STOCHASTIK FÜR DAS LEHRAMT

Vorlesungsskript, Studienjahr 2023/24

Christian Clason

Stand vom 5. Juli 2024

Institut für Mathematik und wissenschaftliches Rechnen
Universität Graz

Inhaltsverzeichnis

EINLEITUNG

Die Mathematik beschäftigt sich damit, aus klar definierten Strukturen logisch notwendige (nicht-triviale) Eigenschaften abzuleiten. Da sich viele Situation in der realen Welt durch solche Strukturen abbilden lassen (bzw. die Mathematik sich über Jahrtausende entwickelt hat, um möglichst viele Situationen so abzubilden), ist sie ein allgegenwertiges mächtiges Werkzeug in Physik, Technik, Wirtschafts-, und Sozialwissenschaften.

Nun ist leider in der Realität nicht immer klar, welche Situation konkret vorliegt. Nur mit partiellen Informationen bleiben also Unsicherheiten zurück, und statt mit Gewissheiten muss man mit *Wahrscheinlichkeiten* schließen. Solche Situationen kommen häufig vor:

- im *Glückspielen*: Zwar folgt ein Münz- oder Würfelwurf den Gesetzen der Mechanik, aber die für eine exakte Vorhersage notwendigen Informationen über den Wurf, die Luft, und die Beschaffenheit des Tisches sind nur im Physikunterricht zu erreichen;[1]
- in der *Physik*: Viele beobachtbare Phänomene etwa in der Wärmeleitung oder der Gasdynamik ergeben sich aus der Interaktion unzähliger mikroskopischer Zustände – viel zu vielen, um sie einzeln zu verfolgen;
- in der *Finanzwirtschaft*: Der Verlauf von Aktienkursen setzt sich aus zahllosen Kauf- und Verkaufsentscheidungen zusammen, die aus den verschiedensten (nicht immer gleichen und nicht immer rationalen) Gründen erfolgen;
- in der *Bildgebung*: Zum Beispiel in der Computertomographie möchte man aus (partiellen, verrauschten) Röntgenmessungen auf die Dichteverteilung im Körper schließen;
- in der *Wettervorhersage*: Der Zustand der Atmosphäre – und damit das Wetter – ist durch ein kompliziertes System partieller Differentialgleichungen vollständig beschrieben; um daraus das Wetter morgen abzuleiten, müssten wir aber das Wetter heute bereits perfekt (überall auf der Welt) kennen;
- in der *Quantenmechanik* ist schließlich nicht mal klar, ob eine objektive, deterministische, Beschreibung der Welt möglich ist.

[1] Tatsächlich motivierten Fragen des Glücksspiels – nicht nur aus hehren akademischen Interessen – viele der frühzeitlichen Mathematiker, die sich mit Wahrscheinlichkeitssrechnung beschäftigt haben.

Die mathematische Theorie für den rigorosen Umgang mit Unsicherheiten nennt man *Wahrscheinlichkeitstheorie*; ihre Anwendung auf konkrete Fragestellung sowie – hier besonders wichtig – der verantwortungsvolle Umgang mit den mathematischen Antworten ist die *Statistik*. Zusammen bezeichnet man dies als *Stochastik* (aus dem Griechischen, sinngemäß „scharfsinniges Vermuten“).

Während Statistiker trefflich darüber streiten können, was nun „Wahrscheinlichkeiten“ oder – allgemeiner – „Zufall“ eigentlich ist, hält sich der Mathematiker wie üblich fein heraus und definiert sich rein axiomatisch, wie sich vernünftige Wahrscheinlichkeiten zu verhalten haben. Der heute verwendete Zugang geht auf Andrei Nikolajewitsch Kolmogorow zurück, der diesen in den 1930er Jahren aufgestellt hat, und basiert auf den beiden Säulen der Mengenlehre und der Maßtheorie (die auch der modernen Integrationstheorie zugrunde liegt). Dies mag auf den ersten Blick (zu) abstrakt erscheinen, hat aber wesentlich dazu beigetragen, viele der naiven Behandlung entstammenden Paradoxa aufzulösen, indem sie implizite Modellannahmen über die *Art* der Unsicherheiten offenlegt.

Dieses Skriptum basiert vor allem auf den folgenden Werken:

- H.-O. Georgii (2015), *Stochastik, Einführung in die Wahrscheinlichkeitstheorie und Statistik*, 5. Aufl., Berlin: De Gruyter, DOI: 10.1515/9783110359701
- N. Henze (2021), *Stochastik für Einsteiger*, 13. Aufl., Berlin: Springer Spektrum, DOI: 10.1007/978-3-662-63840-8
- A. Eberle (2022), Stochastik, Vorlesungsskript, Institut für Angewandte Mathematik, Universität Bonn
- M. Hutzenthaler (2015), Stochastik, Vorlesungsskript, Fakultät für Mathematik, Universität Duisburg-Essen
- U. Krengel (2005), *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, 8. Aufl., Vieweg Stud. Wiesbaden: Vieweg, DOI: 10.1007/978-3-663-09885-0 (ergänzend)
- G. Kersting & A. Wakolbinger (2010), *Elementare Stochastik*, 2. Aufl., Mathematik Kompakt, Basel: Birkhäuser, DOI: 10.1007/978-3-0346-0414-7 (ergänzend)
- W. Linde (2014), *Stochastik für das Lehramt*, De Gruyter Studium, Berlin: De Gruyter, DOI: 10.1524/9783110362411 (ergänzend)
- A. Klenke (2020), *Wahrscheinlichkeitstheorie*, 4. Aufl., Masterclass, Berlin: Springer Spektrum, DOI: 10.1007/978-3-662-62089-2 (vertiefend)

Teil I

WAHRSCHEINLICHKEITSTHEORIE

1 ERGEBNISSE, EREIGNISSE, WAHRSCHEINLICHKEITEN

Wir beginnen mit der mathematischen Formalisierung von Zufallssituationen nach Kolmogorov durch den *Ergebnisraum* der möglichen Ausgänge, der *Ereignisraum* der sinnvoll betrachtbaren Aussagen, und dem *Wahrscheinlichkeitsverteilung*, die jedem Ereignis(!) eine Wahrscheinlichkeit zuordnet. Die Unterscheidung zwischen „Ergebnis“ und „Ereignis“ mag zwar auf den ersten Blick unnötig kompliziert erscheinen, ist aber wesentlich dafür, dass die Theorie auch mit unendlich vielen möglichen Ergebnissen (etwa im Intervall $[0,1]$) umgehen kann. Wir werden uns jedoch schnellstmöglich auf *diskrete* Wahrscheinlichkeiten beschränken, wo wir noch ohne diese technischen Schwierigkeiten auskommen. Dieser Rahmen ist aber notwendig, um auf Fragen wie „Wie lange muss ich den Tumor bestrahlen, um ihn mit über 95% Wahrscheinlichkeit zu zerstören?“ eine sinnvolle mathematische Antwort geben zu können.

1.1 ERGEBNISRÄUME

Der erste Schritt ist festzulegen, welche Ergebnisse in der Situation auftreten können und relevant sind; diese fast man zusammen in einer Menge Ω, genannt *Ergebnisraum*.

Beispiel 1.1. (i) *einmaliges Würfeln:* Bei einem Würfelspiel kommt es in der Regel nur auf die Anzahl der Augen an; der naheliegende Ergebnisraum ist daher

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

(In manchen Situationen kann aber auch

$$\Omega = \{1, 2, 3, 4, 5, 6, \text{„Würfel verloren“}\}$$

relevant sein…)

(ii) *mehrmaliges Würfeln:* Werden n Würfel hintereinander geworfen, gibt es verschiedene Möglichkeiten. Ist man nur an der Summe der Augenzahlen interessiert, so wählt man

$$\Omega = \{n, \dots, n \cdot 6\}.$$

Ist dagegen interessant, an welcher Position welche Augenzahl vorkam, ist der Ergebnisraum

$$\Omega = \{(\omega_1, \dots, \omega_n) \mid \omega_i \in \{1, \dots, 6\}, i = 1, , \dots, n\} =: \{1, \dots, 6\}^n.$$

(iii) *unendlich viele Münzwürfe:* Theoretisch interessant (wenn auch nicht praktisch relevant) ist das wiederholte Werfen einer Münze ohne aufzuhören. Nach Konvention bezeichnet man „Kopf“ mit 0 und „Zahl“ mit 1 und erhält so den (überabzählbaren) Ergebnisraum

$$\Omega = \{\{\omega_n\}_{n\in\mathbb{N}} \mid \omega_i \in \{0, 1\}, n \in \mathbb{N}\} =: \{0, 1\}^{\mathbb{N}}.$$

Wichtig ist nur, dass der Ergebnisraum alle Ergebnisse enthält, die tatsächlich auftreten können; er kann ohne Schaden immer größer gewählt werden. Zum Beispiel kann man für das einmalige Würfeln auch

$$\Omega = \{0, 1, 2, 3, 4, 5, 6\}$$

wählen.

1.2 EREIGNISRÄUME

Der zweite Schritt ist bereits der Kern der modernen Wahrscheinlichkeitstheorie: Wir sind nicht primär an einzelnen Ergebnissen interessiert, sondern daran, welche (beliebig komplizierten) Aussagen über ein Ergebnis zutreffen. (Damit ersparen wir uns, für jede Frage einen neuen Ergebnisraum definieren zu müssen.) Eine Aussage, die für ein konkretes Ergebnis eintreffen kann, nennt man *Ereignis*; mathematisch wird dies beschrieben durch eine Menge A von möglichen Ergebnissen $\omega \in \Omega$, für die die Aussage zutrifft.

Beispiel 1.2 (Würfeln). Wir betrachten für den einfachen Würfelwurf wieder den Ergebnisraum $\Omega = \{1, 2, 3, 4, 5, 6\}$. Mögliche Ereignisse sind dann

(i) „es wurde ein Sechser gewürfelt“: $A = \{6\}$;

(ii) „es wurde eine gerade Zahl gewürfelt“: $A = \{2, 4, 6\}$;

(iii) „es wurde mindestens eine Vier gewürfelt“: $A = \{4, 5, 6\}$;

(iv) „es wurde *kein* Sechser gewürfelt“: $A = \{1, 2, 3, 4, 5\}$.

Für den doppelten Würfelwurf können wir den Ergebnisraum $\Omega = \{1, \dots, 6\}^2$ betrachten. Mögliche Ergebnisse sind dann

(i) „es wurden zwei Sechser gewürfelt“: $A = \{(6, 6)\}$;

(ii) „die Augenzahlsumme beträgt 10“: $A = \{(4, 6), (5, 5), (6, 4)\}$;

(iii) „die Augenzahlsumme beträgt 9“: $A = \{(3,6),(4,5),(5,4),(6,3)\}$;

(iv) „der zweite Wurf ist höher als der erste Wurf“: $A = \{(1,2),(1,3),\ldots,(5,6)\}$.

Offensichtlich kann es beliebig aufwendig sein, Ereignisse durch die komplette Auflistung von Ergebnissen anzugeben. Man möchte daher die üblichen Mengenoperationen verwenden, um aus einfachen Ereignissen kompliziertere zu konstruieren. Als „einfache“ Ereignisse bieten sich an

- die *Elementarereignisse* $A = \{\omega\}$ für $\omega \in \Omega$;
- das *sichere Ereignis* $A = \Omega$;
- das *unmögliche Ereignis* $A = \emptyset$.

Hat man nun Ereignisse A und B, kann man fragen, wann bei einem Ergebnis $\omega \in \Omega$

- A oder B eintreten: das ist dann der Fall, wenn $\omega \in A$ oder $\omega \in B$ gilt, d. h. $\omega \in A \cup B$ gilt;
- A und B eintreten: das ist dann der Fall, wenn $\omega \in A$ und $\omega \in B$ gilt, d. h. $\omega \in A \cap B$ gilt;
- A nicht eintritt: das ist dann der Fall $\omega \notin A$ gilt, d. h. $\omega \in A^c := \Omega \setminus A$ gilt.

In anderen Worten: $A \cup B$ ist das Ereignis, dass A oder B eintreten; $A \cap B$ ist das Ereignis, dass A und B gleichzeitig eintreten; A^c ist das Ereignis, dass A nicht eintritt. Analog kann man über das alternative oder gleichzeitige Eintreten von mehr als zwei Ereignissen sprechen. Können A und B nicht gleichzeitig eintreten, d. h. gilt $A \cap B = \emptyset$, dann nennen wir die Ereignisse *disjunkt*.

