\left(1 - \frac{p}{n}\right)^{-k}}_{\xrightarrow{n \rightarrow \infty} 1} \\
 &\xrightarrow{n \rightarrow \infty} e^{-p} \frac{p^k}{k!}.
 \end{aligned}$$

□

Dieses Ergebnis zur Approximation der Binomial- durch die Poissonverteilung kann in zweierlei Hinsicht verschärft werden. Einerseits kann der dabei auftretende Approximationsfehler explizit abgeschätzt werden und andererseits kann man solch ein Ergebnis auch für Summen unabhängiger Bernoulli-verteilter Zufallsvariablen mit nicht notwendigerweise gleichen Parametern  $p_1, \dots, p_n$  beweisen.

**Lemma 8.2.**  $X_1, \dots, X_n$  seien stochastisch unabhängig,  $X_i \sim \text{Bin}(1, p_i)$  für  $i = 1, \dots, n$ . (Die  $X_i$  sind **Bernoulli-verteilt**.) Außerdem sei  $Y$  eine Poisson-verteilte Zufallsvariable mit Parameter  $p := p_1 + \dots + p_n$ . Dann gilt

$$\sum_{k=0}^{\infty} |P(X_1 + \dots + X_n = k) - P(Y = k)| \leq 2 \sum_{i=1}^n p_i^2.$$

**Bemerkung 8.3.** Im Spezialfall von  $p_1 = \dots = p_n = p/n$  ist  $X_1 + \dots + X_n$  binomialverteilt mit Parametern  $n$  und  $p/n$ . In diesem Fall gilt folgende **quantitative** Abschätzung des Approximationsfehlers:

$$\sum_{k=0}^{\infty} |P(X_1 + \dots + X_n = k) - P(Y = k)| \leq \frac{2p^2}{n}.$$

*Beweis von Lemma 8.2.* Dem Beweis liegt eine interessante und in der Stochastik oftmals schlagkräftige Methode zugrunde. Wir machen **nicht** den Versuch, das Resultat auf rein analytischem Wege herzuleiten, da ein solcher Beweis viel zu komplex und unübersichtlich wäre. Stattdessen konstruieren wir eine sogenanntes „*coupling*“ von geeigneten Zufallsvariablen. Diese Idee kann folgendermaßen skizziert werden. Auf einem geeigneten Wahrscheinlichkeitsraum  $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$  werden stochastisch unabhängige Zufallsvektoren  $(\tilde{X}_1), \dots, (\tilde{X}_n)$  konstruiert, so dass für  $i = 1, \dots, n$  gelten

$$\tilde{X}_i \sim \text{Bin}(1, p_i), \tag{8.3}$$

$$\tilde{Y}_i \sim \text{Poisson}(p_i) \quad (8.4)$$

und

$$\tilde{P}(\tilde{X}_i \neq \tilde{Y}_i) \leq p_i^2. \quad (8.5)$$

Aus (8.3) und (8.4) erhalten wir

$$P(X_1 + \dots + X_n = k) = \tilde{P}(\tilde{X}_1 + \dots + \tilde{X}_n = k) \quad \forall k = 0, 1, \dots, n$$

sowie

$$P(Y = k) = \tilde{P}(\tilde{Y}_1 + \dots + \tilde{Y}_n = k) \quad \forall k \in \mathbb{N}_0.$$

Falls zusätzlich noch (8.5) erfüllt ist, so erhalten wir

$$\begin{aligned} & \sum_{k=0}^{\infty} |P(X_1 + \dots + X_n = k) - P(Y = k)| \\ &= \sum_{k=0}^{\infty} |\tilde{P}(\tilde{X}_1 + \dots + \tilde{X}_n = k) - \tilde{P}(\tilde{Y}_1 + \dots + \tilde{Y}_n = k)| \\ &= \sum_{k=0}^{\infty} \left| \tilde{P}\left(\sum_{i=1}^n \tilde{X}_i = k, \sum_{i=1}^n \tilde{X}_i = \sum_{i=1}^n \tilde{Y}_i\right) + \tilde{P}\left(\sum_{i=1}^n \tilde{X}_i = k, \sum_{i=1}^n \tilde{X}_i \neq \sum_{i=1}^n \tilde{Y}_i\right) \right. \\ & \quad \left. - \tilde{P}\left(\sum_{i=1}^n \tilde{Y}_i = k, \sum_{i=1}^n \tilde{X}_i = \sum_{i=1}^n \tilde{Y}_i\right) - \tilde{P}\left(\sum_{i=1}^n \tilde{Y}_i = k, \sum_{i=1}^n \tilde{X}_i \neq \sum_{i=1}^n \tilde{Y}_i\right) \right| \\ &= \underbrace{\sum_{k=0}^{\infty} \tilde{P}\left(\sum_{i=1}^n \tilde{X}_i = k, \sum_{i=1}^n \tilde{X}_i \neq \sum_{i=1}^n \tilde{Y}_i\right)}_{= \tilde{P}\left(\sum_{i=1}^n \tilde{X}_i \neq \sum_{i=1}^n \tilde{Y}_i\right)} + \underbrace{\sum_{k=0}^{\infty} \tilde{P}\left(\sum_{i=1}^n \tilde{Y}_i = k, \sum_{i=1}^n \tilde{X}_i \neq \sum_{i=1}^n \tilde{Y}_i\right)}_{= \tilde{P}\left(\sum_{i=1}^n \tilde{X}_i \neq \sum_{i=1}^n \tilde{Y}_i\right)} \\ &= 2 \tilde{P}\left(\sum_{i=1}^n \tilde{X}_i \neq \sum_{i=1}^n \tilde{Y}_i\right) \\ &\leq 2 \sum_{i=1}^n \tilde{P}(\tilde{X}_i \neq \tilde{Y}_i) \leq 2 \sum_{i=1}^n p_i^2. \end{aligned}$$

Anstelle möglicherweise länglicher analytischer Herleitungen können wir so auf relativ überschaubare Weise zeigen, dass unsere Behauptung gilt. Der Schlüssel zum Erfolg ist natürlich das durch (8.3) - (8.5) beschriebene „*coupling*“ geeigneter Zufallsgrößen.

Es ist noch zu zeigen, dass (8.3) - (8.5) gleichzeitig erfüllt werden können. Wir beschreiben dazu die Konstruktion eines der Zufallsvektoren  $(\tilde{X}_i)$ . Ausgehend von der Ungleichung  $1 - p \leq e^{-p} \forall p \geq 0$  (Mit  $f(p) := e^{-p} - (1 - p)$  gelten  $f(0) = 0$  sowie  $f'(p) = -e^{-p} + 1 > 0 \forall p > 0$ , woraus diese Ungleichung folgt.) definieren wir die Verteilung des Zufallsvek-

tors  $\binom{\tilde{X}_i}{\tilde{Y}_i}$  folgendermaßen:

$$\tilde{P}\left(\binom{\tilde{X}_i}{\tilde{Y}_i} = \binom{0}{0}\right) = \min\{1 - p_i, e^{-p_i}\} = 1 - p_i, \quad (8.6a)$$

$$\tilde{P}\left(\binom{\tilde{X}_i}{\tilde{Y}_i} = \binom{1}{0}\right) = e^{-p_i} - (1 - p_i), \quad (8.6b)$$

$$\tilde{P}\left(\binom{\tilde{X}_i}{\tilde{Y}_i} = \binom{1}{1}\right) = \min\{p_i, e^{-p_i} p_i\} = e^{-p_i} p_i, \quad (8.6c)$$

$$\tilde{P}\left(\binom{\tilde{X}_i}{\tilde{Y}_i} = \binom{1}{k}\right) = e^{-p_i} p_i^k / k! \quad \forall k \geq 2. \quad (8.6d)$$

Damit erhalten wir

$$\begin{aligned} \tilde{P}(\tilde{X}_i = 0) &= \tilde{P}\left(\binom{\tilde{X}_i}{\tilde{Y}_i} = \binom{0}{0}\right) = 1 - p_i, \\ \tilde{P}(\tilde{X}_i = 1) &= \sum_{k=0}^{\infty} \tilde{P}\left(\binom{\tilde{X}_i}{\tilde{Y}_i} = \binom{1}{k}\right) \\ &= e^{-p_i} - (1 - p_i) + e^{-p_i} p_i + \underbrace{\sum_{k=2}^{\infty} e^{-p_i} p_i^k / k!}_{=1 - e^{-p_i} - e^{-p_i} p_i} = p_i \end{aligned}$$

sowie

$$\begin{aligned} \tilde{P}(\tilde{Y}_i = 0) &= \tilde{P}\left(\binom{\tilde{X}_i}{\tilde{Y}_i} = \binom{0}{0}\right) + \tilde{P}\left(\binom{\tilde{X}_i}{\tilde{Y}_i} = \binom{1}{0}\right) \\ &= 1 - p_i + e^{-p_i} - (1 - p_i) = e^{-p_i} p_i^0 / 0!, \\ \tilde{P}(\tilde{Y}_i = k) &= \tilde{P}\left(\binom{\tilde{X}_i}{\tilde{Y}_i} = \binom{1}{k}\right) = e^{-p_i} p_i^k / k! \quad \forall k \geq 1. \end{aligned}$$

Folglich sind (8.3) und (8.4) erfüllt.

Die Festlegungen in (8.6a) und (8.6c) sind so gewählt, dass die maximalen Wahrscheinlichkeiten auf der „Diagonale“ platziert wurden, d.h.,  $\sum_k \tilde{P}\left(\binom{\tilde{X}_i}{\tilde{Y}_i} = \binom{k}{k}\right) = \sum_k \min\{P(X_i = k), e^{-p_i} p_i^k / k!\}$ . Aus (8.6a) und (8.6c) folgt insbesondere

$$\begin{aligned} \tilde{P}(\tilde{X}_i \neq \tilde{Y}_i) &= 1 - \tilde{P}\left(\binom{\tilde{X}_i}{\tilde{Y}_i} = \binom{0}{0}\right) - \tilde{P}\left(\binom{\tilde{X}_i}{\tilde{Y}_i} = \binom{1}{1}\right) \\ &= 1 - (1 - p_i) - e^{-p_i} p_i = p_i \underbrace{(1 - e^{-p_i})}_{\leq p_i} \leq p_i^2, \end{aligned}$$

d.h., (8.5) ist schließlich auch erfüllt.  $\square$

Während die Poisson-Approximation einer  $\text{Bin}(n, p)$ -Verteilung für kleine  $p$  und großes  $n$  angemessen ist, so ist für moderate Werte von  $p$  und großes  $n$  eine Approximation durch die Normalverteilung geeigneter. Die im Weiteren hergeleiteten Approximationen lassen sich am besten so merken: Für  $S_n \sim \text{Bin}(n, p)$  sind Erwartungswert und Varianz gleich

$$ES_n = np =: \mu_n \quad \text{bzw.} \quad \text{var}(S_n) = np(1-p) =: \sigma_n^2.$$

Die Wahrscheinlichkeit, dass  $S_n$  einen Wert  $k$  annimmt, kann mit Hilfe der Dichte einer Normalverteilung mit **denselbenden ersten zwei Momenten** approximiert werden, d.h.,

$$P(S_n = k) \approx \varphi_{\mu_n, \sigma_n^2}(k) = \frac{1}{\sqrt{2\pi \sigma_n^2}} e^{-\frac{(k-\mu_n)^2}{2\sigma_n^2}}. \quad (8.7)$$

Für Anwendungen noch wichtiger ist eine Approximation der Verteilungsfunktion der Binomialverteilung. Eine einfach zu merkende Form dieser Approximation erhält man, indem man  $S_n$  zunächst **standardisiert**, d.h., man betrachtet  $(S_n - \mu_n)/\sigma_n$ . Mit dieser Standardisierung erreicht man, dass  $E[(S_n - \mu_n)/\sigma_n] = 0$  und  $\text{var}((S_n - \mu_n)/\sigma_n) = 1$  sind. Nun erscheint eine Approximation durch die Standardnormalverteilung natürlich zu sein. Wir werden zeigen, dass

$$P\left(a \leq \frac{S_n - \mu_n}{\sigma_n} \leq b\right) \xrightarrow{n \rightarrow \infty} \Phi(b) - \Phi(a) = \int_a^b \varphi_{0,1}(u) du \quad \forall a, b \in \mathbb{R}, a < b, \quad (8.8)$$

gilt. Mit einer einfachen Zusatzüberlegung lässt sich daraus auch die Approximation

$$P\left(\frac{S_n - \mu_n}{\sigma_n} \leq x\right) \xrightarrow{n \rightarrow \infty} \Phi(x) \quad \forall x \in \mathbb{R} \quad (8.9)$$

herleiten. Wir werden zunächst die Approximation (8.7) für die Binomialwahrscheinlichkeiten herleiten, woraus danach die Approximationen (8.8) und (8.9) folgen werden. Zunächst stellen wir eine Näherungsformel für Fakultäten großer Zahlen bereit, welche dann beim Übergang von den Binomialwahrscheinlichkeiten zur Normalverteilungsdichte benötigt wird.

#### Lemma 8.4. (Stirlingsche Formel)

Für  $n \in \mathbb{N}$  gilt

$$n! = \sqrt{2\pi n} e^{-n} n^n (1 + r(n)),$$

wobei

$$r(n) \xrightarrow{n \rightarrow \infty} 0.$$

Der Beweis der Stirlingschen Formel ist rein analytisch und wird an dieser Stelle nicht gebracht. Er findet sich u.a. am Ende von Kapitel 5 in Krenzel [3]. Als Nächstes formulieren nun die mit (8.7) angekündigte Approximation der Binomialwahrscheinlichkeiten. Im Hinblick auf die Erweiterung zu den globalen Resultaten (8.8) und (8.9) wird das Resultat so formuliert, dass die Approximationseigenschaft **gleichmäßig** auf einem hinreichend großen Bereich erfüllt ist.

**Satz 8.5. (Lokaler Grenzwertsatz von de Moivre-Laplace)**

Es seien  $S_n \sim \text{Bin}(n, p) \forall n \in \mathbb{N}$ , wobei  $0 < p < 1$ .

Dann gilt, mit  $\mu_n = np$  und  $\sigma_n^2 = np(1-p)$ ,

$$P(S_n = k) = \frac{1}{\sqrt{2\pi} \sigma_n} e^{-\frac{(k-\mu_n)^2}{2\sigma_n^2}} (1 + R(n, k)),$$

wobei

$$\max_{k \in K_n} |R(n, k)| \xrightarrow{n \rightarrow \infty} 0,$$

falls  $K_n = \{k \in \{0, 1, \dots, n\} : |k - np| \leq \alpha_n n^{2/3}\}$  und  $(\alpha_n)_{n \in \mathbb{N}}$  eine beliebige Nullfolge (d.h.,  $\alpha_n \xrightarrow{n \rightarrow \infty} 0$ ) ist.

**Bemerkung 8.6.** Dass der Gültigkeitsbereich des in Satz 8.5 formulierten Resultats zur Approximation der Binomialwahrscheinlichkeiten für asymptotische Betrachtungen groß genug ist, kann man anhand folgender Überlegungen bereits erahnen. Wählt man nämlich die Nullfolge  $(\alpha_n)_{n \in \mathbb{N}}$  so, dass  $\alpha_n n^{1/6} \xrightarrow{n \rightarrow \infty} \infty$ , so folgt wegen  $ES_n = \mu_n$  und  $\text{var}(S_n) = \sigma_n^2$  aus der Tschebyscheff-Ungleichung, dass

$$\begin{aligned} P(S_n \notin K_n) &= P(|S_n - \mu_n| > \alpha_n n^{2/3}) \\ &\leq \frac{\sigma_n^2}{\alpha_n^2 n^{4/3}} = \frac{p(1-p)}{\alpha_n^2 n^{1/3}} \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

*Beweis von Satz 8.5.* Wir müssen zeigen, dass  $\sup_{k \in K_n} |R(n, k)| \xrightarrow{n \rightarrow \infty} 0$  gilt. Dazu sei  $(k_n)_{n \in \mathbb{N}}$  eine beliebige Folge mit  $k_n \in K_n \forall n \in \mathbb{N}$ . Wir werden zeigen, dass

$$P(S_n = k_n) = \frac{1}{\sqrt{2\pi} \sigma_n} e^{-\frac{(k_n - \mu_n)^2}{2\sigma_n^2}} (1 + R(n)) \quad (8.10)$$

gilt, wobei  $(R(n))_{n \in \mathbb{N}}$  eine Nullfolge ist. Um die bei den Binomialwahrscheinlichkeiten vorkommenden Fakultäten loszuwerden, benutzen wir die Stirlingsche Formel (siehe Lemma 8.4):  $n! = \sqrt{2\pi} \sqrt{n} e^{-n} n^n (1 + r(n))$ , wobei  $r_n \xrightarrow{n \rightarrow \infty} 0$ . Nun gilt

$$\begin{aligned} P(S_n = k_n) &= \frac{n!}{k_n! (n - k_n)!} p^{k_n} (1 - p)^{n - k_n} \\ &= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{k_n (n - k_n)}} \underbrace{\frac{e^{-n}}{e^{-k_n} e^{-(n - k_n)}}}_{=1} \frac{n^n}{k_n^{k_n} (n - k_n)^{n - k_n}} p^{k_n} (1 - p)^{n - k_n} \frac{1 + r(n)}{(1 + r(k_n))(1 + r(n - k_n))}. \end{aligned}$$

Da  $|k_n - np| \leq \alpha_n n^{2/3}$ , so folgen  $k_n/n \xrightarrow{n \rightarrow \infty} p$  und  $(n - k_n)/n \xrightarrow{n \rightarrow \infty} 1 - p$ . Daraus folgt

$$\sqrt{\frac{n}{k_n (n - k_n)}} = \frac{1}{\sqrt{np(1-p)}} (1 + r_1(n)), \quad r_1(n) \xrightarrow{n \rightarrow \infty} 0.$$

Darüber hinaus gelten  $k_n \xrightarrow[n \rightarrow \infty]{} \infty$  und  $n - k_n \xrightarrow[n \rightarrow \infty]{} \infty$ , woraus

$$\frac{1 + r(n)}{(1 + r(k_n))(1 + r(n - k_n))} = (1 + r_2(n)), \quad r_2(n) \xrightarrow[n \rightarrow \infty]{} 0$$

folgt. Somit erhalten wir die folgende, uns dem angestrebten Ergebnis näher bringende Darstellung der Binomialwahrscheinlichkeit:

$$P(S_n = k_n) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{np(1-p)}} \left(\frac{p}{k_n/n}\right)^{k_n} \left(\frac{1-p}{(n-k_n)/n}\right)^{n-k_n} (1 + r_1(n)) (1 + r_2(n)). \quad (8.11)$$

Wir zeigen jetzt noch, dass

$$\left(\frac{p}{k_n/n}\right)^{k_n} \left(\frac{1-p}{(n-k_n)/n}\right)^{n-k_n} = e^{-\frac{(k_n - np)^2}{2np(1-p)}} (1 + r_3(n)), \quad r_3(n) \xrightarrow[n \rightarrow \infty]{} 0, \quad (8.12)$$

woraus zusammen mit (8.11) die zu zeigende Relation (8.10) folgt.

Zunächst stellen wir fest, dass (8.12) äquivalent ist zu

$$\ln \left[ \left(\frac{p}{k_n/n}\right)^{k_n} \left(\frac{1-p}{(n-k_n)/n}\right)^{n-k_n} \right] = -\frac{(k_n - np)^2}{2np(1-p)} + r_4(n), \quad r_4(n) \xrightarrow[n \rightarrow \infty]{} 0, \quad (8.13)$$

was wir nun im Folgenden zeigen werden. Dazu legen wir uns das Problem noch geeignet zurecht. Es gilt

$$\begin{aligned} & \ln \left[ \left(\frac{p}{k_n/n}\right)^{k_n} \left(\frac{1-p}{(n-k_n)/n}\right)^{n-k_n} \right] \\ &= -n \left\{ \frac{k_n}{n} \ln \left(\frac{k_n/n}{p}\right) - \left(1 - \frac{k_n}{n}\right) \ln \left(\frac{1 - k_n/n}{1-p}\right) \right\} = -n g(k_n/n), \end{aligned}$$

wobei  $g(t) := t \ln(t/p) - (1-t) \ln((1-t)/(1-p))$ : Wir betrachten die Funktion  $g$  und entwickeln sie im Punkt  $t = p$  in eine Taylorreihe. Es gelten

$$\begin{aligned} g(p) &= 0, \\ g'(t) &= \ln\left(\frac{t}{p}\right) + \underbrace{t \frac{p}{t}}_{=1} - \ln\left(\frac{1-t}{1-p}\right) + \underbrace{(1-t) \frac{1-p}{1-t} \frac{-1}{1-p}}_{=-1} \implies g'(p) = 0, \\ g''(t) &= \frac{p}{t} \frac{1}{p} + \frac{1-p}{1-t} \frac{1}{1-p} = \frac{1}{t} + \frac{1}{1-t} \implies g''(p) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}, \\ g^{(3)}(t) & \text{ ist beschränkt in einer Umgebung von } p. \end{aligned}$$

Daraus folgt

$$g(t) = \frac{(t-p)^2}{2} \frac{1}{p(1-p)} + O(|t-p|^3)$$

und somit

$$\ln \left[ \left(\frac{p}{k_n/n}\right)^{k_n} \left(\frac{1-p}{(n-k_n)/n}\right)^{n-k_n} \right] = -n g\left(\frac{k_n}{n}\right) = -n \frac{(k_n/n - p)^2}{2p(1-p)} + O(n |k_n/n - p|^3).$$

Das Restglied konvergiert gegen 0, falls  $|k_n/n - p| = o(n^{-1/3})$ , was äquivalent zu  $|k_n - np| = o(n^{2/3})$  ist. Also gilt (8.13) und somit auch (8.12).  $\square$

Die in Satz 8.5 angegebene Approximation der Binomialwahrscheinlichkeiten kann zur Herleitung einer Approximation der Verteilungsfunktion verwendet werden. Dabei ist zu beachten, dass eine gleichmäßige Abschätzung des Approximationsfehlers in den Grenzen der Bereiche  $K_n \subseteq \{0, 1, \dots, n\}$  gegeben ist. Aus diesem Grund wird zunächst eine Approximation für  $P\left(\frac{S_n - \mu_n}{\sigma_n} \in [a, b]\right)$  hergeleitet; die Approximation der Verteilungsfunktion folgt danach mit einem Extraargument.

**Satz 8.7. (Globaler Grenzwertsatz von de Moivre-Laplace)**

Es seien  $S_n \sim \text{Bin}(n, p) \forall n \in \mathbb{N}$ , wobei  $0 < p < 1$ .

Dann gilt, für  $-\infty < a < b < \infty$ , dass

$$P\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) \xrightarrow{n \rightarrow \infty} \Phi(b) - \Phi(a) = \int_a^b \varphi(u) du,$$

wobei  $\Phi$  und  $\varphi$  die Verteilungsfunktion bzw. Wahrscheinlichkeitsdichte einer Standardnormalverteilung sind. ( $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ ,  $\Phi(x) = \int_{-\infty}^x \varphi(u) du$ )

*Beweis.* Es seien  $\mu_n = np$  und  $\sigma_n^2 = np(1-p)$  Erwartungswert bzw. Varianz von  $S_n$ . Nun gilt

$$a \leq \frac{S_n - \mu_n}{\sigma_n} \leq b$$

genau dann, wenn  $a_n \leq S_n \leq b_n$ , wobei  $a_n := \lceil \mu_n + a\sigma_n \rceil$  die kleinste ganze Zahl, welche nicht kleiner als  $\mu_n + a\sigma_n$  ist und  $b_n := \lfloor \mu_n + b\sigma_n \rfloor$  die größte ganze Zahl, welche nicht größer als  $\mu_n + b\sigma_n$  ist, bezeichnen. Nach dem lokalen Grenzwertsatz von de Moivre-Laplace (Satz 8.5) erhalten wir

$$\begin{aligned} P\left(a \leq \frac{S_n - \mu_n}{\sigma_n} \leq b\right) &= \sum_{k=a_n}^{b_n} P(S_n = k) \\ &= \sum_{k=a_n}^{b_n} \frac{1}{\sqrt{2\pi} \sigma_n} e^{-\frac{(k-\mu_n)^2}{2\sigma_n^2}} (1 + R(n, k)). \end{aligned}$$

Nun gelten  $a_n - \mu_n = O(n^{1/2})$  und  $b_n - \mu_n = O(n^{1/2})$ . Wenn also bei der Definition von  $K_n$  die Nullfolge  $(\alpha_n)_{n \in \mathbb{N}}$  so gewählt wird, dass  $\alpha_n n^{1/6} \xrightarrow{n \rightarrow \infty} 0$  gilt, dann folgt  $a_n, b_n \in K_n$  für hinreichend große  $n$  und wir erhalten nach Satz 8.5, dass

$$\max_{k: a_n \leq k \leq b_n} |R(n, k)| \xrightarrow{n \rightarrow \infty} 0.$$

Somit gilt

$$\begin{aligned}
P\left(a \leq \frac{S_n - \mu_n}{\sigma_n} \leq b\right) &= \sum_{k=a_n}^{b_n} \frac{1}{\sqrt{2\pi} \sigma_n} e^{-\frac{(k-\mu_n)^2}{2\sigma_n^2}} + o(1) \\
&= \sum_{k=a_n}^{b_n} \int_{k-1/2}^{k+1/2} \frac{1}{\sqrt{2\pi} \sigma_n} e^{-\frac{(t-\mu_n)^2}{2\sigma_n^2}} dt + o(1) \\
&= \int_{a_n-1/2}^{b_n+1/2} \frac{1}{\sqrt{2\pi} \sigma_n} e^{-\frac{(t-\mu_n)^2}{2\sigma_n^2}} dt + o(1) \\
&= \int_{\frac{(a_n-1/2)-\mu_n}{\sigma_n}}^{\frac{(b_n+1/2)-\mu_n}{\sigma_n}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du + o(1) \\
&= \int_a^b \varphi(u) du + o(1).
\end{aligned}$$

□

**Folgerung 8.8.** Es seien  $S_n \sim \text{Bin}(n, p) \forall n \in \mathbb{N}$ , wobei  $0 < p < 1$ .

Dann gelten, für  $x \in \mathbb{R}$ , dass

$$(i) \quad P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq x\right) \xrightarrow{n \rightarrow \infty} \Phi(x),$$

$$(ii) \quad P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \geq x\right) \xrightarrow{n \rightarrow \infty} 1 - \Phi(x) = \Phi(-x),$$

wobei  $\Phi$  die Verteilungsfunktion einer Standardnormalverteilung ist.

*Beweis.* (i) Es sei  $\epsilon > 0$  beliebig. Nach der Tschebyscheff-Ungleichung folgt für  $K > 0$ , dass

$$P\left(\left|\frac{S_n - np}{\sigma_n}\right| > K\right) \leq \frac{\text{var}(S_n/\sigma_n)}{K^2} = \frac{1}{K^2}.$$

Da  $\Phi(x) \xrightarrow{x \rightarrow -\infty} 0$ , so existiert ein  $K_\epsilon < \infty$ , sodass

$$P\left(\frac{S_n - np}{\sigma_n} < K_\epsilon\right) + \Phi(-K_\epsilon) \leq \epsilon.$$

Damit folgt jedoch aus Satz 8.7, dass

$$\begin{aligned}
&\left|P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq x\right) - \Phi(x)\right| \\
&\leq \left|P\left(-K_\epsilon \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq x\right) - [\Phi(x) - \Phi(-K_\epsilon)]\right| + \epsilon \\
&\leq 2\epsilon \quad \forall n \geq N_\epsilon
\end{aligned}$$

und  $N_\epsilon$  hinreichend groß. Damit ist (i) bewiesen.

(ii) Analog.

□

## 9 Grundbegriffe der Mathematischen Statistik

In diesem Kapitel werden die beiden wichtigsten Aufgabenstellungen in der Mathematischen Statistik betrachtet, das „Schätzen“ von unbekanntem Parametern und das „Testen“ von Annahmen an das zugrundegelegte Modell. Wir beginnen mit einem einfachen Beispiel.