Nun kommt die wesentliche Schwierigkeit: Es ist im Allgemeinen nicht möglich, für jedes beliebige Ereignis $A \in \mathcal{P}(\Omega) := \{A \mid A \subset \Omega\}$ sinnvoll über dessen Wahrscheinlichkeit zu sprechen. (Konkret ist dies nicht möglich, wenn Ω überabzählbar ist, z. B. $\Omega = [0,1]$ oder $\Omega = \{0,1\}^{\mathbb{N}}$; wir werden das weiter unten sehen.) Wir müssen uns daher vorab auf eine Menge von „vernünftigen“ Ereignissen einschränken, wobei wir garantieren möchten, dass die oben genannten Operationen auf vernünftigen Ereignissen wieder vernünftig sind. Das motiviert die folgende Definition.

Definition 1.3 (σ-Algebra). Sei Ω eine nichtleere Menge. Ein Mengensystem $\mathcal{A} \subset \mathcal{P}(\Omega)$ heißt *σ-Algebra*, falls gilt

(i) $\Omega \in \mathcal{A}$;

(ii) $A \in \mathcal{A}$ impliziert $A^c \in \mathcal{A}$;

(iii) $A_n \in \mathcal{A}$ für alle $n \in \mathbb{N}$ impliziert $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$

Das Paar $(\Omega, \mathcal{A})$ heißt *Ereignisraum*. Ist Ω höchstens abzählbar, so nennt man den Ereignisraum *diskret*.

Dass wir in (iii) die Vereinigung abzählbar viele Mengen betrachten ist wesentlich, um auch unendliche Ergebnisräume wie etwa bei unendlich vielen Münzwürfen betrachten zu können. Direkt aus der Definition folgt, dass für eine σ-Algebra stets gilt

- $\emptyset \in \mathcal{A}$;
- $A_n \in \mathcal{A}$ für alle $n \in \mathbb{N}$ impliziert $\bigcap_{n\in\mathbb{N}} A_n \in \mathcal{A}$;
- $A, B \in \mathcal{A}$ impliziert $A \cup B \in \mathcal{A}$ und $A \cap B \in \mathcal{A}$.

Offensichtlich eine σ-Algebra sind

- $\mathcal{A} = \{\Omega, \emptyset\}$ (die kleinstmögliche σ-Algebra);
- $\mathcal{A} = \mathcal{P}(\Omega)$ (die größtmögliche σ-Algebra, die für das weitere Vorgehen zu groß sein kann).

Auch sonst kann es von Vorteil sein, sich einzuschränken.

Beispiel 1.4. Wir betrachten das Werfen von n Münzen mit Ergebnisraum $\Omega = \{0,1\}^n$. Dann ist für $k \leq n$

$$\mathcal{A} = \left\{ A \subset \{0,1\}^n \,\middle|\, A = B \times \{0,1\}^{n-k} \text{ für ein } B \subset \{0,1\}^k \right\}$$

eine σ-Algebra; ihre Ereignisse sind genau die, die noch eintreten können, wenn wir nur die ersten k Würfe beobachtet haben.
Konkret ist für $n = 2$ und $k = 1$

$$\mathcal{A} = \{\emptyset, \{(0,0),(0,1)\}, \{(1,0),(1,1)\}, \{(0,0),(0,1),(1,0),(1,1)\}\}\,.$$

Dies ist echt kleiner als die Potenzmenge, da z. B. $\{(0,1),(1,1)\} \notin \mathcal{A}$ gilt. Das Ereignis, dass im zweiten Wurf nur die 1 auftritt – d. h. dass wir bereits wissen, dass in jedem Fall Zahl geworfen wird – ist also *nicht* vernünftig.

Üblicherweise wird man die komplette σ-Algebra nicht explizit angeben; stattdessen spezifiziert man nur die eigentlich „interessanten“ Ereignisse und nimmt dann alles dazu, was laut der Definition einer σ-Algebra dazugehört:

Definition 1.5 (erzeugte σ-Algebra). Sei Ω eine nichtleere Menge und $\mathcal{E} \subset \mathcal{P}(\Omega)$ beliebig. Dann ist

$$\sigma(\mathcal{E}) := \bigcap_{\mathcal{E}\subset\mathcal{A}\ \sigma\text{-Algebra}} \mathcal{A}$$

(die kleinste σ-Algebra, die $\mathcal{E}$ enthält), die *von $\mathcal{E}$ erzeugte σ-Algebra*.

Es ist nicht offensichtlich, dass die so erzeugte σ-Algebra stets existiert und eindeutig ist; für einen Beweis siehe [Georgii 2015, Bemerkung 1.6]. Umgekehrt können aber verschiedene $\mathcal{E}$ die selbe σ-Algebra erzeugen!

Diese Konstruktion führt auf die kanonischen Beispiele von σ-Algebren.

Beispiel 1.6. (i) Für Ω abzählbar betrachten wir die Sammlung

$$\mathcal{E} := \{\{\omega\} \mid \omega \in \Omega\}$$

der Elementarereignisse. Dann ist $\sigma(\mathcal{E}) = \mathcal{P}(\Omega)$, da jedes $A \in \mathcal{P}(\Omega)$ abzählbar ist und daher als abzählbare Vereinigung von Elementarereignissen dargestellt werden kann. Wie wir sehen werden, können wir für abzählbare Ergebnisräume stets die Potenzmenge wählen.

(ii) Für $\Omega = \mathbb{R}$ betrachten wir die Sammlung

$$\mathcal{E} := \{(a, b) \mid a, b \in \mathbb{R},\ a < b\}$$

aller offenen Intervalle. Dann heißt $\mathcal{B} := \sigma(\mathcal{E})$ die *Borel-Algebra*. Sie enthält alle offenen, abgeschlossenen, und halboffenen Intervalle, aber auch alle Mengen, die durch (beliebig viele!) Vereinigungen mit abzählbar vielen solchen Intervallen konstruiert werden können. Trotzdem gibt es Mengen (die man nicht explizit angeben kann), die nicht in der Borel-Algebra enthalten sind – darunter genau die, die später zu Problemen führen würden.

Die Borel-Algebra kann auf viele andere Arten erzeugt werden; besonders nützlich ist

$$\mathcal{E}' := \{(-\infty, c] \mid c \in \mathbb{R}\},$$

für das ebenfalls $\mathcal{B} = \sigma(\mathcal{E}')$ gilt.

(iii) Sei

$$\Omega = \bigtimes_{i \in I} \Omega_i := \{(\omega_i)_{i \in I} \mid \omega_i \in \Omega_i \text{ für alle } i \in I\}$$

das kartesische Produkt von beliebig vielen Mengen, und $\mathcal{A}_i \subset \mathcal{P}(\Omega_i)$ jeweils eine σ-Algebra. Bezeichne $\pi_j : \Omega \to \Omega_j$ die Projektion auf j-te Koordinate, d. h. $\pi_j \omega = \omega_j$ für $\omega = (\omega_i)_{i \in I}$. Betrachte

$$\mathcal{E} := \{A \subset \Omega \mid \{\pi_i \omega \mid \omega \in A\} \in \mathcal{A}_i \text{ für alle } i \in I\}.$$

Dann ist $\bigotimes_{i \in I} \mathcal{A}_i := \sigma(\mathcal{E})$ die *Produkt-σ-Algebra*.

Damit kann insbesondere eine σ-Algebra für den unendlich oft wiederholten Münzwurf konstruiert werden. In diesem Fall kann nach (i) $\mathcal{A}_i = \mathcal{P}(\{0, 1\})$ für alle $i \in \mathbb{N}$ gewählt werden; entsprechend wählen wir

$$\mathcal{E} = \{A \subset \Omega \mid A = \{\omega \in \Omega \mid \omega_i = a_i, i \le n\}, n \in \mathbb{N}, a_1, \ldots, a_n \in \{0, 1\}\},$$

d. h. alle Ereignisse, für die das Ereignis der ersten *endlich vielen* Würfe „vernünftig“ ist.

1.3 WAHRSCHEINLICHKEITSVERTEILUNGEN

Wir kommen nun zum letzten und wichtigsten Schritt: Wir legen für alle vernünftigen Ereignisse (und nur diese!) eine *Wahrscheinlichkeit* fest. Dabei interessiert uns nicht, wo diese herkommt; wir wollen nur, dass sie sich vernünftig verhält, d. h. sowohl unsere Intuition widerspiegelt als auch mathematisch konsistent ist. Was ist uns dabei wichtig?

(i) Wahrscheinlichkeiten sollten niemals negativ sein.

(ii) Die Wahrscheinlichkeit, dass zwei disjunkte Ereignisse gleichzeitig eintreffen, sollte gleich der Summe der Wahrscheinlichkeiten der beiden Ereignisse sein (also weder größer noch kleiner).

(iii) Die Wahrscheinlichkeit des sicheren Ereignisses normieren wir auf 1 („100%“).

Damit das auch für unendliche Ergebnismengen funktioniert, müssen wir in (ii) nicht nur zwei, sondern abzählbar viele paarweise disjunkte Ereignisse betrachten. Damit erhalten wir die folgende mathematische Definition.

Definition 1.7 (Wahrscheinlichkeitsverteilung). Sei Ω eine nichtleere Menge und $\mathcal{A}$ eine σ-Algebra. Eine Funktion $\mathbb{P} : \mathcal{A} \to [0, \infty]$ heißt *Wahrscheinlichkeitsverteilung* (oder *Wahrscheinlichkeitsmaß*) auf $\mathcal{A}$, wenn gilt

(i) $\mathbb{P}[\Omega] = 1$;

(ii) Für alle $A_n \in \mathcal{A}$, $n \in \mathbb{N}$, mit $A_i \cap A_j = \emptyset$ für $i \neq j$ gilt

$$\mathbb{P}\left[\bigcup_{n\in\mathbb{N}} A_n\right] = \sum_{n=1}^{\infty} \mathbb{P}[A_n].$$

Das Tripel $(\Omega, \mathcal{A}, \mathbb{P})$ wird *Wahrscheinlichkeitsraum* genannt.

Die Eigenschaft (ii) nennt man *σ-Additivität*. Beachten Sie, wie durch die Definition der σ-Algebra garantiert ist, dass die Eigenschaften (i) und (ii) sinnvoll sind. Weiter sind die Summanden auf der rechten Seite von (ii) alle nichtnegativ. Die Reihe konvergiert also genau dann, wenn sie absolut konvergiert; insbesondere ist der Wert nicht von der Reihenfolge abhängig. (Ansonsten divergiert sie bestimmt gegen ∞, wieder unabhängig von der Reihenfolge; da wir das bislang nicht ausschliessen können, müssen wir in der Definition auch den Wert ∞ zulassen.) Wir schreiben daher auch oft $\sum_{n\in\mathbb{N}} \mathbb{P}[A_n]$.

Direkt aus der Definition folgen weitere nützliche Eigenschaften.

Satz 1.8 (Rechenregeln für Wahrscheinlichkeitsverteilungen). *Sei $\mathbb{P}$ eine Wahrscheinlichkeitsverteilung auf der σ-Algebra $\mathcal{A}$. Dann gilt:*

(i) $\mathbb{P}[\emptyset] = 0$*;*

(ii) Für alle $A, B \in \mathcal{A}$ *ist*

$$\mathbb{P}[A \cap B] + \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$$

und damit insbesondere $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$*;*

(iii) Für alle $A, B \in \mathcal{A}$ *mit* $A \subset B$ *ist* $\mathbb{P}[A] \leq \mathbb{P}[B]$ *und damit insbesondere* $\mathbb{P}[A] \leq 1$*;*

(iv) Für alle $A_n \in \mathcal{A}$*,* $n \in \mathbb{N}$*, gilt*

$$\mathbb{P}\left[\bigcup_{n\in\mathbb{N}} A_n\right] \leq \sum_{n\in\mathbb{N}} \mathbb{P}[A_n].$$

Beweis. Zu (i): Es gibt kein $\omega \in \emptyset$ und damit insbesondere kein $\omega \in \emptyset \cap \emptyset$ und kein $\omega \in \emptyset \cap \Omega$, d. h. die leere Menge ist disjunkt sowohl zu sich selbst als auch zu Ω. Aus der σ-Additivät von $\mathbb{P}$ folgt daher

$$1 = \mathbb{P}[\Omega] = \mathbb{P}[\Omega \cup \emptyset \cup \emptyset \cup \dots] = 1 + \sum_{n=2}^{\infty} P[\emptyset].$$

Das ist wegen $\mathbb{P}[\emptyset] \geq 0$ aber nur möglich für $\mathbb{P}[\emptyset] = 0$.

Zu (ii): Wir nehmen zunächst an, dass A und B disjunkt sind. Dann folgt aus (i)

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

d. h. $\mathbb{P}$ ist auch *endlich additiv*.