### Beispiel: Schätzen eines Fischbestandes (Fang-Wiederfang-Methode)

Wir nehmen an, dass ein Gewässer eine unbekannte Zahl  $N$  von Fischen enthält und dass diese Anzahl experimentell bestimmt werden soll. Da eine direkte Bestimmung dieser Anzahl durch Auspumpen des Gewässers und anschließendes Zählen der Fische nicht praktikabel scheint, wird folgendermaßen vorgegangen. In einem ersten Schritt werden  $M$  Fische gefangen, markiert und anschließend wieder ausgesetzt. Nach einer gewissen Wartezeit wird ein zweiter Fischzug gestartet. Es wird so lange gefischt, bis  $n$  Fische gefangen sind und die Anzahl  $x$  der davon markierten wird bestimmt. Eine plausible Approximation  $\hat{N}(x)$  ergibt sich aus der naheliegenden Überlegung, dass der Anteil  $x/n$  der als markiert identifizierten Fische aus dem zweiten Fang ungefähr gleich dem Anteil  $M/N$  der markierten Fische aus dem ersten Fang sein sollte. Wegen  $x/n \approx M/N$  ist eine Approximation  $\hat{N}(x)$  für  $N$  gegeben durch die zu  $Mn/x$  nächstgelegene ganze Zahl. Eine solche Approximation nennt man in der Statistik **Schätzwert** oder **Schätzung**.

Die so gewonnene heuristisch abgeleitete Schätzung ergibt sich auch aus einem Ansatz, der auf einem weit verbreiteten Prinzip der Statistik beruht. Dazu spezifizieren wir zunächst ein entsprechendes Modell. Wir nehmen an, dass die Anzahl  $M$  bzw.  $n$  der jeweils gefangenen Fische bewusst gewählt ist, da nach Erreichen dieser Anzahl abgebrochen wurde. Dagegen ist jedoch die Anzahl  $x$  der als markiert erkannten Fische beim zweiten Fang zufällig. Es ist naheliegend, diese Anzahl als **Realisierung** einer Zufallsvariable  $X$  anzunehmen. Die zugrundeliegende Situation entspricht zumindest approximativ jener beim zufälligen Ziehen aus einer Urne ohne Zurücklegen. Wir nehmen daher an, dass die Zufallsvariable  $X$  eine hypergeometrische Verteilung mit Parametern  $N$ ,  $M$  und  $n$  besitzt, d.h.

$$P(X = x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} \quad \forall x \in \{0, 1, \dots, n\},$$

wobei der Binomialkoeffizient  $\binom{n}{k}$  durch

$$\binom{n}{k} = \begin{cases} \frac{n!}{k!(n-k)!}, & \text{falls } 0 \leq k \leq n, \\ 0 & \text{sonst} \end{cases}$$

gegeben ist. An dieser Stelle wird ein typischer Unterschied zu klassischen Problemstellungen in der Wahrscheinlichkeitstheorie klar. In der Wahrscheinlichkeitstheorie ist durch die Angabe eines Wahrscheinlichkeitsraumes  $(\Omega, \mathcal{A}, P)$  sowie darauf aufbauender Zufallsvariablen  $X_1, X_2, \dots$  die Ausgangssituation für weitergehende Untersuchungen exakt bestimmt. Nach Fixierung solch einer Ausgangssituation werden dann Gesetzmäßigkeiten hergeleitet. Hier wurde auch eine Modellannahme getroffen, nämlich dass die zufällige Anzahl der als markiert erkannten Fische eine hypergeometrische Verteilung besitzt. (Die explizite Angabe eines geeigneten Wahrscheinlichkeitsraumes  $(\Omega, \mathcal{A}, P)$  ist

an dieser Stelle erst einmal nicht notwendig.) Der wesentliche Unterschied zu typischen Problemstellungen in der Wahrscheinlichkeitstheorie besteht nun darin, dass wir nicht vorab spezifizieren können, wie groß die Gesamtzahl  $N$  von Fischen in dem Gewässer ist. (Andernfalls müssten wir uns auch gar nicht die Mühe machen,  $N$  zu schätzen.) Ein Verfahren zur Schätzung von  $N$  sollte so beschaffen sein, dass es für alle grundsätzlich möglichen Werte von  $N$  zu einer guten Approximation führt. Es ist üblich, die Abhängigkeit der Verteilung von  $X$  durch Hinzufügen des unbekanntes Parameters kenntlich zu machen, d.h., die obige Angabe des Verteilungsgesetzes sollte in der Form

$$P_N(X = x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} \quad \forall x \in \{0, 1, \dots, n\}$$

geschehen, wobei  $N$  den unbekanntes Parameter bezeichnet. Die **Maximum-Likelihood-Methode** zur Schätzung von  $N$  wird durch die Tatsache motiviert, dass Ereignisse mit einer großen Wahrscheinlichkeit häufiger auftreten als solche mit einer kleinen Wahrscheinlichkeit. Dieser Heuristik folgend ist der Schätzwert  $\hat{N}(x)$  nach der Maximum-Likelihood-Methode jener Wert, der  $P_N(X = x)$  maximiert, d.h.

$$P_{\hat{N}(x)}(X = x) = \sup \{P_N(X = x) : N \in \mathbb{N}_0\}.$$

Zur Bestimmung von  $\hat{N}(x)$  betrachten wir die Funktion  $N \mapsto P_N(X = x)$ . Es gilt

$$\begin{aligned} P_N(X = x) &> P_{N-1}(X = x) \\ \iff \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} &> \frac{\binom{M}{x} \binom{N-1-M}{n-x}}{\binom{N-1}{n}} \\ \iff (N-M)(N-n) &> (N-M-(n-x))N \\ \iff N &< Mn/x. \end{aligned}$$

Analog sieht man, dass

$$P_N(X = x) = P_{N-1}(X = x) \iff N = Mn/x$$

und

$$P_N(X = x) < P_{N-1}(X = x) \iff N > Mn/x.$$

Falls nun  $Mn/x$  eine ganze Zahl ist, so lässt sich folgern, dass  $P_N(X = x)$  sowohl durch  $N = Mn/x - 1$  als auch  $N = Mn/x$  maximiert wird. Falls  $Mn/x$  keine ganze Zahl ist, so wird  $P_N(X = x)$  durch  $N = \lfloor Mn/x \rfloor$  maximiert, wobei  $\lfloor k \rfloor$  die größte ganze Zahl ist, welche  $\leq k$  ist. Ein entsprechend gewählter Schätzwert  $\hat{N}(x)$  ist der gesuchte Schätzwert nach der Maximum-Likelihood-Methode.

Das Maximum-Likelihood-Prinzip kann auch zum Schätzen von Parametern bei Verteilungsfamilien, welche Dichten besitzen, angewendet werden. Angenommen,  $X_1, \dots, X_n$  seien stochastisch unabhängige Zufallsvariable, welche jeweils eine Wahrscheinlichkeitsdichte  $p_\theta$  besitzen, wobei  $\theta \in \Theta$  ein unbekanntes Parameter ist. Wir betrachten den

Zufallsvektor  $X := (X_1, \dots, X_n)^T$ . Nun gilt

$$\begin{aligned}
& P_\theta(X \in [a_1, b_1] \times \dots \times [a_n, b_n]) \\
&= P_\theta(X_1 \in [a_1, b_1], \dots, X_n \in [a_n, b_n]) \\
&= \prod_{i=1}^n P_\theta(X_i \in [a_i, b_i]) \\
&= \int_{a_1}^{b_1} p_\theta(x_1) dx_1 \times \dots \times \int_{a_{n-1}}^{b_{n-1}} p_\theta(x_{n-1}) dx_{n-1} \times \int_{a_n}^{b_n} p_\theta(x_n) dx_n \\
&= \int_{a_1}^{b_1} p_\theta(x_1) dx_1 \times \dots \times \int_{a_{n-2}}^{b_{n-2}} p_\theta(x_{n-2}) dx_{n-2} \times \int_{a_{n-1}}^{b_{n-1}} \left[ \int_{a_n}^{b_n} p_\theta(x_n) p_\theta(x_{n-1}) dx_n \right] dx_{n-1} \\
&= \dots = \int_{a_1}^{b_1} \dots \int_{a_n}^{b_n} \underbrace{\prod_{i=1}^n p_\theta(x_i)}_{=: p_\theta^{(n)}(x)} dx_n \dots dx_1.
\end{aligned}$$

Die Wahrscheinlichkeit, dass  $X$  einen Wert in  $[a_1, b_1] \times \dots \times [a_n, b_n]$  annimmt, kann durch iterierte Riemann-Integrale, welche den Bereich  $[a_1, b_1] \times \dots \times [a_n, b_n]$  abdecken, dargestellt werden. Das deutet an, dass  $p_\theta^{(n)}: \mathbb{R}^n \rightarrow [0, \infty)$  mit  $p_\theta^{(n)}(x) = \prod_{i=1}^n p_\theta(x_i)$  als Dichte des Zufallsvektors  $X$  unter  $P_\theta$  angesehen werden kann.

(Mit Hilfsmitteln aus der Maßtheorie kann man zeigen, dass  $P_\theta(X \in B) = \int_B p_\theta^{(n)}(x) d\lambda^n(x)$  für beliebige Borel-Mengen  $B \subseteq \mathbb{R}^n$  gilt, wobei  $\lambda^n$  das sogenannte Lebesgue-Maß auf diesen Mengen ist.)

Falls nun Realisierungen  $x_1, \dots, x_n$  (unsere **Daten**) von  $X_1, \dots, X_n$  vorliegen, so erhält man den Maximum-Likelihood-Schätzwert  $\hat{\theta}(x)$  für  $\theta$  durch Maximierung von  $p_\theta^{(n)}(x)$ . Wir fassen die beiden Fälle durch die folgende Definition nochmals zusammen.

**Definition 9.1.** *Gegeben sei eine Realisierung  $x = (x_1, \dots, x_n)^T$  des Zufallsvektors  $X = (X_1, \dots, X_n)^T$ , wobei  $X_1, \dots, X_n$  stochastisch unabhängig seien. Die  $X_i$  mögen entweder eine von einem Parameter  $\theta \in \Theta$  abhängende diskrete Verteilung oder eine Verteilung mit einer Dichte  $p_\theta$  besitzen.*

Dann ist die **Likelihood-Funktion**  $L(\cdot, x): \Theta \rightarrow [0, \infty)$  gegeben durch

$$L(\theta, x) = \begin{cases} \prod_{i=1}^n P_\theta(X_i = x_i), & \text{falls die } X_i \text{ diskret verteilt sind,} \\ \prod_{i=1}^n p_\theta(x_i), & \text{falls die } X_i \text{ Dichten besitzen.} \end{cases}$$

Falls  $X = x$ , so ist der **Schätzwert**  $\hat{\theta}(x)$  nach der Maximum-Likelihood-Methode gegeben durch

$$L(\hat{\theta}(x), x) = \sup \{L(\theta, x): \theta \in \Theta\}.$$

Der **Schätzer** nach der Maximum-Likelihood-Methode (Maximum-Likelihood-Schätzer)  $\hat{\theta}(X)$  ist die entsprechende Zufallsvariable, d.h.,

$$L(\hat{\theta}(X), X) = \sup \{L(\theta, X): \theta \in \Theta\}.$$

**Bemerkung 9.2.** Schätzungen nach der Maximum-Likelihood-Methode sind **implizit** als Lösung eines Maximierungsproblems definiert. Somit sind die Existenz und Eindeutigkeit einer Maximum-Likelihood-Schätzung im Allgemeinen nicht gesichert. Für die meisten „Lehrbuch-Beispiele“ zeigt sich jedoch, dass Schätzer nach der Maximum-Likelihood-Methode existieren (das entsprechende Supremum wird angenommen) und auch eindeutig bestimmt sind.

**Beispiel: Maximum-Likelihood-Schätzer bei Normalverteilung**

Wir nehmen an, dass Realisierungen  $x_1, \dots, x_n$  von unabhängigen und identisch verteilten Zufallsvariablen  $X_1, \dots, X_n$  beobachtet werden, wobei  $X_i \sim \mathcal{N}(\mu, \sigma^2)$  ( $i = 1, \dots, n$ ) für ein  $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \in \Theta := \mathbb{R} \times (0, \infty)$  gelte. Falls  $X = (X_1, \dots, X_n)^T = x = (x_1, \dots, x_n)^T$ , wie sieht nun der entsprechende Schätzwert  $\hat{\theta}(x)$  nach der Maximum-Likelihood Methode aus?

Die Zufallsvariablen  $X_1, \dots, X_n$  besitzen Dichten. Wir erhalten

$$\begin{aligned} L(\theta, x) &= \prod_{i=1}^n p_{\theta}(x_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x}_n + \bar{x}_n - \mu)^2\right\} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2} \left[ \sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(\bar{x}_n - \mu)^2 \right]\right\}. \end{aligned}$$

Nun ist offensichtlich, dass bei beliebiger Wahl von  $\sigma^2$  die rechte Seite der Gleichungskette durch  $\mu = \bar{x}_n$  maximiert wird. Also gilt  $\hat{\mu}(x) = \bar{x}_n$ . Wir maximieren jetzt noch die Funktion  $\sigma^2 \mapsto L\left(\begin{pmatrix} \bar{x}_n \\ \sigma^2 \end{pmatrix}, x\right)$ . Nach Logarithmieren erhalten wir

$$\ln L\left(\begin{pmatrix} \bar{x}_n \\ \sigma^2 \end{pmatrix}, x\right) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

und es gilt

$$\frac{d}{d\sigma^2} \ln L\left(\begin{pmatrix} \bar{x}_n \\ \sigma^2 \end{pmatrix}, x\right) = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \bar{x}_n)^2.$$

Für  $\sum_{i=1}^n (x_i - \bar{x}_n)^2 > 0$  ist diese Ableitung genau dann Null, wenn  $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$ . Darüber hinaus ist leicht zu sehen, dass  $L\left(\begin{pmatrix} \bar{x}_n \\ \sigma^2 \end{pmatrix}, x\right) \xrightarrow{\sigma^2 \rightarrow \infty} 0$  und  $\ln L\left(\begin{pmatrix} \bar{x}_n \\ \sigma^2 \end{pmatrix}, x\right) \xrightarrow{\sigma^2 \rightarrow 0} -\infty$ .

Somit hat die Funktion  $\sigma^2 \mapsto L\left(\begin{pmatrix} \bar{x}_n \\ \sigma^2 \end{pmatrix}, x\right)$  bei  $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$  ihr Maximum. Zusammenfassend kann man sagen, dass bei  $x$  mit  $\sum_{i=1}^n (x_i - \bar{x}_n)^2 > 0$  der gesuchte Maximum-Likelihood-Schätzwert gegeben ist durch

$$\hat{\theta}(x) = \left( \begin{pmatrix} \bar{x}_n \\ \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \end{pmatrix} \right).$$

Im Falle von  $\sum_{i=1}^n (x_i - \bar{x}_n)^2 = 0$  nimmt die Funktion  $\theta \mapsto L(\theta, x)$  ihr Supremum nicht an und der Maximum-Likelihood-Schätzwert ist folglich nicht definiert. Das ist allerdings zu verschmerzen, da für  $n > 1$   $\sum_{i=1}^n (X_i - \bar{X}_n)^2 > 0$  mit Wahrscheinlichkeit 1 eintritt. Der entsprechende Maximum-Likelihood-Schätzer ist nun

$$\hat{\theta}(X) = \left( \left( \frac{\bar{X}_n}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2} \right) \right).$$

An dieser Stelle soll noch kurz der Frage nachgegangen werden, ob denn der Maximum-Likelihood-Ansatz überhaupt etwas taugt. Diese Vorgehensweise ist sicherlich heuristisch motiviert. Andererseits liegt diesem Prinzip nicht das Streben nach einem Verfahren zugrunde, welches in einem wohldefinierten Sinn optimal ist. Einen gewissen Hinweis auf die Tauglichkeit dieses Ansatzes gibt bereits das eben betrachtete Beispiel. Unter der Voraussetzung, dass  $(X_i)_{i \in \mathbb{N}}$  eine Folge von unabhängigen und identisch verteilten Zufallsvariablen mit  $X_i \sim \mathcal{N}(\mu, \sigma^2)$  ist, folgen nach der Faltungsformel (siehe Satz 7.7)

$$X_1 + \cdots + X_n \sim \mathcal{N}(n\mu, n\sigma^2)$$

sowie aus der Dichtetransformationsformel (siehe Satz 7.10)

$$\bar{X}_n = (X_1 + \cdots + X_n)/n \sim \mathcal{N}(\mu, \sigma^2/n).$$

Falls nun  $\hat{\mu}_n := \bar{X}_n$  den auf einer Stichprobe vom Umfang  $n$  beruhenden Maximum-Likelihood-Schätzer des Lokationsparameters  $\mu$  bezeichnet, so gilt nach der Tschebyscheff-Ungleichung (diese gilt auch für stetige Zufallsvariable)

$$P_\theta(|\hat{\mu}_n - \mu| > \epsilon) \leq \frac{\text{var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n \epsilon^2} \xrightarrow{n \rightarrow \infty} 0 \quad \forall \epsilon > 0.$$

Mit wachsendem Stichprobenumfang  $n$  konvergiert somit  $\mu_n$  gegen den zu schätzenden Wert. Da andererseits mit wachsendem Stichprobenumfang auch ein zunehmendes Maß an Information über den zugrundeliegenden (wahren) Parameter geliefert wird, kann und sollte diese Eigenschaft wohl eine minimale Forderung an jede Schätzmethode sein.

**Definition 9.3.** Gegeben sei eine Folge  $(X_i)_{i \in \mathbb{N}}$  von identisch verteilten Zufallsvariablen, deren jeweilige Verteilung von einem Parameter  $\theta \in \Theta$  abhängt. Für  $n \in \mathbb{N}$  sei  $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$  ein Schätzer, der von  $X_1, \dots, X_n$  abhängt.

Die Folge  $(\hat{\theta}_n)_{n \in \mathbb{N}}$  heißt **konsistent für  $\theta$** , falls

$$\hat{\theta}_n \xrightarrow{P_\theta} \theta \quad \forall \theta \in \Theta$$

gilt, d.h.,

$$P_\theta(\|\hat{\theta}_n - \theta\| \geq \epsilon) \xrightarrow{n \rightarrow \infty} 0 \quad \forall \theta \in \Theta.$$

(Hier bezeichnet  $\|\cdot\|$  eine geeignete, z.B. die Euklidische Norm.)

Wir betrachten noch ein weiteres Beispiel. In Tabelle 2 auf Seite 91 dieses Skriptes sind die Schlusskurse vom 22.01.2021 der 102 im Nasdaq-100 gelisteten Aktien aufgeführt. (Der Nasdaq-100 ist ein Börsenindex, der aus 102 Aktien besteht, die von 100 der größten Nichtfinanz-Unternehmen ausgegeben wurden, die an der Nasdaq-Börse notiert sind.) Dabei werden die in US\$ angegebenen Kurse an der Nasdaq (*National Association of Securities Dealers Automated Quotations*) in New York sowie die in Euro angegebenen Schlusskurse der Frankfurter Börse aufgeführt. Nach unserer Abhandlung zum Benford-Gesetz kann man sich nun die Frage stellen, inwieweit das dadurch beschriebene Phänomen auch für Aktienkurse zutrifft. Wir richten daher den Fokus nicht auf die Aktienkurse selbst, sondern auf die ersten signifikanten Ziffern dieser Kurse. Deren Häufigkeit sind in Tabelle 3 dargestellt. Die Entwicklung von Aktienkursen wird in der Finanzmathematik durch stochastische Modelle beschrieben. Wir treffen hier die vereinfachende Annahme, dass die Häufigkeit des Auftretens des Wertes  $i$  für die erste signifikante Ziffer durch eine Binomialverteilung mit Parametern  $n$  und  $\theta$  beschrieben werden kann. Da wir die Kurse an den Börsen in New York und Frankfurt getrennt betrachten, wählen wir  $n = 102$ . Die Wahrscheinlichkeit des Auftretens des Wertes  $i$  an der ersten Stelle bezeichnen wir mit  $\theta_{i,1}$  (für die Nasdaq in New York) bzw. mit  $\theta_{i,2}$  (für Frankfurt). Basierend auf den in Tabelle 3 angegebenen Daten wollen wir die Parameter  $\theta_{i,j}$  ( $i = 1, 2, \dots, 9, j = 1, 2$ ) mit der Maximum-Likelihood-Methode schätzen. Um zu einer allgemeinen Lösung zu kommen, nehmen wir an, dass die Realisierung  $x$  einer Zufallsvariable  $X \sim \text{Bin}(n, \theta)$  vorliegt, wobei  $\theta \in \Theta := [0, 1]$  der zu schätzende Parameter ist. Die Likelihood-Funktion  $L$  ist gegeben durch  $L(\theta, x) = P_\theta(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$ . Falls  $0 < x < n$ , so gilt

$$\begin{aligned} \frac{d}{d\theta} L(\theta, x) &= \binom{n}{x} \{x\theta^{x-1}(1-\theta)^{n-x} - (n-x)\theta^x(1-\theta)^{n-x-1}\} = 0 \\ \iff x(1-\theta) &= \theta(n-x) \iff x = \theta n, \end{aligned}$$

d.h.  $x/n$  ist die einzige Nullstelle von  $L(\cdot, x)$ . Da andererseits

$$P_0(X = x) = P_1(X = x) = 0$$

gilt, so ist  $x/n$  die gesuchte Maximumstelle von  $L(\cdot, x)$ , d.h.,  $\hat{\theta}(x) = x/n$  ist der Schätzwert für  $\theta$  nach der Maximum-Likelihood-Methode. Für  $x = 0$  gilt

$$L(\theta, x) = (1 - \theta)^n = 1 \iff \theta = 0$$

und für  $x = n$  erhalten wir

$$L(\theta, x) = \theta^n = 1 \iff \theta = 1.$$

In allen möglichen Fällen erhalten wir somit, dass der Maximum-Likelihood-Schätzwert gegeben ist durch

$$\hat{\theta}(x) = \frac{x}{n}.$$

Tabelle 4 zeigt die entsprechenden Schätzwerte für die Wahrscheinlichkeiten des Auftretens der Anfangsziffern  $1, 2, \dots, 9$  an den Börsen in New York bzw. Frankfurt. Ein

Vergleich von  $\hat{\theta}_{i,1}$  mit dem jeweiligen Gegenstück  $\hat{\theta}_{i,2}$  zeigt, dass selbst bei einem Stichprobenumfang von  $n = 102$  noch eine gewisse Schwankungsbreite der Schätzwerte vorliegt. Andererseits war der Autor dieser Zeilen sehr überrascht, wie nah doch  $\hat{\theta}_{i,1}$  bzw.  $\hat{\theta}_{i,2}$  an den durch das Benford-Gesetz vorhergesagten Werten liegen. Das trifft in noch stärkerem Maße auf  $(\hat{\theta}_{i,1} + \hat{\theta}_{i,2})/2$  zu, wo die zusätzliche Mittelung die empirisch gewonnenen Werte noch näher an die Benford-Werte führte.

Aktie	\$	€	Aktie	\$	€	Aktie	\$	€
Act. Blizz.	94,43	77,72	DexCom	370,12	304,70	Monster	88,20	73,01
Adobe	472,44	389,45	DocuSign	255,15	208,05	NetEase	116,76	94,50
AMD	92,79	76,25	DollarTree	107,88	87,90	Netflix	565,17	471,40
Alexion	159,73	129,36	eBay	56,46	46,37	NVIDIA	548,50	453,95
Align Tech	534,08	438,30	Elec. Arts	146,00	119,28	NXP Sem.	172,30	142,00
Alphabet	1.901,05	1557,00	Exelon	42,54	34,96	O'Reilly	457,31	384,85
Alphabet A	1.892,56	1550,60	Facebook	274,50	226,80	Okta	264,00	214,50
Amazon	3.292,23	2709,50	Fastenal	47,72	39,45	PACCAR	98,85	81,00
Am. Elec.	82,08	66,18	Fiserv	107,28	88,75	Paychex	86,87	71,14
Amgen	253,50	207,95	Fox A	30,16	25,40	Paypal	252,00	206,20
Anal. Dev.	155,58	130,14	Fox B	29,17	24,40	Peloton	159,75	129,76
ANSYS	374,57	304,20	Gilead	66,94	55,30	Pepsi	138,59	114,58
Apple	139,07	113,68	IDEXX	492,86	403,50	Pinduoduo	171,89	138,00
Appl. Mat.	106,33	88,98	Illumina	406,25	328,60	QUALC.	162,42	133,30
ASML	569,45	464,00	Incyte	97,89	76,39	RegPharm	537,78	447,15
Atlassian	235,45	192,00	Intel	56,66	46,95	Ross	113,33	94,00
Autodesk	299,17	250,65	Intuit	374,85	310,15	Seagen	182,75	151,76
Aut. Data	161,98	132,24	Int. Surg.	744,08	619,20	Sirius	5,97	4,90
Baidu	252,75	208,50	JD.com	94,91	77,80	Skyworks	161,27	132,62
Biogen	269,44	220,95	Keurig	31,47	25,73	Splunk	170,90	141,80
Booking	2.066,24	1721,60	KLA	305,01	251,90	Starbucks	103,91	85,32
Broadcom	465,02	385,05	Kraft	32,91	26,99	Synopsis	270,14	221,70
Cadence	137,42	113,70	Lam Res.	563,85	466,05	T-Mobile	130,36	106,54
CDW Corp	135,32	111,00	lululemon	346,37	286,85	Tesla	846,64	689,20
Cerner	79,90	65,25	Marriot	122,90	103,10	Texas I.	172,81	143,34
Charter	646,69	531,00	Marvell	52,62	43,78	Trip.com	33,58	27,60
CP Softw.	129,40	104,00	Match	141,00	116,50	VeriSign	194,75	161,52
Cintas	324,48	271,00	Maxim	93,62	78,50	Verisk	192,76	158,50
Cisco	44,77	36,67	MercadoL	1965,05	1599,20	Vertech	237,59	196,80
Cognizant	80,56	65,79	Microchip	150,13	124,72	Walgreens	47,45	37,99
Comcast	48,68	40,00	Micron	82,28	68,10	Workday	235,53	191,80
Copart	114,59	94,50	Microsoft	225,95	186,76	Xcel En.	64,76	52,50
Costco	362,30	298,50	Moderna	131,02	109,06	Xilinx	142,94	121,46
CSX Corp	87,64	74,94	Mondelez	56,25	46,07	Zoom	383,40	314,00

Tabelle 2: Kurse am 22.01.2021 (Nasdaq, Frankfurt)

$i$	1	2	3	4	5	6	7	8	9
Nasdaq	35	15	13	10	11	3	2	7	6
Frankfurt	33	19	12	13	3	6	8	5	3

Tabelle 3: Häufigkeiten der Anfangsziffern

$i$	1	2	3	4	5	6	7	8	9
MLS (Nasdaq)	0,343	0,147	0,127	0,098	0,108	0,029	0,020	0,069	0,059
MLS (Frankfurt)	0,324	0,186	0,118	0,127	0,029	0,059	0,078	0,049	0,029
„Benford“-Wkt.	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046

Tabelle 4: Maximum-Likelihood-Schätzwerte der Wahrscheinlichkeiten  $\theta_{i,j}$