Sind A und B nicht disjunkt, dann können wir disjunkt zerlegen (beachte $A \setminus B, B \setminus A \in \mathcal{A}$!) in

$$\begin{aligned}\mathbb{P}[A \cup B] + \mathbb{P}[A \cap B] &= \mathbb{P}[A \setminus B] + \mathbb{P}[B \setminus A] + 2\,\mathbb{P}[A \cap B] \\ &= (\mathbb{P}[A \setminus B] + \mathbb{P}[A \cap B]) + (\mathbb{P}[B \setminus A] + \mathbb{P}[A \cap B]) \\ &= \mathbb{P}[A] + \mathbb{P}[B].\end{aligned}$$

Für $B = A^c$ folgt daraus mit (i) sofort

$$\mathbb{P}[A] + \mathbb{P}[A^c] = \mathbb{P}[A \cup A^c] + \mathbb{P}[A \cap A^c] = \mathbb{P}[\Omega] + P[\emptyset] = 1 + 0 = 1.$$

Zu (iii): Ist $A \subset B$, dann folgt aus der endlichen Additivität nach (ii) für $B = A \cup B \setminus A$

$$\mathbb{P}[B] = \mathbb{P}[A] + \mathbb{P}[B \setminus A] \geq \mathbb{P}[A]$$

wegen der Nichtnegativität von $\mathbb{P}$. Speziell mit $B = \Omega$ gilt dann wegen $A \subset \Omega$ für alle $A \in \mathcal{A}$

$$\mathbb{P}[A] \leq \mathbb{P}[\Omega] = 1.$$

Zu (iv): Auch wenn die A_n nicht disjunkt sind, können wir ihre Vereinigung trotzdem schreiben als Vereinigung disjunkter Mengen: Wir ersetzen A_n einfach durch die Teilmenge der Elemente, die nicht schon in einem „früheren“ A_j mit $j < n$ enthalten sind. Aus der σ-Additivität und (iii) folgt dann

$$\mathbb{P}\left[\bigcup_{n\in\mathbb{N}} A_n\right] = \mathbb{P}\left[\bigcup_{n\in\mathbb{N}}\left(A_n \setminus \bigcup_{j<n} A_j\right)\right] = \sum_{n\in\mathbb{N}} \mathbb{P}\left[A_n \setminus \bigcup_{j<n} A_j\right] \leq \sum_{n\in\mathbb{N}} \mathbb{P}[A_n]. \qquad \square$$

Aussage (ii) lässt sich auf endlich viele Ereignisse verallgemeinern. Für drei Ereignisse $A, B, C \in \mathcal{A}$ erhält man zum Beispiel durch zweimaliges Anwenden von (ii)

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

Durch Induktion erhält man daraus die folgende Rechenregel (die wir später deutlich einfacher beweisen werden).

Folgerung 1.9 (Einschluss-Ausschluss-Prinzip). *Für alle $n \in \mathbb{N}$ und $A_1, \ldots, A_n \in \mathcal{A}$ gilt*

$$\mathbb{P}\left[\bigcup_{k=1}^{n} A_k\right] = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1\leq i_1<\cdots<i_k\leq n} \mathbb{P}\left[\bigcap_{j=1}^{k} A_{i_j}\right].$$

Für folgende, noch recht triviale, Beispiele lässt sich leicht nachprüfen, dass sie die Definition erfüllen.

Beispiel 1.10 (Wahrscheinlichkeitsverteilungen). (i) Ist Ω eine nichtleere Menge und $\omega \in \Omega$ beliebig, dann ist für jede σ-Algebra $\mathcal{A}$ durch

$$\mathbb{P}[A] = \begin{cases} 1 & \text{falls } \omega \in A, \\ 0 & \text{sonst,} \end{cases}$$

eine Wahrscheinlichkeitsverteilung, genannt *Dirac-Maß*, definiert. Sie drückt aus, dass das Ergebnis ω sicher ist, d. h. dass das Ergebnis gar nicht zufällig ist.

(ii) Wir betrachten mal wieder den Münzwurf. Für jedes $p \in [0,1]$ wird auf der σ-Algebra

$$\mathcal{A} = \mathcal{P}(\{0,1\}) = \{\emptyset, \{0\}, \{1\}, \{0,1\}\}$$

durch die Zuweisung

$$\mathbb{P}[\emptyset] = 0, \quad \mathbb{P}[\{0,1\}] = 1, \quad \mathbb{P}[\{1\}] = p, \quad \mathbb{P}[\{0\}] = 1 - p,$$

eine Wahrscheinlichkeitsverteilung, genannt *Bernoulli-Verteilung mit Parameter* p, definiert. Ist $p = \frac{1}{2}$, so ist der Münzwurf *fair*.

Für den Münzwurf hat es offensichtlich gereicht, die Wahrscheinlichkeit der Elementarereignisse anzugeben. Das ist allgemein für abzählbare Ereignisräume möglich.

Satz 1.11. *Sei Ω eine abzählbare Menge und $p : \Omega \to \mathbb{R}$ eine* Zähldichte*, d. h. erfüllt*

(i) $0 \le p(\omega) \le 1$ *für alle* $\omega \in \Omega$*;*

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

Dann wird durch

$$\mathbb{P}[A] := \sum_{\omega\in A} p(\omega), \qquad A \subset \Omega,$$

eine Wahrscheinlichkeitsverteilung auf $\mathcal{P}(\Omega)$ definiert. Umgekehrt hat jede Wahrscheinlichkeitsverteilung auf $\mathcal{P}(\Omega)$ eine Zähldichte $p(\omega) := \mathbb{P}[\{\omega\}]$.

Beweis. Zunächst sind wegen (i) alle Summanden nichtnegativ, so dass die Reihen konvergieren genau dann, wenn sie absolut konvergieren, und ihr Wert hängt nicht von der Reihenfolge der Summanden ab; dies rechtfertigt auch die Schreibweise. Weiter gilt nach Definition

$$\mathbb{P}[\Omega] = \sum_{\omega\in\Omega} p(\omega) = 1,$$

also ist $\mathbb{P}$ normiert, und alle Reihen sind absolut konvergent.

Für die σ-Additivität seien $A_n \subset \Omega, n \in \mathbb{N}$, paarweise disjunkt. Da die Vereinigung abzählbar vieler Mengen wieder abzählbar ist, können wir schreiben

$$\mathbb{P}\left[\bigcup_{n\in\mathbb{N}} A_n\right] = \sum_{\omega\in\bigcup_{n\in\mathbb{N}} A_n} p(\omega) = \sum_{n\in\mathbb{N}} \sum_{\omega\in A_n} p(\omega) = \sum_{n\in\mathbb{N}} \mathbb{P}[A_n].$$

Ist umgekehrt $\mathbb{P}$ eine Wahrscheinlichkeitsverteilung auf $\mathcal{P}(\Omega)$, dann gilt insbesondere $\{\omega\} \in \mathcal{P}(\Omega)$ für alle $\omega \in \Omega$. Also folgt für alle $A \subset \Omega$ aus der σ-Additivität

$$\mathbb{P}[A] = \mathbb{P}\left[\bigcup_{\omega\in A} \{\omega\}\right] = \sum_{\omega\in A} \mathbb{P}[\{\omega\}],$$

da A auch höchstens abzählbar und die Vereinigung offensichtlich paarweise disjunkt ist. Insbesondere gilt für $A = \Omega$

$$\sum_{\omega\in\Omega} \mathbb{P}[\{\omega\}] = \mathbb{P}[\Omega] = 1.$$

Außerdem folgt aus Satz 1.8 (iii) auch

$$0 \le \mathbb{P}[\{\omega\}] \le \mathbb{P}[\Omega] = 1 \quad \text{für alle } \omega \in \Omega.$$

Also ist $p(\omega) := \mathbb{P}[\{\omega\}]$ eine Zähldichte. □

Eine Wahrscheinlichkeitsverteilung, die durch eine Zähldichte definiert ist, nennt man *diskrete Wahrscheinlichkeitsverteilung*.

Beispiel 1.12. Für $\Omega = \mathbb{N}$ betrachte auf $\mathcal{P}(\Omega)$ die Zähldichte

$$p(n) = 2^{-n}, \qquad n \in \mathbb{N}.$$

Dann gilt für die entsprechende diskrete Wahrscheinlichkeitsverteilung zum Beispiel für die Menge der geraden Zahlen $A = \{2, 4, 6, \dots\}$ mit Hilfe der geometrischen Reihe

$$\mathbb{P}[A] = \sum_{n \in A} \frac{1}{2^n} = \sum_{j=1}^{\infty} \frac{1}{2^{2j}} = \sum_{j=1}^{\infty} \left(\frac{1}{4}\right)^j = \frac{1}{1 - \frac{1}{4}} - 1 = \frac{1}{3}.$$

Speziell für endliche Mengen kann man die Zähldichte konstant wählen; dies definiert eine eindeutige diskrete Wahrscheinlichkeitsverteilung, die in der naiven Wahrscheinlichkeitsrechnung oft implizit angesetzt wird.

Definition 1.13 (diskrete Gleichverteilung). Sei Ω endlich. Dann definiert

$$p(\omega) := \frac{1}{|\Omega|}, \qquad \omega \in \Omega,$$

eine Zähldichte, wobei $|\Omega|$ die (endliche) Anzahl der Elemente von Ω bezeichnet. Die dadurch definierte Wahrscheinlichkeitsverteilung

$$\mathbb{P}[A] = \sum_{\omega \in A} \frac{1}{|\Omega|} = \frac{|A|}{|\Omega|}, \quad A \subset \Omega, \tag{1.1}$$

wird *Gleichverteilung* oder *Laplace-Verteilung* auf $\mathcal{P}(\Omega)$ genannt und auch mit $\mathcal{U}(\Omega)$ bezeichnet.

Die Definition (1.1) begründet die bekannte Rechenformel „Anzahl günstiger Ergebnisse dividiert durch Anzahl aller Ergebnisse“ – allerdings nur für gleichverteilte Ereignisse!

Beispiel 1.14 (Gleichverteilungen). (i) *Münzwürfe:* Für das Werfen von n Münzen hat der Ergebnisraum $\Omega = \{0, 1\}^n$ genau 2^n Elemente; die Gleichverteilung auf $\mathcal{P}(\Omega)$ hat also die Zähldichte $p(\omega) = 2^{-n}$.

Damit hat für $k \le n$ das Ereignis

$$A_k = \{(\omega_1, \dots, \omega_n) \in \{0, 1\}^n \mid \omega_k = 1\}$$

im k-ten Wurf „Zahl“ zu erhalten, die Wahrscheinlichkeit

$$\mathbb{P}[A_k] = \frac{|A_k|}{|\Omega|} = \frac{2^{n-1}}{2^n} = \frac{1}{2},$$

denn durch Festlegen von ω_k bleiben noch jeweils zwei Möglichkeiten für die $n-1$ anderen Würfe übrig.

(ii) *Würfeln:* Wir betrachten wieder den doppelten Würfelwurf aus Beispiel 1.2 mit Ergebnisraum $\Omega = \{1,\dots,6\}^2$. Die Gleichverteilung auf $\mathcal{P}(\Omega)$ hat dann die Zähldichte $p(\omega) = \frac{1}{36}$. Wir erhalten dann die Wahrscheinlichkeit für das Ergebnis

a) „die Augenzahlsumme beträgt 10“: $A = \{(4,6),(5,5),(6,4)\}$,

$$\mathbb{P}[A] = \frac{3}{36} = \frac{1}{12};$$

b) „die Augenzahlsumme beträgt 9“: $A = \{(3,6),(4,5),(5,4),(6,3)\}$,

$$\mathbb{P}[A] = \frac{4}{36} = \frac{1}{9} > \frac{1}{12};$$

c) „der zweite Wurf ist höher als der erste Wurf“: $A = \{(1,2),(1,3),\dots,(5,6)\}$,

$$\mathbb{P}[A] = \frac{5+4+3+2+1}{36} = \frac{15}{36} < \frac{1}{2}.$$

Für Wahrscheinlichkeitsverteilungen auf überabzählbaren Mengen kann man nur schwächere Aussagen machen. Für den Beweis etwa der folgenden Aussage sei auf [Georgii 2015, Satz 1.12] verwiesen.

Satz 1.15. *Sei Ω eine nichtleere Menge und $\mathcal{E} \subset \mathcal{P}(\Omega)$ mit der Eigenschaft, dass für $A, B \in \mathcal{E}$ auch $A \cap B \in \mathcal{E}$ liegt. Seien $\mathbb{P}, \mathbb{P}'$ zwei Wahrscheinlichkeitsverteilungen auf $\sigma(\mathcal{E})$. Gilt $\mathbb{P}[A] = \mathbb{P}'[A]$ für alle $A \in \mathcal{E}$, so ist $\mathbb{P} = \mathbb{P}'$.*

Zum Abschluß geben wir noch ein Resultat an, das zeigt, dass es unmöglich ist, für den fairen Münzwurf ohne Aufhören eine Wahrscheinlichkeitsverteilung für *alle* Ereignisse zu definieren. Den Beweis (der auf dem Auswahlaxiom beruht) findet man z. B. in [Georgii 2015, Satz 1.5].