Bisher hatten wir uns damit begnügt, unbekannte Parameter mit Hilfe eines lediglich heuristisch motivierten Verfahrens, der Maximum-Likelihood-Methode, zu schätzen. Damit ist jedoch nicht klar, ob solche Schätzer auch tatsächlich „gut“ sind oder ob es nicht doch bessere Möglichkeiten zur Approximation des zu schätzenden Parameters gibt. Um nun „gute“ von „schlechten“ Schätzern unterscheiden zu können, braucht man zunächst ein Kriterium zur Messung der Güte. Wir führen die Diskussion zunächst anhand eines Beispiels. Passend zum bereits diskutierten Fall mit der Verteilung der Anfangsziffern bei den Werten der Nasdaq-100-Aktien nehmen wir an, dass wir Realisierungen  $x_1, \dots, x_n$  von unabhängigen und identische  $\text{Bin}(1, \theta)$ -verteilten Zufallsvariablen  $X_1, \dots, X_n$  zur Verfügung haben. (Im obigen Beispiel wäre  $n = 102$ . Wenn wir die Wahrscheinlichkeit des Auftretens von  $i$  als Anfangsziffer an der New Yorker Nasdag schätzen wollen, so wäre  $\theta_{i,1}$  der uns interessierende Parameter und  $\Theta = [0, 1]$  die Menge der grundsätzlich möglichen Werte für  $\theta_{i,1}$ .) Für eine kompakte Schreibweise definieren wir  $x := (x_1, \dots, x_n)^T$  und  $X := (X_1, \dots, X_n)^T$ . Zur Beurteilung der Güte eines Schätzverfahrens ist es naheliegend, auf die Abweichung des Schätzwertes  $\hat{\theta}(x)$  von seiner Zielgröße  $\theta$  zu schauen. So böten sich beispielsweise der Absolutbetrag  $|\hat{\theta}(x) - \theta|$  der Abweichung oder auch dessen Quadrat  $(\hat{\theta}(x) - \theta)^2$  als sinnvolle Maße für die Genauigkeit an. Dies hilft zunächst einmal nicht weiter, da die obigen Größen den unbekannt Parameter  $\theta$  enthalten und somit nicht bestimmt werden können. Darüber hinaus ist der Wert  $x$ , welchen die Zufallsvariable  $X$  annimmt, zufällig. Da wir uns für die Tauglichkeit des Schätzverfahrens interessieren, können wir zur Charakterisierung seiner Güte besser den **mittleren absoluten Fehler**  $E_\theta|\hat{\theta}(X) - \theta|$  oder auch den **mittleren quadratischen Fehler**  $E_\theta[(\hat{\theta}(X) - \theta)^2]$  heranziehen. Letzterer erweist sich bei konkreten Berechnungen als besonders bequem und wird daher überwiegend als Maß für die Güte eines Schätzverfahrens verwendet. Anzumerken ist natürlich, dass  $R(\hat{\theta}, \theta) := E_\theta[(\hat{\theta}(X) - \theta)^2]$  vom unbekannt Parameter  $\theta$  abhängt. Daher stellt sich zunächst die naheliegende Frage, ob es einen Schätzer  $\hat{\theta}$  gibt, welcher für alle möglichen Werte von  $\theta$  den geringsten mittleren quadratischen Fehler aufweist. Die Antwort wird negativ ausfallen, was die folgende kurze Argumentation zeigt: Es seien  $\theta_1, \theta_2 \in (0, 1)$  mit  $\theta_1 \neq \theta_2$  zwei mögliche Werte des Parameters  $\theta$ . Zum Beweis der Nichtexistenz eines gleichmäßig besten Schätzers betrachten wir die Funktionen  $T_1, T_2: \{0, 1\}^n \rightarrow [0, 1]$  mit  $T_1(x) := \theta_1$  und  $T_2(x) := \theta_2$  für alle  $x \in \{0, 1\}^n$ . Nun gelten für die mittleren quadratischen Fehler der entsprechenden Schätzer offensichtlich

$$R(T_1(X), \theta_1) = E_{\theta_1} \left[ \underbrace{(T_1(X) - \theta_1)^2}_{=0} \right] = 0$$

sowie

$$R(T_2(X), \theta_2) = E_{\theta_2} \left[ \underbrace{(T_2(X) - \theta_2)^2}_{=0} \right] = 0.$$

Wir nehmen nun an, dass  $\hat{\theta} = \hat{\theta}(X)$  ein **gleichmäßig bester Schätzer** für  $\theta$  ist, d.h., es muss gelten

$$R(\hat{\theta}(X), \theta) = \inf \{ R(T(X), \theta) : T(X) \text{ Schätzer für } \theta \}. \quad (9.1)$$

Daher muss insbesondere gelten, dass  $R(\hat{\theta}(X), \theta_1) \leq R(T_1(X), \theta_1) = 0$  und  $R(\hat{\theta}(X), \theta_2) \leq$

$R(T_2(X), \theta_2) = 0$ . Daraus folgt aber

$$\begin{aligned}
0 &= R(\widehat{\theta}(X), \theta_1) + R(\widehat{\theta}(X), \theta_2) \\
&= \sum_{x \in \{0,1\}^n} (\widehat{\theta}(x) - \theta_1)^2 P_{\theta_1}(X = x) + \sum_{x \in \{0,1\}^n} (\widehat{\theta}(x) - \theta_2)^2 P_{\theta_2}(X = x) \\
&\geq \sum_{x \in \{0,1\}^n} \underbrace{\left[ (\widehat{\theta}(x) - \theta_1)^2 + (\widehat{\theta}(x) - \theta_2)^2 \right]}_{>0} \underbrace{\min \{P_{\theta_1}(X = x), P_{\theta_2}(X = x)\}}_{>0} \\
&> 0.
\end{aligned} \tag{9.2}$$

Damit ist jedoch ein Widerspruch zu (9.1) erreicht und ein gleichmäßig bester Schätzer für  $\theta$  existiert somit **nicht**. Die hier gezeigte Nichtexistenz eines universell besten Verfahrens ist keinesfalls auf das speziell gewählte Binomialmodell beschränkt. Die Abschätzung in (9.2) zeigt, dass bei einem beliebigen Schätzproblem mit zwei möglichen diskreten Verteilungen  $P_{\theta_1}$  und  $P_{\theta_2}$ , welche die Eigenschaft  $\sum_x \min\{P_{\theta_1}(X = x), P_{\theta_2}(X = x)\} > 0$  haben, kein Schätzer  $\widehat{\theta}$  existiert mit  $R(\widehat{\theta}, \theta_1) = R(\widehat{\theta}, \theta_2) = 0$ . Da jedoch stets Konkurrenz durch die „entarteteten“ Schätzer  $T_1, T_2: \{0, 1\}^n \rightarrow [0, 1]$  mit  $T_1(x) := \theta_1$  und  $T_2(x) := \theta_2$  für alle  $x \in \{0, 1\}^n$  besteht, ist die Existenz eines gleichmäßig besten Schätzverfahrens ausgeschlossen. Der einzige Fall, bei dem diese Argumentation nicht greift ist der, wo die möglichen Verteilungen voneinander separiert sind, d.h., für beliebige  $\theta_1, \theta_2 \in \Theta$  mit  $\theta_1 \neq \theta_2$  gilt  $\sum_x \min\{P_{\theta_1}(X = x), P_{\theta_2}(X = x)\} = 0$ . In diesem Fall existiert natürlich ein perfekter Schätzer: Falls  $X = x$  eintritt, so wählen wir als Schätzwert  $\widehat{\theta}(x)$  dasjenige  $\theta \in \Theta$ , welches die Eigenschaft  $P_{\theta}(X = x) > 0$  besitzt. In diesem Fall braucht es jedoch keine tiefliegende Theorie, um dieses offensichtlich vernünftige Schätzverfahren zu finden. Wir können somit festhalten, dass gleichmäßig beste Schätzverfahren nur bei trivialen Schätzproblemen existieren.

Ein weit verbreiteter Ausweg aus dem Dilemma, dass ein universell bestes Verfahren nicht existiert, besteht darin, dass man die Klasse der betrachteten Schätzer auf eine sinnvolle Weise einschränkt und unter den verbleibenden Schätzern nach einem besten sucht. Die obigen Schätzer  $T_1$  und  $T_2$  mit  $T_1(x) = \theta_1$  und  $T_2(x) = \theta_2$  für alle  $x$  sind zwar unschlagbar, wenn der wahre Wert von  $\theta$  gleich  $\theta_1$  bzw.  $\theta_2$  ist. Für andere Werte von  $\theta$  sind entsprechende Schätzerfolgen nicht einmal konsistent, falls der Stichprobenumfang  $n$  gegen unendlich strebt. Es erscheint durchaus akzeptabel, derartige Schätzer von vornherein aus der Menge der infrage kommenden Schätzer auszuschließen. Die folgende Einschränkung führt bei einer Reihe von Schätzproblemen bereits zu einer Situation, wo dann ein universell bestes Verfahren gefunden werden kann.

**Definition 9.4.** *Es möge die Realisierung  $x$  einer Zufallsvariable  $X$  mit einer Verteilung  $P_{\theta}$  vorliegen, wobei  $\theta \in \Theta$  gilt.*

*Ein Schätzer  $\widehat{\theta} = \widehat{\theta}(X)$  heißt **erwartungstreu für den Parameter  $\theta$** , falls*

$$E_{\theta}[\widehat{\theta}(X)] = \theta \quad \forall \theta \in \Theta$$

*gilt.*

Falls eine von  $\theta$  abgeleitete Größe  $g(\theta)$  geschätzt werden soll, so heißt ein Schätzer  $\widehat{g} = \widehat{g}(X)$  erwartungstreu für  $g(\theta)$ , falls

$$E_{\theta}[\widehat{g}(X)] = g(\theta) \quad \forall \theta \in \Theta$$

gilt.

Wir betrachten nochmals das Binomialbeispiel, d.h., wir setzen voraus, dass die vorliegenden Daten aus Realisierungen  $x_1, \dots, x_n$  von unabhängigen und identisch verteilten Zufallsvariablen  $X_1, \dots, X_n$  bestehen, wobei  $X_i \sim \text{Bin}(1, \theta)$ . Wir setzen der Einfachheit halber voraus, dass  $\theta \in \Theta := (0, 1)$ , d.h., die Fälle  $\theta = 0, 1$  schließen wir aus. Zu schätzen sei  $\theta$  und als Gütekriterium wählen wir den mittleren quadratischen Fehler. Es sei  $X = (X_1, \dots, X_n)^T$ . Im Folgenden zeigen wir, dass  $\bar{X}_n$  der eindeutig bestimmte **beste erwartungstreue Schätzer** für  $\theta \in \Theta$  ist.

$$R(\bar{X}_n, \theta) = \inf \{ R(T(X), \theta) : T(X) \text{ erwartungstreuer Schätzer für } \theta \}. \quad (9.3)$$

Es sei also  $T(X)$  mit  $T: \{0, 1\}^n \rightarrow [0, 1]$  ein beliebiger erwartungstreuer Schätzer für  $\theta$ . Wir gehen in zwei Schritten vor:

- 1) Wir betrachten einen weiteren Schätzer  $\bar{T}(X)$ , wobei

$$\bar{T}(x) = h\left(\sum_{i=1}^n x_i\right)$$

und, für  $k = 0, 1, \dots, n$ ,

$$h(k) := \frac{1}{\binom{n}{k}} \sum_{x: \sum_{i=1}^n x_i = k} T(x).$$

$\bar{T}(x)$  ist eine „symmetrisierte“ Version von  $T(x)$ , denn es gilt

$$\bar{T}((x_1, \dots, x_n)^T) = \bar{T}((x_{\pi(1)}, \dots, x_{\pi(n)})^T) \quad \forall x \in \{0, 1\}^n, \forall \pi \in \mathcal{P}_n, \quad (9.4)$$

wobei  $\mathcal{P}_n$  die Menge aller Permutationen  $\pi: \{1, \dots, n\} \rightarrow \{1, \dots, n\}$  von  $n$  Elementen bezeichnet.

Wir zeigen, dass  $\bar{T}(X)$  ebenfalls erwartungstreu für  $\theta$  ist und, falls (9.4) durch  $T(x)$  verletzt ist, so folgt

$$R(\bar{T}(X), \theta) < R(T(X), \theta) \quad \forall \theta \in \Theta. \quad (9.5)$$

- 2)  $\bar{X}_n$  ist der **einzige** erwartungstreu Schätzer für  $\theta$ , der die Symmetriebedingung (9.4) aus 1) erfüllt.

**Details:**

zu 1)  $\bar{T}(x)$  kann nur die Werte  $h(0), \dots, h(n)$  annehmen, wobei  $\bar{T}(x) = h(k)$  gilt, falls  $\sum_{i=1}^n x_i = k$  ist. Daraus folgt

$$\begin{aligned}
E_\theta \bar{T}(X) &= \sum_{k=0}^n h(k) P_\theta \left( \sum_{i=1}^n X_i = k \right) \\
&= \sum_{k=0}^n \frac{1}{\binom{n}{k}} \sum_{x: \sum_{i=1}^n x_i = k} T(x) \binom{n}{k} \theta^k (1-\theta)^{n-k} \\
&= \sum_{k=0}^n \sum_{x: \sum_{i=1}^n x_i = k} T(x) \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} \\
&= \sum_{x \in \{0,1\}^n} T(x) P_\theta(X = x) = E_\theta T(X).
\end{aligned}$$

Da jedoch nach Voraussetzung  $E_\theta T(X) = \theta$  für all  $\theta \in \Theta$  gilt, so folgt auch die Erwartungstreue von  $\bar{T}(X)$ .

Wir nehmen nun an, dass  $T$  (9.4) verletzt. Dann gilt

$$\begin{aligned}
R(T(X), \theta) &= \sum_{x \in \{0,1\}^n} (T(x) - \theta)^2 P_\theta(X = x) \\
&= \sum_{k=0}^n \sum_{x: \sum_{i=1}^n x_i = k} (T(x) - \bar{T}(x) + \bar{T}(x) - \theta)^2 P_\theta(X = x) \\
&= \sum_{k=0}^n \sum_{x: \sum_{i=1}^n x_i = k} (\bar{T}(x) - \theta)^2 P_\theta(X = x) \\
&\quad + \sum_{k=0}^n \sum_{x: \sum_{i=1}^n x_i = k} \underbrace{(T(x) - \bar{T}(x))^2}_{>0 \text{ für ein } x} \underbrace{P_\theta(X = x)}_{>0} \\
&\quad + 2 \underbrace{\sum_{k=0}^n \sum_{x: \sum_{i=1}^n x_i = k} (T(x) - \bar{T}(x)) (\bar{T}(x) - \theta)^2 P_\theta(X = x)}_{=0} \\
&> \sum_{x \in \{0,1\}^n} (\bar{T}(x) - \theta)^2 P_\theta(X = x) = R(\bar{T}(X), \theta),
\end{aligned}$$

d.h.,  $\bar{T}(X)$  verbessert  $T(X)$ .

zu 2) Es sei nun  $\tilde{T}(X)$  ein beliebiger erwartungstreuer Schätzer für  $\theta$ , welcher die Symmetriebedingung (9.4) erfüllt. Dann existiert  $\tilde{h}: \{0, 1, \dots, n\} \rightarrow [0, 1]$  mit  $\tilde{T}(x) =$

$\tilde{h}(\sum_{i=1}^n x_i)$  für alle  $x \in \{0, 1\}^n$ . Nun gilt

$$\begin{aligned} E_\theta[\bar{X}_n] - E_\theta[\tilde{T}(X)] &= \sum_{k=0}^n [k/n - \tilde{h}(k)] \binom{n}{k} \theta^k (1-\theta)^{n-k} \\ &= (1-\theta)^n \sum_{k=0}^n [k/n - \tilde{h}(k)] \binom{n}{k} \left(\frac{\theta}{1-\theta}\right)^k = 0 \quad \forall \theta \in (0, 1). \end{aligned}$$

Das ist äquivalent zu

$$\sum_{k=0}^n [k/n - \tilde{h}(k)] \binom{n}{k} \rho^k = 0 \quad \forall \rho \in (0, \infty),$$

woraus  $[k/n - \tilde{h}(k)] \binom{n}{k} = 0$  für alle  $k = 0, \dots, n$  folgt. Also gilt  $\tilde{h}(k) = k/n$  und somit  $\tilde{T}(x) = \bar{x}_n$  für alle  $x \in \{0, 1\}^n$ . Daher ist  $\bar{X}_n$  der einzige erwartungstreue Schätzer für  $\theta$ , der (9.4) erfüllt.

## Das Testen statistischer Hypothesen

Im letzten Teil des Kurses wollen wir noch die wichtigsten Konzepte aus dem Bereich des Testens von Hypothesen einführen. Wir nehmen an, dass wir als Datengrundlage  $x_1, \dots, x_n$  haben, welche wir als Realisierungen entsprechender Zufallsvariablen  $X_1, \dots, X_n$  auffassen. Wir nehmen an, dass die Verteilung der  $X_i$  einer bestimmten Menge von Verteilungen angehört und wollen diese Annahme unter Ausnutzung der Information, welche uns  $x_1, \dots, x_n$  liefern, überprüfen. Bevor wir uns dieser Fragestellung systematisch nähern, betrachten wir ein einfaches Beispiel.

Es wurde bereits der Frage nachgegangen, inwieweit Börsenkurse dem Benford-Gesetz folgen. Dabei wurden unter anderem die Schlusskurse der 102 an der Nasdaq in New York gehandelten Werte aus dem Nasdaq-100-Index herangezogen und es wurden damit die Wahrscheinlichkeiten geschätzt, dass die Anfangsziffern dieser zufälligen Kurse die jeweiligen Werte  $1, 2, \dots, 9$  annehmen. Tabelle 4 auf Seite 92 dieses Skriptes enthält die so gewonnenen Schätzwerte  $\hat{\theta}_{1,1} = 0,343, \dots, \hat{\theta}_{9,1} = 0,059$  für die jeweiligen unbekanntenen Wahrscheinlichkeiten  $\theta_{1,1}, \dots, \theta_{9,1}$ . Ein Vergleich mit den in der letzten Zeile dieser Tabelle aufgelisteten Wahrscheinlichkeiten nach dem Benford-Gesetz zeigt einerseits eine überraschend gute Übereinstimmung mit diesen Werten. Andererseits kann es auch sein, dass das Benford-Gesetz für zufällige Daten dieser Art gar nicht gilt und uns nur der Zufall einen Streich gespielt hat. Beispielsweise könnte man „aus Symmetriegründen“ annehmen, dass jede der möglichen Anfangsziffern mit einer Wahrscheinlichkeit von  $1/9$  vorkommt. (Weitere Alternativen sind sicherlich auch denkbar.) Zur Vereinfachung richten wir hier den Fokus auf die Wahrscheinlichkeit des Auftretens der Eins als erste signifikante Ziffer. Wir nehmen an, dass die in Tabelle 3 angegebene Zahl von  $x = 35$  die Realisierung einer  $\text{Bin}(102, \theta)$ -verteilten Zufallsvariable  $X$  ist. Wir wollen die Hypothese

überprüfen, dass  $X$  eine  $\text{Bin}(102, \theta_0)$ -Verteilung besitzt, wobei  $\theta_0 = 0,301$  der durch das Benford-Gesetz vorgeschriebene Wert für diese Wahrscheinlichkeit ist. Als Alternative steht die Vermutung, dass  $\theta = \theta_1 = 1/9$  ist. Damit ergibt sich folgendes Testproblem:

$$H_0: \theta = \theta_0 \quad \text{gegen} \quad H_1: \theta = \theta_1.$$

$H_0$  ist hierbei unsere **Hypothese („Nullhypothese“)** und  $H_1$  die **Alternative („Alternativhypothese“)**. Wenn man lediglich das Benford-Gesetz anzweifelt und sich nicht auf einen konkreten Alternativwert  $\theta_1$  festlegen will, so könnte das Testproblem auch folgendermaßen aussehen:

$$H_0: \theta = \theta_0 \quad \text{gegen} \quad H'_1: \theta \neq \theta_0.$$

In der Fachsprache sind  $H_0$  und  $H_1$  sogenannte **einfache Hypothesen**, da unter diesen Annahmen der das Modell spezifizierende Parameter  $\theta$  jeweils nur einen einzigen Wert annehmen kann. Unter  $H'_1$  kann dagegen  $\theta$  alle Werte aus der Menge  $[0, 1] \setminus \{\theta_0\}$  annehmen und man spricht von einer sogenannten **zusammengesetzten Hypothese**.

Wenden wir uns zunächst dem Problem des Testens von  $H_0$  gegen  $H_1$  zu. Da  $\theta_0 > \theta_1$  ist und eine  $\text{Bin}(102, \theta_0)$ -verteilte Zufallsvariable tendenziell größere Werte als eine  $\text{Bin}(102, \theta_1)$ -verteilte annimmt, erscheint es sinnvoll,  $H_0$  abzulehnen, falls die Realisierung  $x$  der Zufallsvariable  $X$  einen kleinen Wert angenommen hat. Aber welcher Wert ist nun klein genug, sodass man sich gegen  $H_0$  entscheiden sollte? Wenn man die Entscheidung über Annahme bzw. Ablehnung von  $H_0$  nicht komplett aus dem Bauch heraus treffen möchte, so kann man sich auf eine von dem polnischen Mathematiker Jerzy Neyman und dem britischen Statistiker Egon Sharpe Pearson in den dreißiger Jahren des vorigen Jahrhunderts vorgeschlagene Strategie stützen. In der vorliegenden Situation können aus zweierlei Art falsche Entscheidungen getroffen werden. Wenn man die Nullhypothese  $H_0$  ablehnt obwohl sie richtig ist, begeht man einen **Fehler erster Art**. Ist dagegen die Alternativhypothese  $H_1$  korrekt und man nimmt stattdessen die Nullhypothese an, so spricht man von einem **Fehler zweiter Art**. Nun ist man ausschließlich daran interessiert, die richtige Entscheidung zu treffen. Da die durch die Realisierung  $x$  gelieferte Information zufallsbehaftet ist, versucht man nun, eine solche Entscheidungsregel zu finden, sodass die Wahrscheinlichkeiten einer Fehlentscheidung möglichst klein sind. Eine solche Entscheidungsregel beschreibt man am einfachsten durch eine Funktion  $\varphi: \Omega_X \rightarrow \{0, 1\}$ , wobei  $\Omega_X$  die Menge der möglichen Werte der Zufallsvariable  $X$  ist (hier ist  $\Omega_X = \{0, 1, \dots, 102\}$ ) und  $\varphi(x) = 1$  bedeutet, dass im Falle von  $X = x$  die Nullhypothese  $H_0$  abgelehnt wird, wogegen  $\varphi(x) = 0$  bedeutet, dass bei Eintreten des Ereignisses  $X = x$   $H_0$  angenommen wird. Eine solche Funktion  $\varphi$  nennt man kurz **Test**. Ideal wäre nun ein solcher Test, für den sowohl die Wahrscheinlichkeit eines Fehlers erster Art als auch die Wahrscheinlichkeit eines Fehlers zweiter Art minimal sind. Bei dem betrachteten Testproblem ist es möglich, jede einzelne der beiden Irrtumswahrscheinlichkeiten auf Null zu drücken. Ein Test  $\varphi_0$ , der niemals ablehnt ( $\varphi_0(x) = 0$  für alle  $x = 0, 1, \dots, 102$ ) hat die Eigenschaft, dass die Wahrscheinlichkeit eines Fehlers erster Art Null ist (also  $P_{\theta_0}(\varphi_0(X) = 1) = 0$ ). Umgekehrt, ein Test  $\varphi_1$ , der stets ablehnt ( $\varphi_1(x) = 1$  für alle  $x = 0, 1, \dots, 102$ ) besitzt die Eigenschaft, dass die Wahrscheinlichkeit eines Fehlers zweiter Art Null ist ( $P_{\theta_1}(\varphi_1(X) = 0) = 0$ ). So wie es üblicherweise keine gleichmäßig besten Schätzverfahren gibt, existiert hier auch

kein universell bestes Testverfahren, bei dem beide Irrtumswahrscheinlichkeiten simultan minimiert werden. Sei dazu  $\varphi: \{0, 1, \dots, 102\} \rightarrow \{0, 1\}$  ein beliebiger Test. Dann gilt für die Summe beider Irrtumswahrscheinlichkeiten

$$\begin{aligned}
& P_{\theta_0}(\varphi(X) = 1) + P_{\theta_1}(\varphi(X) = 0) \\
&= \sum_{x: \varphi(x)=1} \binom{n}{x} \theta_0^x (1 - \theta_0)^{n-x} + \sum_{x: \varphi(x)=0} \binom{n}{x} \theta_1^x (1 - \theta_1)^{n-x} \\
&= \sum_{x=0}^{102} \left\{ \varphi(x) \binom{n}{x} \theta_0^x (1 - \theta_0)^{n-x} + (1 - \varphi(x)) \binom{n}{x} \theta_1^x (1 - \theta_1)^{n-x} \right\} \\
&\geq \sum_{x=0}^{102} \underbrace{[\varphi(x) + (1 - \varphi(x))]}_{=1} \underbrace{\min \left\{ \binom{n}{x} \theta_0^x (1 - \theta_0)^{n-x}, \binom{n}{x} \theta_1^x (1 - \theta_1)^{n-x} \right\}}_{>0} > 0.
\end{aligned}$$

D.h., es gibt keinen Test, für den beide Irrtumswahrscheinlichkeiten Null sind. Angesichts dieser Tatsache schlugen Neyman und Pearson vor, nach einem geeigneten Kompromiss zwischen den beiden Irrtumswahrscheinlichkeiten zu suchen. Im vorliegenden Fall liegt der Nullhypothese das inzwischen durch viele Publikationen gestützte Benford-Gesetz zugrunde. Daher ist es wünschenswert, dass im Falle ihrer Korrektheit eine Ablehnung nur mit einer sehr begrenzten Wahrscheinlichkeit vorkommt. Man wählt also ein üblicherweise kleines  $\alpha$  ( $\alpha = 0,05$  oder  $\alpha = 0,01$  werden in der Praxis häufig gewählt) und lässt nur Tests  $\varphi$  zu, deren jeweilige Wahrscheinlichkeiten für einen Fehler erster Art den Wert  $\alpha$  nicht überschreiten. Ein solches Vorgehen sichert den Anwender dahingehend ab, dass er im Falle der tatsächlichen Gültigkeit des Benford-Gesetzes dessen Widerlegung lediglich mit einer Wahrscheinlichkeit, die  $\alpha$  nicht übersteigt, verkündet. Wenn es dagegen zur Ablehnung von  $H_0$  kommt, so kann man das als einen Beweis gegen die Gültigkeit von  $H_0$  ansehen. In der Klasse der Tests welche die Eigenschaft besitzen, dass die deren Wahrscheinlichkeiten eines Fehlers erster Art das gewählte **Niveau**  $\alpha$  nicht überschreiten, wählt man dann einen der Tests aus, deren Wahrscheinlichkeiten eines Fehlers zweiter Art minimal sind. Diese Vorgehensweise wird durch die folgende Definition nochmals zusammengefasst.

**Definition 9.5.** *Es möge eine Realisierung  $x$  ( $x \in \mathbb{R}$  oder auch  $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$ ) einer Zufallsvariable  $X$  mit möglichen Werten in  $\Omega_X$  vorliegen, wobei  $X \sim P_\theta$ ,  $\theta \in \Theta$ . Ein Test  $\varphi: \Omega_X \rightarrow \{0, 1\}$  für das Problem*

$$H_0: \theta \in \Theta_0 \quad \text{gegen} \quad H_1: \theta \in \Theta_1,$$

wobei  $\Theta_0 \cap \Theta_1 = \emptyset$  und  $\Theta_0 \cup \Theta_1 = \Theta$ , heißt

- **Test zum Niveau  $\alpha$  ( $\alpha$ -Test), falls**

$$\sup_{\theta \in \Theta_0} P_\theta(\varphi(X) = 1) \leq \alpha,$$

- **gleichmäßig bester  $\alpha$ -Test**, falls  $\varphi$  ein  $\alpha$ -Test ist und

$$P_{\theta}(\varphi(X) = 0) = \inf \{P_{\theta}(\tilde{\varphi}(X) = 0) : \tilde{\varphi} \text{ ist } \alpha\text{-Test}\} \quad \forall \theta \in \Theta_1$$

*gilt.*

Im Spezialfall einer einfachen Alternative  $\Theta_1 = \{\theta_1\}$  spricht man von einem **besten  $\alpha$ -Test**.