Satz 1.16 (Vitali). *Sei $\Omega = \{0,1\}^{\mathbb{N}}$. Dann existiert keine Abbildung $P : \mathcal{P}(\Omega) \to [0,1]$ mit*

(i) $P(\Omega) = 1$;

(ii) P *ist σ-additiv;*

(iii) Für alle $A \subset \Omega$ und $n \in \mathbb{N}$ gilt $P(A) = P(T_n A)$ für

$$T_n A := \{(\omega_1,\dots,\omega_{n-1}, 1-\omega_n, \omega_{n+1},\dots) \mid \omega \in A\}.$$

Die letzte Bedingung sagt aus, dass für jeden einzelnen Wurf sowohl Kopf als auch Zahl mit gleicher Wahrscheinlichkeit auftauchen können und drückt damit die Fairness und Unabhängigkeit der einzelnen Würfe aus. Analog kann man zeigen, dass z. B. auf $\mathcal{P}([0,1])$ keine Gleichverteilung existieren kann. Man muss sich daher zwingend auf eine strikt kleinere σ-Algebra (z. B. die Borel-Algebra) einschränken.

2 DISKRETE ZUFALLSVARIABLEN

Wir haben im letzten Kapitel aus dem *Ergebnis* eines doppelten Würfelwurfs das *Ereignis* „Summe der Augen ist 9“ gebildet. Wir wollen dies nun verallgemeinern, indem wir die Augensumme selbst als Ergebnis betrachten, so dass wir daraus wieder Ereignisse (etwa „die Summe ist größer als 6“) bilden können; das Ziel ist dann, aus der Verteilung der ursprünglichen Ereignisse für den Würfelwurf die (komplette!) Verteilung der Augensummen zu bestimmen. Mathematisch bedeutet das, *Funktionen* auf Ereignisräumen zu betrachten (in unserem Beispiel etwa $\omega = (\omega_1, \omega_2) \mapsto \omega_1 + \omega_2$. Natürlich ist dabei wieder wichtig, dass wir nur „vernünftige“ Ereignisse betrachten müssen. Das führt auf die folgende Definition.

Definition 2.1 (Zufallsvariable). Seien $(\Omega, \mathcal{A})$ und $(S, \mathcal{S})$ zwei Ereignisräume. Eine Funktion

$$X : \Omega \to S$$

heißt *Zufallsvariable* von $(\Omega, \mathcal{A})$ nach $(S, \mathcal{S})$, wenn gilt

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

Ist S abzählbar, so nennt man die Zufallsvariable *diskret*.

Wir schreiben für $X^{-1}(B)$ in Folge oft auch kurz $\{X \in B\}$. Da die Potenzmenge nach Beispiel 1.6 (i) durch die Elementarereignisse erzeugt wird, genügt es für diskrete Zufallsvariablen, Mengen der Form

$$X^{-1}(\{s\}) = \{\omega \in \Omega \mid X(\omega) = s\} =: \{X = s\} \qquad \text{für alle } s \in S$$

zu betrachten.

Einfache Beispiele von Zufallsvariablen sind die folgenden.

Beispiel 2.2. (i) Für jeden Ereignisraum $(\Omega, \mathcal{A})$ ist die *Identität*

$$\mathrm{Id}_\Omega : \Omega \to \Omega, \qquad \mathrm{Id}_\Omega(\omega) = \omega,$$

eine Zufallsvariable von $(\Omega, \mathcal{A})$ nach $(\Omega, \mathcal{A})$.

(ii) Für jeden Ereignisraum $(\Omega, \mathcal{A})$ und jedes $A \in \mathcal{A}$ ist die *Indikatorfunktion*

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

eine diskrete Zufallsvariable von $(\Omega, \mathcal{A})$ nach $(\{0,1\}, \mathcal{P}(\{0,1\}))$.

(iii) Ist Ω abzählbar, so ist *jede* Funktion $X : \Omega \to S$ eine Zufallsvariable von $(\Omega, \mathcal{P}(\Omega))$ nach $(S, \mathcal{P}(S))$.

Die Indikatorfunktion erlaubt insbesondere, „mit Ereignissen zu rechnen“. Z.B. rechnet man leicht nach, dass gilt

$$\mathbb{1}_{A^c} = 1 - \mathbb{1}_A, \qquad \mathbb{1}_{A\cap B} = \min\{\mathbb{1}_A, \mathbb{1}_B\} = \mathbb{1}_A \cdot \mathbb{1}_B, \qquad \mathbb{1}_{A\cup B} = \max\{\mathbb{1}_A, \mathbb{1}_B\} \le \mathbb{1}_A + \mathbb{1}_B,$$

wobei letztere Ungleichung mit Gleichheit gilt, falls A und B disjunkt sind.

Wir kommen nun zum wesentlichen Schritt: die Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$ definiert über die Zufallsvariable $X : \Omega \to S$ eine Wahrscheinlichkeitsverteilung auf $(S, \mathcal{S})$.

Satz 2.3 (Verteilung einer Zufallsvariablen). *Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und X eine Zufallsvariable von $(\Omega, \mathcal{A})$ nach $(S, \mathcal{S})$. Dann wird durch*

$$\mathbb{P}_X : \mathcal{S} \to [0,\infty], \qquad \mathbb{P}_X(B) := \mathbb{P}[X^{-1}(B)] = \mathbb{P}[\{X \in B\}] \qquad \textit{für alle } B \in \mathcal{S},$$

eine Wahrscheinlichkeitsverteilung auf $(S, \mathcal{S})$ definiert, genannt Verteilung von X.

Beweis. Wir müssen lediglich die beiden Eigenschaften aus Definition 1.7 überprüfen.

Zu (i): Da $\mathbb{P}$ eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$ ist, folgt sofort

$$\mathbb{P}_X[S] = \mathbb{P}[\{X \in S\}] = \mathbb{P}[\Omega] = 1,$$

da $X(\omega) \in S$ für alle $\omega \in \Omega$ gilt.

Zu (ii): Seien $B_n \in \mathcal{S}$, $n \in \mathbb{N}$, paarweise disjunkte Ereignisse. Dann gilt für alle $i \neq j$

$$\begin{aligned} X^{-1}(B_i) \cap X^{-1}(B_j) &= \{\omega \in \Omega \,|\, X(\omega) \in B_i\} \cap \{\omega \in \Omega \,|\, X(\omega) \in B_j\} \\ &= \{\omega \in \Omega \,|\, X(\omega) \in B_i \text{ und } X(\omega) \in B_j\} \\ &= \{\omega \in \Omega \,|\, X(\omega) \in B_i \cap B_j\} \\ &= \{\omega \in \Omega \,|\, X(\omega) \in \emptyset\} = \emptyset, \end{aligned}$$

d. h. auch die $X^{-1}(B_n)$ sind paarweise disjunkt. Weiter gilt analog

$$\begin{aligned} X^{-1}\left(\bigcup_{n\in\mathbb{N}} B_n\right) &= \left\{\omega \in \Omega \,\middle|\, X(\omega) \in \bigcup_{n\in\mathbb{N}} B_n\right\} \\ &= \bigcup_{n\in\mathbb{N}} \{\omega \in \Omega \mid X(\omega) \in B_n\} = \bigcup_{n\in\mathbb{N}} X^{-1}(B_n). \end{aligned}$$

Da $\mathbb{P}$ Wahrscheinlichkeitsverteilung ist, folgt dann

$$\begin{aligned} \mathbb{P}_X\left[\bigcup_{n\in\mathbb{N}} B_n\right] &= \mathbb{P}\left[X^{-1}\left(\bigcup_{n\in\mathbb{N}} B_n\right)\right] = \mathbb{P}\left[\bigcup_{n\in\mathbb{N}} X^{-1}(B_n)\right] \\ &= \sum_{n\in\mathbb{N}} \mathbb{P}[X^{-1}(B_n)] = \sum_{n\in\mathbb{N}} \mathbb{P}_X[B_n]. \end{aligned}$$

□

Der Übersichtlichkeit halber lässt man oft die Mengenklammern weg und schreibt

$$\mathbb{P}[X \in B] := \mathbb{P}[\{X \in B\}], \qquad \mathbb{P}[X = s] := \mathbb{P}[\{X = s\}], \qquad \text{etc.}$$

Beispiel 2.4. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum.

(i) Für die Identität Id_Ω auf $(\Omega, \mathcal{A})$ gilt offensichtlich für alle $A \in \mathcal{A}$

$$\mathbb{P}_{\mathrm{Id}_\Omega}[A] = \mathbb{P}[\mathrm{Id}_\Omega^{-1}(A)] = \mathbb{P}[\{\omega \in \Omega \mid \omega \in A\}] = \mathbb{P}[A],$$

d. h. $\mathbb{P}_{\mathrm{Id}_\Omega} = \mathbb{P}$.

(ii) Für die Indikatorfunktion $\mathbb{1}_A$ auf $(\Omega, \mathcal{A})$ für $A \in \mathcal{A}$ gilt offensichtlich

$$\begin{aligned} \mathbb{P}_{\mathbb{1}_A}[\{1\}] &= \mathbb{P}[\mathbb{1}_A = 1] = \mathbb{P}[A], \\ \mathbb{P}_{\mathbb{1}_A}[\{0\}] &= \mathbb{P}[\mathbb{1}_A = 0] = \mathbb{P}[A^c]. \end{aligned}$$

(iii) Ist $X : \Omega \to S$ eine diskrete Zufallsvariable von $(\Omega, \mathcal{A})$ nach $(S, \mathcal{P}(S))$, dann ist nach Satz 1.11 die Verteilung $\mathbb{P}_X$ eindeutig festgelegt durch

$$\mathbb{P}_X[B] = \sum_{s\in B} p_X(s)$$

für die Zähldichte

$$p_X(s) = \mathbb{P}_X[\{s\}] = \mathbb{P}[X = s] \qquad \text{für alle } s \in S.$$

Betrachte zum Beispiel den dreifachen fairen Münzwurf, beschrieben durch den Wahrscheinlichkeitsraum $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ für $\Omega = \{0,1\}^3$ und $\mathbb{P} = \mathcal{U}(\Omega)$ die Gleichverteilung, und als Zufallsvariable

$$X : \{0,1\}^3 \to \{0,1,2,3\}, \qquad X(\omega) = \omega_1 + \omega_2 + \omega_3,$$

d. h. die (zufällige) Anzahl, wie oft Zahl geworfen wurde. Wegen $|\{0,1\}^3| = 8$ ist dann

$$\mathbb{P}_X[\{0\}] = \mathbb{P}[X=0] = \mathbb{P}[\{(0,0,0)\}] = \frac{1}{8},$$
$$\mathbb{P}_X[\{1\}] = \mathbb{P}[X=1] = \mathbb{P}[\{(1,0,0),(0,1,0),(0,0,1)\}] = \frac{3}{8},$$
$$\mathbb{P}_X[\{2\}] = \mathbb{P}[X=2] = \mathbb{P}[\{(1,1,0),(1,0,1),(0,1,1)\}] = \frac{3}{8},$$
$$\mathbb{P}_X[\{3\}] = \mathbb{P}[X=3] = \mathbb{P}[\{(1,1,1)\}] = \frac{1}{8}.$$

Also ist $\mathbb{P}_X$ nicht mehr die Gleichverteilung!

Definition 2.5. Sei $(S, \mathcal{S})$ ein Ereignisraum und X und Y Zufallsvariablen nach $(S, \mathcal{S})$. Gilt $\mathbb{P}_X = \mathbb{P}_Y$, dann heißen X und Y *identisch verteilt*, geschrieben $X \sim Y$.

Gilt insbesondere $\mathbb{P}_X = \mathcal{U}(S)$, so sagt man, X ist *gleichverteilt*.

Beachten Sie, dass X und Y weder die selbe Abbildungsvorschrift noch den selben Definitionsbereich haben müssen – es ist lediglich die Verteilung der Werte zu vergleichen.

DISKRETE WAHRSCHEINLICHKEITSVERTEILUNGEN

Durch Betrachten geeigneter Zufallsvariablen können wir nun aus der bereits bekannten (und leicht anzugebenden) Gleichverteilung weitere Wahrscheinlichkeitsverteilungen konstruieren. Wir beginnen mit den klassischen *Urnenmodellen*. Angenommen, in einer Urne liegen Kugeln, die von außen nicht unterscheidbar sind, von denen aber ein gewisser Anteil $p \in (0,1)$ irgendwie markiert ist. Aus dieser Urne entnehmen wir nun blind n Kugeln, wobei wir dies entweder mit Zurücklegen oder ohne tun können.