Beim oben diskutierten Beispiel haben wir das Testproblem

$$H_0: \theta = \theta_0 \quad \text{gegen} \quad H_1: \theta = \theta_1,$$

wobei  $\theta_0 = 0,301$  und  $\theta_1 = 1/9$  sind. Beobachtet wird eine Realisierung einer Zufallsvariable  $X \sim \text{Bin}(102, \theta)$ . Wir wollen uns hier zunächst auf Tests  $\varphi$  beschränken, welche die Gestalt

$$\varphi(x) = \begin{cases} 1, & \text{falls } x \leq c, \\ 0, & \text{falls } x > c \end{cases} \quad (9.6)$$

besitzen. Ein solcher Test hat eine Irrtumswahrscheinlichkeit erster Art von

$$P_{\theta_0}(\varphi(X) = 1) = P_{\theta_0}(X \leq c) = \sum_{x=0}^c \binom{n}{x} \theta_0^x (1 - \theta_0)^{n-x} \quad (9.7a)$$

und eine Irrtumswahrscheinlichkeit zweiter Art von

$$P_{\theta_1}(\varphi(X) = 0) = P_{\theta_1}(X > c) = \sum_{x=c+1}^n \binom{n}{x} \theta_1^x (1 - \theta_1)^{n-x}. \quad (9.7b)$$

Die Schwelle  $c \in \mathbb{N}_0$ , welche den Ablehn- vom Annahmehereich trennt, heißt **kritischer Wert**. Nachdem man sich für ein Niveau  $\alpha$  entschieden hat, kann man nach einem besten  $\alpha$ -Test suchen. (9.7b) besagt, dass fallendes  $c$  zu einer wachsenden Irrtumswahrscheinlichkeit zweiter Art führt, wohingegen (9.7a) den umgekehrten Effekt für die Irrtumswahrscheinlichkeit erster Art zeigt. Wenn wir uns beispielsweise ein Niveau von  $\alpha = 0,05$  vorgeben, dann ist der beste  $\alpha$ -Test innerhalb der Tests mit der Gestalt (9.6) durch den größtmöglichen kritischen Wert  $c_{\alpha}$  gegeben, so dass der entsprechende Test noch das Niveau  $\alpha = 0,05$  einhält, d.h.

$$c_{\alpha} = \max \{c: P_{\theta_0}(X \leq c) \leq \alpha\}.$$

An dieser Stelle ist nun eine exakte Bestimmung des gesuchten kritischen Wertes  $c_{\alpha}$  recht mühsam, da die entsprechenden Binomialwahrscheinlichkeiten Fakultäten großer Zahlen beinhalten. Wie werden daher die Normalapproximation der Binomialverteilung aus Satz 8.7 (Globaler Grenzwertsatz von de Moivre-Laplace) verwenden. Nach diesem Satz gilt

$$P_{\theta_0} \left( \frac{X - n\theta_0}{\sqrt{n\theta_0(1 - \theta_0)}} \leq c \right) \approx \Phi(c),$$

wobei  $\Phi$  die Verteilungsfunktion einer Standardnormalverteilung ist. Wegen  $\Phi(-1,64) \approx 0,05$  erhalten wir

$$P_{\theta_0}\left(X \leq n\theta_0 - 1,64\sqrt{n\theta_0(1-\theta_0)}\right) \approx 0,05.$$

Der gesuchte kritische Wert ist somit

$$c_\alpha = 102 \cdot 0,301 - 1,64\sqrt{102 \cdot 0,301 \cdot 0,699} \approx 23,$$

d.h., wir lehnen die dem Benford-Gesetz entsprechende Nullhypothese ab, wenn die Zufallsvariable  $X$  einen Wert  $x \leq 23$  annimmt. Der gesuchte Test  $\varphi$  hat die Gestalt

$$\varphi(x) = \begin{cases} 1, & \text{falls } x \leq 23, \\ 0, & \text{falls } x > 23 \end{cases}$$

und besitzt die Eigenschaft

$$P_{\theta_0}(\varphi(X) = 1) = P_{\theta_0}(X \leq 23) \approx 0,05.$$

Im vorliegenden Fall hatte die Zufallsvariable den Wert  $x = 35$  angenommen. Wir können also auf Grund unserer Daten die Hypothese  $H_0: \theta = 0,301$  **nicht ablehnen**.

## Optimale Tests

Nachdem die grundlegenden Konzepte zum Testen von Hypothesen eingeführt wurden, werden wir uns nun der Suche nach optimalen Testverfahren zuwenden. Dabei sei noch einmal an die bereits auf den Seiten 99/100 gegebene Definition erinnert:

**Definition 9.6.** *Es möge eine Realisierung  $x$  ( $x \in \mathbb{R}$  oder auch  $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$ ) einer Zufallsvariable  $X$  mit möglichen Werten in  $\Omega_X$  vorliegen, wobei  $X \sim P_\theta$ ,  $\theta \in \Theta$ . Ein Test  $\varphi: \Omega_X \rightarrow \{0, 1\}$  für das Problem*

$$H_0: \theta \in \Theta_0 \quad \text{gegen} \quad H_1: \theta \in \Theta_1,$$

wobei  $\Theta_0 \cap \Theta_1 = \emptyset$  und  $\Theta_0 \cup \Theta_1 = \Theta$ , heißt

- **Test zum Niveau  $\alpha$  ( $\alpha$ -Test), falls**

$$\sup_{\theta \in \Theta_0} P_\theta(\varphi(X) = 1) \leq \alpha,$$

- **gleichmäßig bester  $\alpha$ -Test, falls  $\varphi$  ein  $\alpha$ -Test ist und**

$$P_\theta(\varphi(X) = 0) = \inf \{P_\theta(\tilde{\varphi}(X) = 0) : \tilde{\varphi} \text{ ist } \alpha\text{-Test}\} \quad \forall \theta \in \Theta_1$$

*gilt.*

*Im Spezialfall einer einfachen Alternative  $\Theta_1 = \{\theta_1\}$  spricht man von einem **besten  $\alpha$ -Test**.*

Um eventuelle technische Schwierigkeiten zu vermeiden, beschränken wir uns auf den Fall, dass die Zufallsvariable  $X$  **diskret** ist. Zur weiteren Vereinfachung betrachten wir hier nur das Problem des Testens **einfacher Hypothesen** (d.h.  $\Theta_0 = \{\theta_0\}$  und  $\Theta_1 = \{\theta_1\}$ ).

Wir nehmen also an, dass die Realisierung  $x$  einer diskreten Zufallsvariable  $X \sim P_\theta$  mit Werten in  $\Omega_X$  beobachtet werde, wobei  $\theta \in \{\theta_0, \theta_1\}$ ,  $\theta_0 \neq \theta_1$  ist. Das Testproblem sei

$$H_0: \theta = \theta_0 \quad \text{gegen} \quad H_1: \theta = \theta_1.$$

Bevor wir uns nun auf die Suche nach einem besten  $\alpha$ -Test begeben, sei noch einmal an die jeweiligen Irrtumswahrscheinlichkeiten erinnert. Für einen Test  $\varphi: \Omega_X \rightarrow \{0, 1\}$  ist die Wahrscheinlichkeit eines Fehlers erster Art ist gegeben durch

$$P_{\theta_0}(\varphi(X) = 1) = \sum_{x: \varphi(x)=1} P_{\theta_0}(X = x) = \sum_{x \in \Omega_X} \varphi(x) P_{\theta_0}(X = x)$$

und die eines Fehlers zweiter Art durch

$$P_{\theta_1}(\varphi(X) = 0) = \sum_{x: \varphi(x)=0} P_{\theta_1}(X = x) = \sum_{x \in \Omega_X} (1 - \varphi(x)) P_{\theta_1}(X = x).$$

Damit deutet sich an, dass ein bester Test zu einem vorgegebenen Niveau  $\alpha$  so sein sollte, dass  $H_0$  bei Eintreten des Ereignisses  $X = x$  dann abgelehnt wird, wenn  $P_{\theta_1}(X = x)$  im Verhältnis zu  $P_{\theta_0}(X = x)$  groß ist. Das folgende Lemma präzisiert diese Aussage.

**Lemma 9.7.**  $X$  sei eine diskrete Zufallsvariable mit Werten in  $\Omega_X$ . Ein Test  $\varphi: \Omega_X \rightarrow \{0, 1\}$  für das Testproblem

$$H_0: \theta = \theta_0 \quad \text{gegen} \quad H_1: \theta = \theta_1.$$

möge die Struktur

$$\varphi(x) = \begin{cases} 1, & \text{falls } P_{\theta_1}(X = x) > c P_{\theta_0}(X = x), \\ 0, & \text{falls } P_{\theta_1}(X = x) < c P_{\theta_0}(X = x) \end{cases}$$

besitzen, wobei  $c \in [0, \infty)$ .

(Für alle  $x$  mit  $P_{\theta_1}(X = x) = c P_{\theta_0}(X = x)$  kann  $\varphi(x)$  sowohl den Wert 1 als auch den Wert 0 annehmen.)

Dann gilt: Falls  $\varphi^*: \Omega_X \rightarrow \{0, 1\}$  ein weiterer Test ist mit

$$P_{\theta_0}(\varphi^*(X) = 1) \leq P_{\theta_0}(\varphi(X) = 1),$$

so gilt

$$P_{\theta_1}(\varphi^*(X) = 0) \geq P_{\theta_1}(\varphi(X) = 0).$$

Folglich ist  $\varphi$  ein bester  $\bar{\alpha}$ -Test zum Niveau  $\bar{\alpha} := P_{\theta_0}(\varphi(X) = 1)$ .

*Beweis.* Der Beweis dieser Aussage beruht auf folgender Ungleichung.

$$(\varphi(x) - \varphi^*(x)) c P_{\theta_0}(X = x) \leq (\varphi(x) - \varphi^*(x)) P_{\theta_1}(X = x) \quad \forall x \in \Omega_X. \quad (9.8)$$

Der Beweis dieser Ungleichung folgt aus einer einfachen Fallunterscheidung.

- Falls  $c P_{\theta_0}(X = x) < P_{\theta_1}(X = x)$  ist, so gilt wegen  $\varphi(x) = 1$ , dass  $(\varphi(x) - \varphi^*(x)) \geq 0$ .
- Falls  $c P_{\theta_0}(X = x) > P_{\theta_1}(X = x)$  ist, so gilt wegen  $\varphi(x) = 0$ , dass  $(\varphi(x) - \varphi^*(x)) \leq 0$ .
- Falls  $c P_{\theta_0}(X = x) = P_{\theta_1}(X = x)$  ist, so gilt in (9.8) sogar “=”.

In allen drei Fällen folgt also die Gültigkeit von (9.8).

Nun folgt der zentrale Schritt des Beweises. Wegen  $P_{\theta_0}(\varphi^*(X) = 1) \leq P_{\theta_0}(\varphi(X) = 1)$  und  $c \geq 0$  erhalten wir

$$\begin{aligned} 0 &\leq c \left\{ P_{\theta_0}(\varphi(X) = 1) - P_{\theta_0}(\varphi^*(X) = 1) \right\} \\ &= \sum_{x \in \Omega_X} \underbrace{(\varphi(x) - \varphi^*(x)) c P_{\theta_0}(X = x)}_{\leq (\varphi(x) - \varphi^*(x)) P_{\theta_1}(X = x)} \\ &\leq \sum_{x \in \Omega_X} (\varphi(x) - \varphi^*(x)) P_{\theta_1}(X = x) \\ &= P_{\theta_1}(\varphi(X) = 1) - P_{\theta_1}(\varphi^*(X) = 1), \end{aligned}$$

d.h., die **Macht** von  $\varphi$  gegen die Alternative  $\theta_1$  ist nicht kleiner als die des Tests  $\varphi^*$ . Daraus folgt

$$P_{\theta_1}(\varphi(X) = 0) = \left(1 - P_{\theta_1}(\varphi(X) = 1)\right) \leq \left(1 - P_{\theta_1}(\varphi^*(X) = 1)\right) = P_{\theta_1}(\varphi^*(X) = 0),$$

d.h., die Wahrscheinlichkeit eines Fehlers zweiter Art durch den Test  $\varphi$  ist kleiner oder gleich der Wahrscheinlichkeit eines Fehlers zweiter Art durch  $\varphi^*$ .  $\square$

Wir kommen zurück auf das Beispiel mit der Häufigkeit der 1 als erster signifikanter Ziffer bei den New Yorker Kursen der 102 Aktien im Nasdaq-100-Index. Wir hatten  $x = 35$  beobachtet und dieses  $x$  als Realisierung einer Zufallsvariable  $X \sim \text{Bin}(n, \theta)$  modelliert, wobei  $n = 102$  ist,  $\theta = \theta_0 := 0,301$  dem Benford-Gesetz entspricht und  $\theta = \theta_1 := 1/9$  unsere Alternativhypothese war. Angenommen, wir suchen nach einem besten  $\alpha$ -Test für

$$H_0: \theta = \theta_0 \quad \text{gegen} \quad H_1: \theta = \theta_1,$$

so können wir versuchen, einen solchen mit Hilfe von Lemma 9.7 zu finden. Die Zufallsvariable  $X$  kann Werte aus  $\Omega_X = \{0, 1, \dots, 102\}$  annehmen und Lemma 9.7 legt einen Test  $\varphi: \Omega_X \rightarrow \{0, 1\}$  mit folgender Struktur nahe:

$$\varphi(x) = \begin{cases} 1, & \text{falls } P_{\theta_1}(X = x) > c P_{\theta_0}(X = x), \\ 0, & \text{falls } P_{\theta_1}(X = x) < c P_{\theta_0}(X = x) \end{cases}$$

Da wir  $\varphi(x)$  auch für den Fall  $P_{\theta_1}(X = x) = c P_{\theta_0}(X = x)$  irgendwie festlegen müssen, betrachten wir zunächst alle Tests der Art

$$\varphi(x) = \begin{cases} 1, & \text{falls } P_{\theta_1}(X = x) \geq c P_{\theta_0}(X = x), \\ 0, & \text{falls } P_{\theta_1}(X = x) < c P_{\theta_0}(X = x) \end{cases} \quad (9.9)$$

Nun gilt

$$\frac{P_{\theta_1}(X = x)}{P_{\theta_0}(X = x)} = \frac{\binom{n}{x} \theta_1^x (1 - \theta_1)^{n-x}}{\binom{n}{x} \theta_0^x (1 - \theta_0)^{n-x}} = \left( \frac{1 - \theta_1}{1 - \theta_0} \right)^n \underbrace{\left( \frac{\theta_1 (1 - \theta_0)}{\theta_0 (1 - \theta_1)} \right)^x}_{\in (0,1)}.$$

Demzufolge ist die Funktion  $x \mapsto P_{\theta_1}(X = x)/P_{\theta_0}(X = x)$  streng monoton fallend und die durch (9.9) beschriebenen Tests können so dargestellt werden:

$$\varphi^{(d)}(x) = \begin{cases} 1, & \text{falls } x \leq d, \\ 0, & \text{falls } x > d, \end{cases}$$

für ein  $d \in \mathbb{Z}$ . Damit das geforderte Niveau  $\alpha$  eingehalten wird, muss gelten

$$P_{\theta_0}(\varphi^{(d)}(X) = 1) = \sum_{k=0}^d \binom{n}{k} \theta_0^k (1 - \theta_0)^{n-k} \leq \alpha. \quad (9.10)$$

Die Wahrscheinlichkeit eines Fehlers zweiter Art ist gleich

$$P_{\theta_1}(\varphi^{(d)}(X) = 0) = \sum_{k=d+1}^n \binom{n}{k} \theta_1^k (1 - \theta_1)^{n-k}.$$

Die optimale Wahl des kritischen Wertes  $d$ , welcher die primäre Forderung (9.10) erfüllt, ist gegeben durch

$$d_\alpha = \max \left\{ d: \sum_{k=0}^d \binom{n}{k} \theta_0^k (1 - \theta_0)^{n-k} \leq \alpha \right\}.$$

Falls nun

$$P_{\theta_1}(\varphi^{(d_\alpha)}(X) = 0) = \alpha,$$

so ist  $\varphi^{(d_\alpha)}$  nach Lemma 9.7 der gesuchte beste  $\alpha$ -Test für  $H_0$  gegen  $H_1$ . Falls dagegen

$$P_{\theta_1}(\varphi^{(d_\alpha)}(X) = 0) < \alpha,$$

so greift Lemma 9.7 nicht und es ist nicht klar, ob der Test  $\varphi^{(d_\alpha)}$  ein bester  $\alpha$ -Test ist.