URNENMODELLE MIT ZURÜCKLEGEN

Sei K die Menge der Kugeln in der Urne (die wir uns z. B. durchnummeriert vorstellen können, so dass $K = \{1, \dots, m\}$ gilt für ein $m \in \mathbb{N}$). Wenn wir n Kugeln daraus ziehen, ist der Ergebnisraum

$$\Omega = K^n = \{(\omega_1, \dots, \omega_n) \mid \omega_i \in K \text{ für } i = 1, \dots, n\},$$

da eine Kugel durch Zurücklegen mehrfach gezogen werden kann. Dies ist eine endliche Menge mit $|\Omega| = |K|^n$; wir können als Ereignisraum also die Potenzmenge $\mathcal{P}(\Omega)$ wählen.

Da die Kugeln von außen nicht unterscheidbar sein sollen, nehmen wir als Wahrscheinlichkeitsverteilung die Gleichverteilung

$$\mathbb{P}[A] = \frac{|A|}{|\Omega|}, \qquad A \subset \Omega,$$

an. Nun seien $M \subset K$ dieser Kugeln markiert, und wir fragen uns, wie viele davon unter den n gezogenen Kugeln sind. Dafür betrachten wir die Zufallsvariable

$$X : \Omega \to \{0, 1, \ldots, n\}, \qquad X(\omega) = |\{i \mid \omega_i \in M\}|.$$

Dann hat X nach Satz 2.3 die Verteilung

$$\mathbb{P}_X[B] = \mathbb{P}[X \in B] = \frac{|\{X \in B\}|}{|\Omega|}, \qquad B \subset \{0, 1, \ldots, n\},$$

für die wir nach Satz 1.11 nur die Zähldichte $p_X(k) := \mathbb{P}_X[\{k\}] = \mathbb{P}[X = k]$ kennen müssen. Dazu berechnen wir

$$|\{X = k\}| = \{(\omega_1, \ldots, \omega_n) \mid \omega_i \in M \text{ für genau } k \text{ verschiedene } i\} = \binom{n}{k} |M|^k |K \setminus M|^{n-k},$$

denn $\binom{n}{k}$ ist die Anzahl der Möglichkeiten, k Indizes aus $\{1, \ldots, n\}$ zu wählen. Für diese k Indizes gibt es $|M|^k$ Möglichkeiten, markierte Kugeln zu wählen; für die restlichen $n - k$ Indizes gibt es dann $|K \setminus M|^{n-k}$ Möglichkeiten, unmarkierte Kugeln zu wählen. Wegen $|\Omega| = |K|^n$ folgt

$$\mathbb{P}_X[\{k\}] = \frac{\binom{n}{k} |M|^k |K \setminus M|^{n-k}}{|K|^n} = \binom{n}{k} \left(\frac{|M|}{|K|}\right)^k \left(\frac{|K \setminus M|}{|K|}\right)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k},$$

da $p = |M|/|K|$ genau der Anteil der markierten Kugeln ist.

Die dadurch definierte Wahrscheinlichkeitsverteilung $\mathrm{Bin}(n, p)$ auf $\mathcal{P}(\{0, \ldots, n\})$ mit Zähldichte $b_{n,p}(k) := \binom{n}{k} p^k (1-p)^{n-k}$ wird *Binomialverteilung* mit Parametern n und p genannt. (Beachten Sie, dass wir nicht von Hand nachweisen mussten, dass dies tatsächlich eine Wahrscheinlichkeitsverteilung ist, da das nach Konstruktion durch Satz 1.11 garantiert ist!)

Aus der Kenntnis der Zähldichte lassen sich nun mit Satz 1.11 und den Rechenregeln aus Satz 1.8 die Wahrscheinlichkeiten beliebiger Ereignisse berechnen.

Beispiel 2.6. Angenommen, die Hälfte der Kugeln in der Urne ist markiert, d. h. $p = \frac{1}{2}$. Was ist die Wahrscheinlichkeit, dass von $n = 10$ gezogenen Kugeln mindestens 3 markiert sind? Dies entspricht dem Ereignis

$$A = \{3, 4, 5, 6, 7, 8, 9, 10\} = \{0, 1, 2\}^c.$$

Dann gilt

$$\mathbb{P}_{\mathrm{Bin}(10,0.5)}[\{0,1,2\}] = \sum_{k=0}^{2} b_{10,0.5}(k) = \frac{1}{2^{10}} \sum_{k=0}^{2} \binom{10}{k} = \frac{1+10+45}{1024} = \frac{56}{1024}$$

und damit

$$\mathbb{P}_{\mathrm{Bin}(10,0.5)}[A] = 1 - \mathbb{P}_{\mathrm{Bin}(10,0.5)}[\{0,1,2\}] = \frac{968}{1024} > 0.94.$$

URNENMODELLE OHNE ZURÜCKLEGEN

Wir betrachten nun den Fall, dass keine Kugel zweimal gezogen werden kann (entweder, weil sie nach dem Ziehen beiseite gelegt wird, oder weil alle n Kugeln mit einem Griff gezogen werden). Hier ist also der Ergebnisraum

$$\Omega = \left\{(\omega_1, \dots, \omega_n) \in K^n \,\middle|\, \omega_i \neq \omega_j \text{ für alle } i \neq j\right\} \subsetneq K^n.$$

Dies ist wieder eine endliche Menge mit $|\Omega| = n!\binom{|K|}{n}$ Elementen (genau der Anzahl von Möglichkeiten, *n verschiedene* Elemente aus K zu wählen, multipliziert mit der Anzahl der möglichen Reihenfolgen dieser Elemente), und wir können auf $\mathcal{P}(\Omega)$ wieder die Gleichverteilung annehmen. Was ist nun die Wahrscheinlichkeitsverteilung von

$$X : \Omega \to \{0, 1, \dots, \}, \qquad X(\omega) = |\,\{i \,|\, \omega_i \in M\}\,|,$$

der Anzahl der markierten Kugeln, die wir gezogen haben? Wieder müssen wir

$$\mathbb{P}_X[\{k\}] = \mathbb{P}[X = k] = \frac{|\{X = k\}|}{|\Omega|}, \qquad k = 0, \dots, n,$$

berechnen. Für den Zähler überlegen wir uns, dass um genau k markierte Kugeln zu ziehen, wir k verschiedene Kugeln aus den $|M|$ markierten und $n-k$ verschiedene Kugeln aus den $|K \setminus M|$ unmarkierten ziehen müssen. Wieder gibt es $n!$ verschiedene Reihenfolgen, in denen diese Kugeln gezogen werden können. Also gilt

$$|\{X = k\}| = n!\binom{r}{k}\binom{|K \setminus M|}{n-k}.$$

Setzen wir $m := |K|$ und $r := |M|$, so folgt damit

$$\mathbb{P}_X[\{k\}] = \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}}.$$

Hier ist also die konkrete Anzahl der markierten Kugeln und nicht nur ihr Anteil relevant! Dafür erhalten wir die selbe Wahrscheinlichkeitsverteilung, egal ob wir die Reihenfolge der gezogenen Kugeln berücksichtigen oder nicht.

Die dadurch definierte Wahrscheinlichkeitsverteilung $\mathrm{H}(m, r, n)$ auf $\mathcal{P}(\{0, \dots, n\})$ mit Zähldichte $h_{m,r,n}(k) := \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}}$ wird *hypergeometrische Verteilung* mit Parametern m, r, und n genannt.

POISSONVERTEILUNG

Aus der Binomialverteilung lässt sich eine weitere Verteilung ableiten, die die Häufigkeit *seltener* zufälliger Vorfälle beschreibt wie zum Beispiel Unfälle, Naturkatastrophen, (damit verbundene) Schadensmeldungen bei Versicherungen, oder dem Zerfall radioaktiver Teilchen. Letzteres ist am einfachsten zu modellieren, daher betrachten wir diese Situation. Angenommen, wir haben eine Laborprobe mit radioaktivem Material, die wir für eine Zeitspanne – die wir ohne Beschränkung der Allgemeinheit als Intervall $[0, 1]$ annehmen – beobachten. Was ist die Wahrscheinlichkeit, in dieser Zeit k radioaktive Zerfälle zu messen? Dazu zerlegen wir das Intervall in n gleichlange Teilintervalle $(\frac{k-1}{n}, \frac{k}{n}]$, $k = 1, \dots, n$, die wir separat betrachten. Wir machen nun folgende intuitive Modellannahmen:

(i) *In jedem Teilintervall findet höchstens ein Zerfall statt.* Dies ist plausibel, wenn die Intervalle klein genug und die Zerfälle selten genug sind.

(ii) *Die Wahrscheinlichkeit, dass ein Zerfall stattfindet, ist proportional zur Länge des Teilintervalls.* Dies ist plausibel, wenn der Zerfall prinzipiell zu jedem Zeitpunkt gleich wahrscheinlich ist. (Dies entspricht intuitiv einer Gleichverteilung der Zerfallszeitpunkte, die wir aber wie im letzten Kapitel beschrieben nicht so einfach direkt verwenden können!)

(iii) *Was in jedem Teilintervall passiert, hängt nicht von den anderen Teilintervallen ab.* Das ist plausibel, wenn die einzelnen Zerfälle (kausal) unabhängig voneinander sind.

Dies entspricht aber genau dem Urnenmodell mit Zurücklegen: Für jedes der n Teilintervalle der Länge $\frac{1}{n}$ ziehen wir blind eine Kugel, die mit Wahrscheinlichkeit $p = \frac{\lambda}{n}$ für einen Proportionalitätsfaktor $\lambda > 0$ (der von dem speziellen radioaktiven Material und der Probengröße abhängt) als „Zerfall“ markiert wurde. Die Anzahl der gemessenen Zerfälle ist also binomialverteilt mit Zähldichte $b_{n,\frac{\lambda}{n}}$.

Da die Einteilung in Teilintervalle willkürlich war, lassen wir nun $\frac{1}{n} \to 0$, d. h. $n \to \infty$, gehen.

Lemma 2.7. *Für alle $\lambda > 0$ und $k \in \mathbb{N} \cup \{0\}$ gilt*

$$\lim_{n\to\infty} b_{n,\frac{\lambda}{n}}(k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Beweis. Nach Definition ist

$$\begin{aligned} b_{n,\frac{\lambda}{n}}(k) &= \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \frac{n(n-1)\cdots(n-k+1)}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \to \frac{\lambda^k}{k!} e^{-\lambda} \end{aligned}$$

für $n \to \infty$, da

- $n(n-1)\cdots(n-k+1) = n^k + O(n^{k-1})$ ist und daher der zweite Bruch gegen 1 geht;
- bekanntermaßen $(1+\frac{x}{n})^n \to e^x$ geht;
- die letzte Klammer und damit der gesamte letzte Term gegen 1 geht. □

Bleibt noch nachzuweisen, dass dies tatsächlich eine Zähldichte ist und damit eine Wahrscheinlichkeitsverteilung definiert (da wir diesmal keine Zufallsvariable für die Konstruktion verwendet haben, mit der dies nach Satz 2.3 immer der Fall ist).

Satz 2.8 (Poissonverteilung). *Für alle $\lambda > 0$ wird durch*

$$p_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, 2, \ldots,$$

eine Zähldichte definiert. Die zugehörige Wahrscheinlichkeitsverteilung $\mathcal{P}_\lambda$ auf $\mathcal{P}(\{0, 1, 2, \ldots\})$ wird Poissonverteilung *mit Parameter (oder* Intensität*) λ genannt.*

Beweis. Wegen $\lambda > 0$ und $e^x > 0$ für alle $x \in \mathbb{R}$ ist stets $p_\lambda(k) > 0$. Weiter gilt

$$\sum_{k=0}^{\infty} p_\lambda(k) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.$$

Damit folgt die Behauptung aus Satz 1.11. □

3 ERWARTUNGSWERT UND VARIANZ

Wir beschäftigen uns nun mit *numerischen* diskreten Zufallsvariablen, d. h. solchen, für die der Wertebereich S eine (höchstens abzählbare) Teilmenge von $\mathbb{R}$ ist, und möchten diese anhand von „Kenngrößen“ vergleichen. Angenommen jemand bietet uns zwei Würfelspiele an:

(i) Es wird ein Würfel gewürfelt; bei gerader Augenzahl zahlen wir ein Euro, bei ungerader Augenzahl bekommen wir einen Euro.

(ii) Es wird ein Würfel gewürfelt; wir müssen einen Einsatz von drei Euro zahlen, bekommen dann aber für jedes Auge einen Euro.

Welches Spiel würden wir wählen? Beide Spiele können wir als Zufallsvariable X bzw. Y, modellieren, die die Augenzahl $\omega \in \Omega := \{1, 2, 3, 4, 5, 6\}$ auf den Gewinn $X(\omega), Y(\omega) \in S := \mathbb{Z}$ (negativer Gewinn heißt dabei, dass wir zahlen müssen) abbildet. Wir fragen uns dann in welchem Spiel wir

(i) einen höheren Gewinn erwarten können;

(ii) die Schwankung des erwarteten Gewinns höher ist.

Ersteres wird durch den *Erwartungswert*, Letzteres durch die *Varianz* der entsprechenden Zufallsvariable beschrieben.

3.1 ERWARTUNGSWERT DISKRETER ZUFALLSVARIABLEN

Sei also $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $S \subset \mathbb{R}$ abzählbar, und X eine diskrete Zufallsvariable von $(\Omega, \mathcal{A})$ nach $(S, \mathcal{P}(S))$.

Definition 3.1 (Erwartungswert). Ist die Summe

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

wohldefiniert, so nennt man $\mathbb{E}[X]$ *Erwartungswert* von X (bezüglich $\mathbb{P}$).

Dies ist insbesondere dann der Fall, wenn gilt

$$\sum_{s\in S} |s| \cdot \mathbb{P}[X = s] < \infty;$$

wir schreiben dafür kurz $X \in \ell^1(\mathbb{P})$ (oder nur ℓ^1, falls die Wahrscheinlichkeitsverteilung aus dem Kontext klar ist).

Der Erwartungswert ist also wohldefiniert, falls die Reihe absolut konvergiert (in diesem Fall unabhängig von der Reihenfolge). Ist $X \notin \ell^1$, so kann der Erwartungswert immer noch wohldefiniert sein, wenn $X(\omega) \geq 0$ für alle $\omega \in \Omega$ gilt; in diesem Fall ist $\mathbb{E}[X] = \infty$, wieder unabhängig von der Reihenfolge.

Eine wichtige Beobachtung ist, dass die Definition des Erwartungswerts nicht vom zugrundeliegenden Wahrscheinlichkeitsraum abhängt; der Erwartungswert ist also (nur) *verteilungsabhängig*.

Folgerung 3.2. *Seien X, Y zwei diskrete Zufallsvariablen mit Wertebereich S. Gilt $X \sim Y$ (d. h $\mathbb{P}_X = \mathbb{P}_Y$), dann ist $\mathbb{E}[X] = \mathbb{E}[Y]$.*

Wir betrachten nun einige einfache aber nützliche Beispiele.

Beispiel 3.3 (Gleichverteilungen). Sei $S = \{s_1, \dots, s_n\} \subset \mathbb{R}$ endlich und $X \sim \mathcal{U}(S)$ gleichverteilt. Dann ist

$$\mathbb{E}[X] = \sum_{s\in S} s \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} s_n,$$

d. h. das arithmetische Mittel (Mittelwert) der s_i.

Beispiel 3.4 (Bernoulli-Verteilung). Sei $X \sim B_p$ Bernoulli-verteilt mit Parameter $p \in (0, 1)$.

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

Beispiel 3.5 (Poisson-Verteilung). Sei $X \sim \mathcal{P}_\lambda$ Poisson-verteilt mit Parameter $\lambda > 0$. Dann ist

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

Der Parameter λ gibt also die zu erwartende Anzahl der Ereignisse pro Zeiteinheit an.

Beispiel 3.6 (konstante Zufallsvariablen). Sei $c \in \mathbb{R}$ beliebig. Wir identifizieren c mit

der konstanten Zufallsvariable $X_c : \Omega \to \{c\}$, $X(\omega) = c$ für alle $\omega \in \Omega$. Dann ist

$$\mathbb{E}[c] = \mathbb{E}[X_c] = c \cdot \mathbb{P}[\Omega] = c.$$

Beispiel 3.7 (Indikatorfunktion). Sei $A \in \mathcal{A}$ beliebig. Dann ist

$$\mathbb{E}[\mathbb{1}_A] = 1 \cdot \mathbb{P}[X = 1] + 0 \cdot \mathbb{P}[X = 0] = \mathbb{P}[X = 1] = \mathbb{P}[A].$$

Letzteres erlaubt, Wahrscheinlichkeiten mit Hilfe von Erwartungswerten zu berechnen. Dafür leiten wir nun fundamentale Rechenregeln her.

Satz 3.8 (Transformationssatz). *Sei X eine diskrete Zufallsvariable von $(\Omega, \mathcal{A}, \mathbb{P})$ nach $(S, \mathcal{P}(S))$ und $g : S \to \mathbb{R}$ eine beliebige Funktion. Dann ist auch*

$$g(X) := g \circ X : \Omega \to g(S) \subset \mathbb{R}$$

eine diskrete Zufallsvariable mit Erwartungswert

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

falls diese Summe wohldefiniert ist.

Beweis. Zunächst gilt für alle $z \in g(S)$

$$\{g(X) = z\} = \bigcup_{s \in g^{-1}(z)} \{X = s\} \in \mathcal{A}, \tag{3.1}$$

da die Vereinigung höchstens abzählbar ist und $\mathcal{A}$ eine σ-Algebra ist. Also ist $g(X)$ wieder eine Zufallsvariable. Weiter ist mit S auch $g(S) = \{g(s) \mid s \in S\}$ abzählbar und damit $g(X)$ eine diskrete Zufallsvariable.

Außerdem ist die Vereinigung in (3.1) disjunkt, so dass aus der σ-Additivität von $\mathbb{P}$ folgt

$$\begin{aligned}\mathbb{E}[g(X)] = \sum_{z \in g(S)} z \cdot \mathbb{P}[g(X) = z] &= \sum_{z \in g(S)} z \sum_{s \in g^{-1}(z)} \mathbb{P}[X = s] \\ &= \sum_{z \in g(S)} \sum_{s \in g^{-1}(z)} g(s) \cdot \mathbb{P}[X = s] \\ &= \sum_{s \in S} g(s) \cdot \mathbb{P}[X = s].\end{aligned}$$

□

Damit kann man nun einige nützliche Eigenschaften beweisen.

Folgerung 3.9. *Es ist $X \in \ell^1$ genau dann, wenn gilt $\mathbb{E}[|X|] < \infty$.*

Beweis. Wende Satz 3.8 an auf $g(x) = |x|$. □

Folgerung 3.10. *Ist Ω abzählbar, so gilt*

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

falls die Summe wohldefiniert ist. Ist Ω endlich und $\mathbb{P} = \mathcal{U}(\Omega)$ die Gleichverteilung, so gilt insbesondere

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

Beweis. Für alle Zufallsvariablen gilt $X = X \circ \mathrm{Id}_\Omega$ und damit nach Satz 3.8 (mit X an Stelle von g)

$$\mathbb{E}[X] = \mathbb{E}[X \circ \mathrm{Id}_\Omega] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}[\mathrm{Id}_\Omega = \omega] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}[\{\omega\}].$$ □

Wir kommen nun zur zentralen Eigenschaft des Erwartungswerts: der Linearität.

Satz 3.11. *Seien $X : \Omega \to S_X \subset \mathbb{R}$ und $Y : \Omega \to S_Y \subset \mathbb{R}$ diskrete Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit $X, Y \in \ell^1$. Dann gilt für alle $\alpha, \beta \in \mathbb{R}$ auch $\alpha X + \beta Y \in \ell^1$ und*

$$\mathbb{E}[\alpha X + \beta Y] = \alpha\, \mathbb{E}[X] + \beta\, \mathbb{E}[Y].$$

Beweis. Wir betrachten

$$(X, Y) : \Omega \to S_X \times S_Y, \qquad \omega \mapsto (X(\omega), Y(\omega)).$$

Da X, Y diskrete Zufallsvariablen sind, ist auch (X, Y) eine diskrete Zufallsvariable von $(\Omega, \mathcal{A})$ nach $(S_X \times S_Y, \mathcal{P}(S_X \times S_Y))$. Wir können also Satz 3.8 anwenden auf

$$g : S_X \times S_Y \to \mathbb{R}, \qquad (s, t) \mapsto \alpha s + \beta t,$$

und erhalten (falls alle Reihen absolut konvergieren)

$$\begin{aligned}
\mathbb{E}[\alpha X + \beta Y] = \mathbb{E}[g(X, Y)] &= \sum_{(s,t) \in S_X \times S_Y} g(s, t) \cdot \mathbb{P}[(X, Y) = (s, t)] \\
&= \sum_{s \in S_X} \sum_{t \in S_Y} (\alpha s + \beta t) \cdot \mathbb{P}[X = s, Y = t] \\
&= \alpha \sum_{s \in S_X} s \sum_{t \in S_Y} \mathbb{P}[X = s, Y = t] + \beta \sum_{t \in S_Y} t \sum_{s \in S_X} \mathbb{P}[X = s, Y = t] \\
&= \alpha \sum_{s \in S_X} s \cdot \mathbb{P}[X = s] + \beta \sum_{t \in S_Y} t \cdot \mathbb{P}[Y = t] \\
&= \alpha\, \mathbb{E}[X] + \beta\, \mathbb{E}[Y],
\end{aligned}$$

wobei wir im vorletzten Schritt die σ-Additivität von $\mathbb{P}$ benutzt haben.

Es bleibt zu zeigen, dass für $X, Y \in \ell^1$ die Summe wohldefiniert (und damit die obige Rechnung gerechtfertigt) ist. Aber das folgt analog aus

$$\begin{aligned}\sum_{s\in S_X}\sum_{t\in S_Y} |\alpha s + \beta t| \cdot \mathbb{P}[X = s, Y = t] &\leq |\alpha| \sum_{s\in S_X} |s| \cdot \mathbb{P}[X = s] + |\beta| \sum_{t\in S_Y} |t| \cdot \mathbb{P}[Y = t] \\ &= |\alpha|\,\mathbb{E}[|X|] + |\beta|\,\mathbb{E}[|Y|] < \infty\end{aligned}$$

nach Voraussetzung. □

Daraus folgt direkt eine weitere fundamentale Eigenschaft des Erwartungswerts: die Monotonie. Wir zeigen zuerst einen einfachen Spezialfall.

Lemma 3.12. *Sei $X : \Omega \to S$ eine diskrete Zufallsvariable mit $X(\omega) \geq 0$ für alle $\omega \in \Omega$. Dann gilt*

(i) $\mathbb{E}[X] \geq 0$;

(ii) $\mathbb{E}[X] = 0$ *genau dann, wenn* $\mathbb{P}[X = 0] = 1$ *gilt.*

Beweis. Aus $X(\omega) \geq 0$ folgt $\mathbb{P}[X = s] = 0$ für alle $s < 0$. Damit enthält

$$\mathbb{E}[X] = \sum_{s\in S} s \cdot \mathbb{P}[X = s]$$

nur nichtnegative Summanden und ist daher selber nichtnegativ. Die zweite Aussage folgt analog aus $\mathbb{P}[X = s] = 0$ für alle $s \neq 0$ wegen der σ-Additivität. □

Folgerung 3.13. *Seien $X : \Omega \to S_X \subset \mathbb{R}$ und $Y : \Omega \to S_Y \subset \mathbb{R}$ diskrete Zufallsvariablen auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit $X, Y \in \ell^1$. Gilt $X(\omega) \geq Y(\omega)$ für alle $\omega \in \Omega$, dann gilt auch*

$$\mathbb{E}[X] \geq \mathbb{E}[Y].$$

Beweis. Wir wenden Satz 3.11 an mit $\alpha = 1$ und $\beta = -1$ und erhalten $X - Y \in \ell^1$ sowie $X - Y \geq 0$ und daher

$$\mathbb{E}[X] - \mathbb{E}[Y] = \mathbb{E}[X - Y] \geq 0$$

nach Lemma 3.12 (i). □

Die Linearität liefert auch einen eleganten Beweis für das Einschluss-Ausschluss-Prinzip.