Nun könnte man versuchen, dieses Problem mit der „Brute-Force-Methode“ (Methode der rohen Gewalt) zu lösen, indem man **alle** Tests  $\varphi: \Omega_X \rightarrow \{0, 1\}$  betrachtet, davon jene aussortiert, die das Niveau  $\alpha$  nicht einhalten, und aus der verbliebenen Menge von Tests schließlich den mit der kleinsten Irrtumswahrscheinlichkeit zweiter Art auswählt. Im vorliegenden Beispiel mit den Nasdaq-Kursen ist  $\Omega_X = \{0, 1, \dots, 102\}$  und die Anzahl aller Tests ist somit gleich  $2^{103}$ , also ungefähr  $10^{31}$ . Damit wird klar, dass eine Betrachtung aller Tests praktisch unmöglich ist. Ein vielleicht überraschender Ausweg wird durch sogenannte **randomisierte Tests**  $\varphi: \Omega_X \rightarrow [0, 1]$  geboten, wobei bei Eintreten des Ereignisses  $X = x$  die Nullhypothese mit einer Wahrscheinlichkeit von  $\varphi(x)$  abgelehnt wird. Zur praktischen Durchführung müsste dann ein weiteres unabhängiges Zufallsexperiment mit einer Erfolgswahrscheinlichkeit von  $\varphi(x)$  durchgeführt werden. Zeigt dieses dann einen Erfolg an, so wird  $H_0$  schließlich abgelehnt, andernfalls angenommen. Die folgende Definition beschreibt die angekündigte Erweiterung des Testbegriffs.

**Definition 9.8.** *Es werde die Realisierung  $x$  einer Zufallsvariable  $X \sim P_\theta$  mit möglichen Werten in  $\Omega_X$  beobachtet.*

(i) *Ein **Test** für das Entscheidungsproblem*

$$H_0: \theta \in \Theta_0 \quad \text{gegen} \quad H_1: \theta \in \Theta_1$$

*ist eine Funktion  $\varphi: \Omega_X \rightarrow [0, 1]$ , wobei  $\varphi(x)$  die bedingte Wahrscheinlichkeit angibt, mit der bei Eintreten des Ereignisses  $X = x$  die Nullhypothese  $H_0$  abgelehnt wird.*

(ii) *Ein Test  $\varphi$  heißt **nicht randomisiert**, falls  $\varphi(x) \in \{0, 1\}$  für alle  $x \in \Omega_X$  gilt. Andernfalls heißt  $\varphi$  **randomisiert**.*

Falls  $X$  eine diskrete Zufallsvariable ist, so ist im Falle einer einfachen Nullhypothese  $H_0: \theta = \theta_0$  die Wahrscheinlichkeit eines Fehlers erster Art gegeben durch

$$\begin{aligned} & P_{\theta_0}(\text{„}H_0 \text{ wird abgelehnt“}) \\ &= \sum_{x \in \Omega_X} P_{\theta_0}(X = x \text{ und } H_0 \text{ wird abgelehnt}) \\ &= \sum_{x \in \Omega_X} \underbrace{P_{\theta_0}(H_0 \text{ wird abgelehnt} \mid X = x)}_{= \varphi(x)} P_{\theta_0}(X = x) \\ &= \sum_{x \in \Omega_X} \varphi(x) P_{\theta_0}(X = x) = E_{\theta_0} \varphi(X). \end{aligned}$$

Analog gilt im Falle einer einfachen Alternativhypothese  $H_1: \theta = \theta_1$  für die Wahrscheinlichkeit eines Fehlers zweiter Art, dass

$$\begin{aligned} & P_{\theta_1}(\text{„}H_0 \text{ wird angenommen“}) \\ &= \sum_{x \in \Omega_X} \underbrace{P_{\theta_1}(H_0 \text{ wird angenommen} \mid X = x)}_{= 1 - \varphi(x)} P_{\theta_1}(X = x) \\ &= \sum_{x \in \Omega_X} (1 - \varphi(x)) P_{\theta_1}(X = x) = E_{\theta_1}[1 - \varphi(X)]. \end{aligned}$$

Der nun folgende Satz zeigt, dass beim Testen einfacher Hypothesen mit dem auf randomisierte Tests erweiterten Testbegriff stets ein bestes Verfahren gefunden werden kann.

**Satz 9.9. (Lemma von Neyman-Pearson)**

Es werde die Realisierung  $x$  einer diskreten Zufallsvariable  $X \sim P_\theta$  mit Werten in  $\Omega_X$  beobachtet, wobei  $\theta \in \{\theta_0, \theta_1\}$ ,  $\theta_0 \neq \theta_1$ . Gegeben sei das Testproblem

$$H_0: \theta = \theta_0 \quad \text{gegen} \quad H_1: \theta = \theta_1.$$

Für beliebiges  $\alpha > 0$  existieren  $c_\alpha \in [0, \infty)$  und  $\gamma_\alpha \in [0, 1]$ , sodass der Test  $\varphi_\alpha: \Omega_X \rightarrow [0, 1]$  mit

$$\varphi_\alpha(x) = \begin{cases} 1, & \text{falls } P_{\theta_1}(X = x) > c_\alpha P_{\theta_0}(X = x), \\ \gamma_\alpha, & \text{falls } P_{\theta_1}(X = x) = c_\alpha P_{\theta_0}(X = x), \\ 0, & \text{falls } P_{\theta_1}(X = x) < c_\alpha P_{\theta_0}(X = x) \end{cases}$$

die Eigenschaft  $E_{\theta_0} \varphi_\alpha(X) = \alpha$  besitzt.

$\varphi_\alpha$  ist ein bester  $\alpha$ -Test für  $H_0$  gegen  $H_1$ .

*Beweis.* Wir betrachten Tests der Gestalt

$$\varphi(x) = \begin{cases} 1, & \text{falls } P_{\theta_1}(X = x) > c P_{\theta_0}(X = x), \\ \gamma, & \text{falls } P_{\theta_1}(X = x) = c P_{\theta_0}(X = x), \\ 0, & \text{falls } P_{\theta_1}(X = x) < c P_{\theta_0}(X = x) \end{cases}$$

und zeigen, dass  $c =: c_\alpha \in [0, \infty)$  und  $\gamma =: \gamma_\alpha \in [0, 1]$  existieren, sodass

$$\begin{aligned} \alpha &= E_{\theta_0} \varphi(X) \\ &= P_{\theta_0}^X \left( \{x \in \Omega_X: P_{\theta_1}(X = x) > c_\alpha P_{\theta_0}(X = x)\} \right) \\ &\quad + \gamma_\alpha P_{\theta_0}^X \left( \{x \in \Omega_X: P_{\theta_1}(X = x) = c_\alpha P_{\theta_0}(X = x)\} \right). \end{aligned} \quad (9.11)$$

Wir betrachten dazu die Funktion  $g$  mit

$$g(c) := P_{\theta_0}^X \left( \{x \in \Omega_X: P_{\theta_1}(X = x) \geq c P_{\theta_0}(X = x)\} \right).$$

$g$  is offensichtlich monoton nichtwachsend und es gelten  $g(0) = 1$  sowie, da das Wahrscheinlichkeitsmaß  $P_{\theta_0}^X$  die Eigenschaft der Stetigkeit von oben besitzt,

$$\begin{aligned} \lim_{n \rightarrow \infty} g(n) &= \lim_{n \rightarrow \infty} P_{\theta_0}^X \left( \{x \in \Omega_X : P_{\theta_1}(X = x) \geq n P_{\theta_0}(X = x)\} \right) \\ &= P_{\theta_0}^X \left( \underbrace{\bigcap_{n=1}^{\infty} \{x \in \Omega_X : P_{\theta_1}(X = x) \geq n P_{\theta_0}(X = x)\}}_{=\{x: P_{\theta_0}(X=x)=0\}} \right) \\ &= \sum_{x: P_{\theta_0}(X=x)=0} P_{\theta_0}(X = x) = 0. \end{aligned}$$

Es sei nun

$$c_\alpha := \sup \{c : g(c) \geq \alpha\}.$$

Wegen  $g(0) = 1$ ,  $g(n) \xrightarrow{n \rightarrow \infty} 0$  und  $\alpha > 0$  folgt, dass  $c_\alpha \in [0, \infty)$ .

Mit Stetigkeit von oben erhalten wir

$$\begin{aligned} &P_{\theta_0}^X \left( \{x \in \Omega_X : P_{\theta_1}(X = x) \geq c_\alpha P_{\theta_0}(X = x)\} \right) \\ &= P_{\theta_0}^X \left( \bigcap_{n=1}^{\infty} \{x \in \Omega_X : P_{\theta_1}(X = x) \geq (c_\alpha - 1/n) P_{\theta_0}(X = x)\} \right) \\ &= \lim_{n \rightarrow \infty} \underbrace{P_{\theta_0}^X \left( \{x \in \Omega_X : P_{\theta_1}(X = x) \geq (c_\alpha - 1/n) P_{\theta_0}(X = x)\} \right)}_{\geq \alpha} \geq \alpha \quad (9.12a) \end{aligned}$$

und mit Stetigkeit von unten

$$\begin{aligned} &P_{\theta_0}^X \left( \{x \in \Omega_X : P_{\theta_1}(X = x) > c_\alpha P_{\theta_0}(X = x)\} \right) \\ &= P_{\theta_0}^X \left( \bigcup_{n=1}^{\infty} \{x \in \Omega_X : P_{\theta_1}(X = x) \geq (c_\alpha + 1/n) P_{\theta_0}(X = x)\} \right) \\ &= \lim_{n \rightarrow \infty} \underbrace{P_{\theta_0}^X \left( \{x \in \Omega_X : P_{\theta_1}(X = x) \geq (c_\alpha + 1/n) P_{\theta_0}(X = x)\} \right)}_{< \alpha} \leq \alpha. \quad (9.12b) \end{aligned}$$

Ausgehend von (9.12a) und (9.12b) können wir nun noch  $\gamma_\alpha$  so bestimmen, dass (9.11) gilt. Falls  $P_{\theta_0}^X \left( \{x \in \Omega_X : P_{\theta_1}(X = x) > c_\alpha P_{\theta_0}(X = x)\} \right) = \alpha$ , so wählen wir  $\gamma_\alpha = 0$ .

Falls dagegen  $P_{\theta_0}^X \left( \{x \in \Omega_X : P_{\theta_1}(X = x) > c_\alpha P_{\theta_0}(X = x)\} \right) < \alpha$ , so wählen wir

$$\gamma_\alpha = \frac{\alpha - P_{\theta_0}^X \left( \{x \in \Omega_X : P_{\theta_1}(X = x) > c_\alpha P_{\theta_0}(X = x)\} \right)}{P_{\theta_0}^X \left( \{x \in \Omega_X : P_{\theta_1}(X = x) = c_\alpha P_{\theta_0}(X = x)\} \right)}.$$

Dann gilt insbesondere  $\gamma_\alpha \in [0, 1]$  und (9.11) ist auch in diesem Fall erfüllt.

Um die Optimalität von  $\varphi_\alpha$  zu beweisen nehmen wir an, dass  $\varphi^*: \Omega_X \rightarrow [0, 1]$  ein beliebiger (möglicherweise randomisierter)  $\alpha$ -Test ist. Wie im Beweis von Lemma 9.7 kann man die Ungleichung

$$(\varphi(x) - \varphi^*(x)) c P_{\theta_0}(X = x) \leq (\varphi(x) - \varphi^*(x)) P_{\theta_1}(X = x) \quad \forall x \in \Omega_X$$

beweisen. Daraus folgt, wiederum analog zum Beweis von Lemma 9.7, dass

$$\begin{aligned} 0 &\leq c (E_{\theta_0} \varphi(X) - E_{\theta_0} \varphi^*(X)) \\ &= \sum_{x \in \Omega_X} \underbrace{(\varphi(x) - \varphi^*(x)) c P_{\theta_0}(X = x)}_{\leq (\varphi(x) - \varphi^*(x)) P_{\theta_1}(X = x)} \\ &\leq \sum_{x \in \Omega_X} (\varphi(x) - \varphi^*(x)) P_{\theta_1}(X = x) \\ &= E_{\theta_1} \varphi(X) - E_{\theta_1} \varphi^*(X), \end{aligned}$$

woraus schließlich

$$E_{\theta_1} [1 - \varphi(X)] = 1 - E_{\theta_1} \varphi(X) \leq 1 - E_{\theta_1} \varphi^*(X) = E_{\theta_1} [1 - \varphi^*(X)],$$

folgt. □