Folgerung 3.14 (Einschluss-Ausschluss-Prinzip). *Für alle $n \in \mathbb{N}$ und $A_1, \ldots, A_n \in \mathcal{A}$ gilt*

$$\mathbb{P}\left[\bigcup_{k=1}^{n} A_k\right] = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1\leq i_1<\cdots<i_k\leq n} \mathbb{P}\left[\bigcap_{j=1}^{k} A_{i_j}\right].$$

Beweis. Wir verwenden Beispiel 3.7 zusammen mit Satz 3.11. Aus den de Morganschen Gesetzen und den Rechenregeln für Indikatorfunktionen folgt direkt mit Hilfe der multinomischen Formel

$$\begin{aligned}\mathbb{P}\left[\left(\bigcup_{k=1}^{n} A_k\right)^c\right] &= \mathbb{P}\left[\left(\bigcap_{k=1}^{n} A_k^c\right)\right] = \mathbb{E}\left[\mathbb{1}_{\bigcap_{k=1}^{n} A_k^c}\right] = \mathbb{E}\left[\prod k = 1^n \mathbb{1}_{A_k^c}\right] = \mathbb{E}\left[\prod k = 1^n (1 - \mathbb{1}_{A_k})\right] \\ &= \sum_{k=0}^{n} (-1)^k \sum_{1 \le i_1 < \cdots < i_k \le n} \mathbb{E}\left[\Pi_{j=1}^{k} \mathbb{1}_{A_{i_j}}\right] \\ &= \sum_{k=0}^{n} (-1)^k \sum_{1 \le i_1 < \cdots < i_k \le n} \mathbb{P}\left[\bigcap_{j=1}^{k} A_{i_j}\right].\end{aligned}$$

Also ist

$$\begin{aligned}\mathbb{P}\left[\bigcup_{k=1}^{n} A_k\right] = 1 - \mathbb{P}\left[\left(\bigcup_{k=1}^{n} A_k\right)^c\right] &= 1 - \sum_{k=0}^{n} (-1)^k \sum_{1 \le i_1 < \cdots < i_k \le n} \mathbb{P}\left[\bigcap_{j=1}^{k} A_{i_j}\right] \\ &= \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < \cdots < i_k \le n} \mathbb{P}\left[\bigcap_{j=1}^{k} A_{i_j}\right]. \qquad \square\end{aligned}$$

Mit Hilfe der Linearität können wir bequem Erwartungswerte von Verteilungen für Urnenmodelle ausrechnen.

Beispiel 3.15 (Binomialverteilung). Wir betrachten wieder die Situation, dass aus einer Urne mit m Kugeln, von denen r markiert sind, n Kugeln mit Zurücklegen gezogen werden. Die Zufallsvariable X, die die Anzahl der markierten gezogenen Kugeln angibt, ist binomialverteilt mit Zähldichte

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

für $p := \frac{m}{r}$. Mit den Rechenregeln für Binomialkoeffizienten und einiger Mühe kann man ausrechnen, dass gilt

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \cdot b_{n,p}(k) = np.$$

Es geht aber auch leichter: Wir betrachten jeden Zug separat. Sei $X_i : \Omega \to \{0,1\}$ die Zufallsvariable die angibt, im i-ten Zug eine markierte Kugel (aus M) zu ziehen. Dann gilt

$$\mathbb{P}[X_i = 1] = \frac{|\{\omega \in \Omega \mid \omega_i \in M\}|}{m^n} = \frac{m \cdots m \cdot r \cdot m \cdots m}{m^n} = \frac{r}{m} = p,$$

da wir für den $i$te Zug r markierte Kugeln und für alle anderen n (markierte oder unmarkierte) zur Auswahl haben. Daraus folgt auch $\mathbb{P}[X_i = 0] = 1 - p$. Für alle

$i = 1, \dots, n$ ist X_i also Bernoulli-verteilt mit Parameter p. Die Anzahl der gezogenen Kugeln ist dann also gegeben durch

$$X = X_1 + \cdots + X_n,$$

so dass nach Satz 3.11 und Beispiel 3.4 gilt

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = np.$$

Dies gilt, obwohl die Binomial*verteilung* nicht die Summe von Bernoulli-Verteilungen ist!

Beispiel 3.16 (Hypergeometrische Verteilung). Wir betrachten nun die Situation *ohne* Zurücklegen. Hier ist die Zufallsvariable X hypergeometrisch verteilt mit Zähldichte

$$h_{m,r,n}(k) = \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}}.$$

Wieder rechnet man mit einigem Aufwand nach, dass gilt

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \cdot h_{m,r,n}(k) = n\frac{r}{m} = np$$

für $p := \frac{r}{m}$. Auch hier kommt man mit etwas Nachdenken schneller zum Zug: Wir haben bereits gesehen, dass die Reihenfolge der Züge keine Rolle spielt; die Wahrscheinlichkeit, im i-ten Zug eine markierte Kugel zu erhalten (egal was in den restlichen Zügen passiert!), ist also für alle Züge die selbe – insbesondere für den ersten Zug, wo Zurücklegen keinen Unterscheid macht. (Dies ist kein rigoroser Beweis, kann aber analog zu Beispiel 3.15 gerechtfertigt werden.) Es gilt also wieder $\mathbb{P}[X_i = 1] = p$ und damit

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = np,$$

obwohl die Verteilungen sehr verschieden sind.

Wir werden die Frage, wie man die Verteilung von Summen von Zufallsvariablen aus den einzelnen Verteilungen erhält, später noch genauer untersuchen.

3.2 VARIANZ UND KOVARIANZ

Wir kommen nun zur zweiten Kenngröße, die für eine Zufallsvariable X die zu erwartende Abweichung $X - \mathbb{E}[X]$ vom Mittelwert $\mathbb{E}[X]$ angibt. Da uns das Vorzeichen der Abweichung hier nicht interessiert, betrachten wir den Betrag oder – einfacher – das Quadrat der Abweichung.

Sei wieder $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $S \subset \mathbb{R}$ abzählbar, und X eine diskrete Zufallsvariable von $(\Omega, \mathcal{A})$ nach $(S, \mathcal{P}(S))$. Dann führt dies direkt auf die folgenden Definition.

Definition 3.17 (Varianz). Sei X eine Zufallsvariable mit Erwartungswert $\mathbb{E}[X] < \infty$. Ist

$$\mathbb{V}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_{s \in S} (s - \mathbb{E}[X])^2 p_X(s)$$

wohldefiniert, so nennt man $\mathbb{V}[X]$ *Varianz* von X (bezüglich $\mathbb{P}$).

Genau wie der Erwartungswert ist also die Varianz (nur) verteilungsabhängig; $X \sim Y$ impliziert daher $\mathbb{V}[X] = \mathbb{V}[Y]$. (Verschiedene Verteilungen können aber zur selben Varianz führen.)

Aus der Linearität des Erwartungswerts folgt sofort die fundamentale Gleichung

$$\begin{aligned} \mathbb{E}[(X - \mathbb{E}[X])^2] &= \mathbb{E}[X^2] - 2\,\mathbb{E}[X \cdot \mathbb{E}[X]] + \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2; \end{aligned} \tag{3.2}$$

die Varianz ist also endlich genau dann, wenn $\mathbb{E}[X^2] < \infty$ ist. Wir schreiben in dem Fall auch kurz $X \in \ell^2$. (Wir werden später sehen, dass $X \in \ell^2$ bereits $X \in \ell^1$ impliziert.)

Wir betrachten mit Hilfe dieser Darstellung nun wieder einfache Beispiele.

Beispiel 3.18 (Gleichverteilungen). Sei $S = \{s_1, \ldots, s_n\} \subset \mathbb{R}$ endlich und $X \sim \mathcal{U}(S)$ gleichverteilt. Dann ist $\mathbb{E}[X] = \frac{1}{n} \sum_{i=1}^n s_n =: \bar{x}$ und damit

$$\mathbb{V}[X] = \frac{1}{n} \sum_{i=1}^n (s_n - \bar{x})^2$$

d. h. die mittlere quadratische Abweichung vom Mittelwert.

Beispiel 3.19 (Bernoulli-Verteilung). Sei X Bernoulli-verteilt mit Parameter $p \in (0,1)$. Dann ist $\mathbb{E}[X^2] = \mathbb{E}[X] = p$ wegen $X(\omega) \in \{0,1\}$ und damit

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

Beispiel 3.20 (Poisson-Verteilung). Sei $X \sim \mathcal{P}_\lambda$ Poisson-verteilt mit Parameter $\lambda > 0$.

Dann ist $\mathbb{E}[X] = \lambda$, und aus dem Transformationssatz Satz 3.8 folgt

$$\begin{aligned}\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 &= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2 \\ &= \lambda \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} - \lambda^2 \\ &= \lambda \sum_{k=0}^{\infty} (k+1) \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2 \\ &= \lambda \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} + \lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2 \\ &= \lambda \cdot \lambda + \lambda \cdot 1 - \lambda^2 = \lambda.\end{aligned}$$

Der Parameter λ gibt also nicht nur die zu erwartende Anzahl der Ereignisse pro Zeiteinheit sondern auch ihre zu erwartende Streuung an.

Aus (3.2) erhält man auch eine analoge Rechenregel für die Varianz.

Lemma 3.21. *Sei $X \in \ell^2$ und $a, b \in \mathbb{R}$. Dann gilt*

$$\mathbb{V}[aX + b] = a^2 \, \mathbb{V}[X].$$

Beweis. Aus (3.2) folgt zusammen mit der Linearität des Erwartungswerts direkt

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

□

Da Quadrate stets nicht-negativ sind, erhalten wir aus der ursprünglichen Definition der Varianz mit Lemma 3.12 die folgenden Eigenschaften.

Lemma 3.22. *Sei $X \in \ell^2$. Dann gilt stets*

(i) $\mathbb{V}[X] \geq 0$;

(ii) $\mathbb{V}[X] = 0$ *genau dann, wenn* $\mathbb{P}[X = \mathbb{E}[X]] = 1$ *gilt (d. h.* $X = \mathbb{E}[X]$ *konstant ist).*

Ist X nicht deterministisch, erhalten wir die folgende sehr nützliche Abschätzung.

Satz 3.23 (Chebyshev-Ungleichung). *Sei $X \in \ell^2$ und $\varepsilon > 0$. Dann gilt*

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \varepsilon\right] \leq \frac{\mathbb{V}[X]}{\varepsilon^2}.$$

Beweis. Sei $\varepsilon > 0$. Dann gilt für alle $\omega \in \Omega$

$$\frac{(X(\omega) - \mathbb{E}[X])^2}{\varepsilon^2} \geq \begin{cases} 1 & \text{falls } |X(\omega) - \mathbb{E}[X]| \geq \varepsilon, \\ 0 & \text{sonst,} \end{cases}$$
$$= \mathbb{1}_{\{|X-\mathbb{E}[X]|\geq\varepsilon\}}(\omega).$$

Mit Beispiel 3.7 folgt daraus zusammen mit der Monotonie und Linearität des Erwartungswerts

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \varepsilon\right] = \mathbb{E}\left[\mathbb{1}_{\{|X-\mathbb{E}[X]|\geq\varepsilon\}}\right] \leq \mathbb{E}[\varepsilon^{-2}(X - \mathbb{E}[X])^2] = \frac{\mathbb{V}[X]}{\varepsilon^2}. \qquad \square$$

Lemma 3.22 (i) erlaubt auch die folgende Definition, die manchmal handlicher ist als die Varianz.

Definition 3.24 (Standardabweichung). Die *Standardabweichung* von $X \in \ell^2$ ist definiert als $\sigma(X) := \sqrt{\mathbb{V}[X]}$.

Wir möchten nun, analog zum Erwartungswert, eine Summenformel für Varianzen herleiten. Diese kann im Allgemeinen nicht linear sein; ist $X \in \ell^2$ mit $\mathbb{V}[X] > 0$, so folgt aus Lemma 3.21 sofort

$$\mathbb{V}[X + X] = \mathbb{V}[2X] = 4\,\mathbb{V}[X] > \mathbb{V}[X] + \mathbb{V}[X],$$
$$\mathbb{V}[X - X] = \mathbb{V}[0] = 0 < \mathbb{V}[X] + \mathbb{V}[X] = \mathbb{V}[X] + \mathbb{V}[-X].$$

Wir müssen daher mögliche *Abhängigkeiten* zwischen den zu summierenden Zufallsvariablen berücksichtigen. Dies führt auf die folgende Definition.

Definition 3.25 (Kovarianz). Seien $X, Y \in \ell^2$. Dann ist die *Kovarianz* von X und Y definiert als

$$\mathrm{Cov}[X, Y] := \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

Gilt $\mathrm{Cov}[X, Y] = 0$, dann nennt man X und Y *unkorreliert*.

Aus $X, Y \in \ell^2$ folgt nach der Cauchy-Schwarz-Ungleichung für Reihen $XY \in \ell^1$, so dass die Kovarianz unter der Voraussetzung wohldefiniert ist. Direkt aus der Definition erhalten wir die folgende nützliche Charakterisierung unkorrelierter Zufallsvariablen.

Folgerung 3.26. *Zwei Zufallsvariablen $X, Y \in \ell^2$ sind unkorreliert genau dann, wenn gilt*

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

Beweis. Dies folgt sofort aus

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

□

Für korrelierte Zufallsvariablen ist die *Richtung* der Korrelation interessant.

Definition 3.27 (Korrelationskoeffizient). Sind $\mathbb{V}[X], \mathbb{V}[Y] > 0$, dann heißt

$$\rho(X, Y) := \frac{\operatorname{Cov}[X, Y]}{\sqrt{\mathbb{V}[X]\,\mathbb{V}[Y]}}$$

Korrelationskoeffizient von X und Y. Man nennt X und Y

- positiv korreliert, wenn gilt $\rho(X, Y) > 0$;
- negativ korreliert, wenn gilt $\rho(X, Y) < 0$.

Die Kovarianz hat einige nützliche Eigenschaften, die aus der linearen Algebra bekannt sein sollten.

Lemma 3.28. *Die Kovarianz ist*

(i) symmetrisch, d. h. $\operatorname{Cov}[X, Y] = \operatorname{Cov}[Y, X]$ *für alle* $X, Y \in \ell^2$*;*

(ii) bilinear, d. h. für alle $X, Y, Z \in \ell^2$ *und* $\alpha, \beta \in \mathbb{R}$ *gilt*

$$\operatorname{Cov}[X, \alpha Y + \beta Z] = \alpha \operatorname{Cov}[X, Y] + \beta \operatorname{Cov}[X, Z];$$

(iii) positiv semidefinit, d. h.

$$\operatorname{Cov}[X, X] = \mathbb{V}[X] \geq 0 \qquad \text{für alle } X \in \ell^2.$$

Beweis. *(i)* folgt direkt aus der Definition.

(ii) folgt aus der Linearität des Erwartungswerts:

$$\begin{aligned}\operatorname{Cov}[X, \alpha Y + \beta Z] &= \mathbb{E}\left[(X - \mathbb{E}[X])(\alpha Y + \beta Z - \mathbb{E}[\alpha Y + \beta Z])\right] \\ &= \mathbb{E}\left[(X - \mathbb{E}[X])\left(\alpha(Y - \mathbb{E}[Y]) + \beta(Z - \mathbb{E}[Z])\right)\right] \\ &= \alpha\,\mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] + \beta\,\mathbb{E}\left[(X - \mathbb{E}[X])(Z - \mathbb{E}[Z])\right] \\ &= \alpha \operatorname{Cov}[X, Y] + \beta \operatorname{Cov}[X, Z];\end{aligned}$$

(iii) folgt wieder aus der Definition zusammem mit Lemma 3.22 (ii). □

Wie ein Skalarprodukt erfüllt daher auch die Kovarianz eine Cauchy-Schwarz-Ungleichung.

Lemma 3.29 (Cauchy-Schwarz-Ungleichung). *Für alle $X, Y \in \ell^2$ gilt*

$$|\operatorname{Cov}[X, Y]| \leq \sqrt{\mathbb{V}[X]\,\mathbb{V}[Y]}.$$

Beweis. Ist $\mathbb{V}[X] = 0$, so folgt aus Lemma 3.22 (ii) dass $X = \mathbb{E}[X]$ mit Wahrscheinlichkeit 1 und damit auch $\operatorname{Cov}[X, Y] = \mathbb{E}[0 \cdot (Y - \mathbb{E}[Y])] = 0$ gilt. Damit gilt die Ungleichung trivialerweise.

Wir können also annehmen, dass $\mathbb{V}[X] > 0$ ist. Wir setzen nun

$$a := \frac{\operatorname{Cov}[X, Y]}{\mathbb{V}[X]}.$$

Dann folgt aus Lemma 3.28 durch quadratisches Ergänzen

$$\begin{aligned}
0 &\leq \operatorname{Cov}[Y - aX, Y - aX] \\
&= a^2 \operatorname{Cov}[X, X] - 2a \operatorname{Cov}[X, Y] + \operatorname{Cov}[Y, Y] \\
&= \left(a\sqrt{\mathbb{V}[X]} - \frac{\operatorname{Cov}[X, Y]}{\sqrt{\mathbb{V}[X]}}\right)^2 + \mathbb{V}[Y] - \frac{\operatorname{Cov}[X, Y]^2}{\mathbb{V}[X]} \\
&= \mathbb{V}[Y] - \frac{\operatorname{Cov}[X, Y]^2}{\mathbb{V}[X]}
\end{aligned}$$

und damit nach Wurzelziehen die Behauptung. □

Auf ähnliche Weise folgt aus Lemma 3.28 auch die gesuchte Summenformel: Für $X, Y \in \ell^2$ gilt

$$\begin{aligned}
\mathbb{V}[X + Y] = \operatorname{Cov}[X + Y, X + Y] &= \operatorname{Cov}[X, X] + 2\operatorname{Cov}[X, Y] + \operatorname{Cov}[Y, Y] \\
&= \mathbb{V}[X] + \mathbb{V}[Y] + 2\operatorname{Cov}[X, Y].
\end{aligned}$$

Allgemein gilt die folgende Formel.

Satz 3.30. *Für alle $X_1, \ldots, X_n \in \ell^2$ gilt*

$$\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}[X_i] + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \operatorname{Cov}[X_i, X_j].$$

Sind die X_i paarweise unkorreliert, gilt insbesondere die Bienaymé-Formel

$$\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}[X_i].$$

Beweis. Aus der Bilinearität und Symmetrie der Kovarianz folgt

$$\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \operatorname{Cov}\left[\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right] = \sum_{i,j=1}^{n} \operatorname{Cov}[X_i, X_j] = \sum_{i=1}^{n} \mathbb{V}[X_i] + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \operatorname{Cov}[X_i, X_j]. \qquad \square$$

Beispiel 3.31 (Binomialverteilung). Wie in Beispiel 3.15 betrachten wir wieder eine binomialverteilte Zufallsvariable $X \sim \operatorname{Bin}(n, p)$ als Summe von n Bernoulli-verteilten Zufallsvariablen X_i. Die X_i sind paarweise unkorreliert, da ein Zug nicht von dem Ergebnis eines anderen abhängt: Seien X_i, X_j mit $i \neq j$. Dann gilt

$$X_i X_j(\omega) = \begin{cases} 1 & \text{falls } X_i(\omega) = X_j(\omega) = 1, \\ 0 & \text{sonst.} \end{cases}$$

und damit

$$\mathbb{E}[X_i X_j] = 1 \cdot \mathbb{P}[X_i = 1, X_j = 1] = \mathbb{P}[X_i = 1]\, \mathbb{P}[X_j = 1] = p^2 = \mathbb{E}[X_i]\, \mathbb{E}[X_j].$$

(Dies ist eine Modell-Annahme!) Also folgt aus der Bienaymé-Formel

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

nach Beispiel 3.4.

3.3 GESETZ DER GROSSEN ZAHL

Wir haben bereits gesehen, dass man den Erwartungswert sowohl der Binomialverteilung als auch der hypergeometrischen Verteilung als Summe von Bernoulli-Verteilungen berechnen kann; in beiden Fällen gilt

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = np,$$

wenn n die Anzahl der gezogenen Kugeln und X_i die Zufallsvariable „i-te gezogene Kugel ist markiert“ ist. Dies kann man auch andersherum auffassen: für alle $n \in \mathbb{N}$ gilt

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

d. h. der Erwartungswert des *Mittelwerts* der Zufallsvariablen ist gleich dem Erwartungswert der einzelnen Zufallsvariablen.

Was können wir nun über den Mittelwert selber (der dann ja wieder eine Zufallsvariable ist) aussagen? Die Antwort auf solche Fragen geben die *Gesetze der großen Zahl*, die Aussagen machen über den Grenzwert $n \to \infty$ solcher Mittelwerte. Wir betrachten hier ein einfaches Beispiel.

Satz 3.32 (schwaches Gesetz der großen Zahl). *Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $\{X_n\}_{n\in\mathbb{N}} \subset \ell^2$ eine Folge von diskreten Zufallsvariablen. Gilt*

(i) $\operatorname{Cov}[X_i, X_j] = 0$ *für alle $i \neq j$ (d. h. die X_n sind unkorreliert);*

(ii) $\mathbb{V}[X_n] \leq M$ *für alle $n \in \mathbb{N}$ (d. h. die X_n haben beschränkte Varianz),*

dann gilt für alle $\varepsilon > 0$ und $n \in \mathbb{N}$

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| \geq \varepsilon\right] \leq \frac{M}{n\varepsilon^2}.$$

Beweis. Sei $n \in \mathbb{N}$ und $\varepsilon > 0$ beliebig. Aus der Chebyshev-Ungleichung (Satz 3.23) folgt dann

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2}\,\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right].$$

Weiter folgt aus Lemma 3.21 und Satz 3.30 mit Annahme (i)

$$\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\,\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i] \leq \frac{M}{n}$$

nach Annahme (ii) und damit die Behauptung. □

Für $n \to \infty$ geht die Wahrscheinlichkeit, dass der Mittelwert der X_i weiter als ein beliebiges $\varepsilon > 0$ vom Erwartungswert entfernt ist, wie $\frac{1}{n}$ gegen Null. Ein besonderer Spezialfall ist, wenn – wie bei der Binomialverteilung – alle X_i den gleichen Erwartungswert und die gleiche Varianz besitzen.

Folgerung 3.33. *Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $\{X_n\}_{n\in\mathbb{N}} \subset \ell^2$ eine Folge von unkorrelierten diskreten Zufallsvariablen mit $\mathbb{E}[X_n] = m$ und $\mathbb{V}[X_n] = M$ für alle $n \in \mathbb{N}$. Dann gilt*

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - m\right| \geq \varepsilon\right] = 0 \qquad \text{für alle } \varepsilon > 0.$$

Man sagt in diesem Fall auch, die Mittelwerte *konvergieren stochastisch* gegen den Erwartungswert m.

Beachten Sie, dass das schwache Gesetz der Großen Zahl keine Aussage über den Grenzwert $\lim_n \to \infty \frac{1}{n} \sum_{i=1}^n X_i$ selber macht. Dafür braucht man ein *starkes Gesetz der großen Zahl*, welches aufwendiger zu beweisen ist. Den Beweis der folgenden Aussage findet man z. B. in [Krengel 2005, Satz 12.4].

Satz 3.34 (starkes Gesetz der großen Zahl). *Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $\{X_n\}_{n\in\mathbb{N}} \subset \ell^2$ eine Folge von unkorrelierten diskreten Zufallsvariablen mit $\mathbb{E}[X_n] = m$ und $\mathbb{V}[X_n] = M$ für alle $n \in \mathbb{N}$. Dann gilt*

$$\mathbb{P}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = m\right] = 1.$$

Man sagt in diesem Fall auch, die Mittelwerte *konvergieren fast sicher* gegen den Erwartungswert m. (Aus fast sicherer Konvergenz folgt stochastische Konvergenz, aber nicht umgekehrt; siehe z. B. [Krengel 2005, Satz 12.1].) Andere Gesetze der großen Zahl zeigen analoge Aussagen unter anderen, teils schwächeren, Voraussetzungen an die Zufallsvariablen.

Beispiel 3.35 (Schätzen von Erwartungswerten). Angenommen, wir wollen wissen, ob eine gegebene Münze fair ist, d. h. ob der Parameter p einer Bernoulli-verteilten Zufallsvariable X gleich 0.5 ist. Wegen $\mathbb{E}[X] = p$ entspricht dies der Aufgabe, den Erwartungswert zu *schätzen*. Dafür können wir wie folgt vorgehen: Wir werfen die entsprechende Münze n mal und mitteln das Ergebnis, d. h. analog zu Beispiel 3.15 betrachten wir die einzelnen Münzwürfe als (unkorrelierte) Bernoulli-verteilte Zufallsvariablen $X_i, \dots, X_n$ und bilden daraus die Zufallsvariable

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Nach Satz 3.32 gilt dann wegen $\mathbb{E}[\overline{X}_n] = \mathbb{E}[X_i] = p$ und $\mathbb{V}[X_i] = p(1-p) \leq \frac{1}{4}$ für alle $i = 1, \dots, n$

$$\mathbb{P}\left[|\overline{X}_n - p| \geq \varepsilon\right] \leq \frac{1}{4n\varepsilon^2} \to 0 \qquad \text{für alle } \varepsilon > 0,$$

d. h. $\overline{X}_n$ konvergiert stochastisch gegen p. Je öfter wir werfen, desto sicherer können wir also sein, dass der *empirische Mittelwert* $\overline{X}_n$ – obwohl er zufällig ist – mit hoher Wahrscheinlichkeit nicht weit weg von p liegt. (Man nennt daher $\overline{X}_n$ einen *konsistenten Schätzer*.)
Dies kann man wie folgt quantifizieren: Für gegebenes $\varepsilon > 0$ wählt man $n \in \mathbb{N}$ (oder umgekehrt) so, dass z. B. $\frac{1}{4n\varepsilon^2} < 0.05$ ist. Dann ist

$$\mathbb{P}\left[p \in (\overline{X}_n - \varepsilon, \overline{X}_n + \varepsilon)\right] = 1 - \mathbb{P}\left[|\overline{X}_n - p| \geq \varepsilon\right] \geq 1 - 0.05 = 0.95,$$

d. h. der echte Erwartungswert p liegt mit mindestens 95% Wahrscheinlichkeit in dem (zufälligen!) Intervall $(\overline{X}_n - \varepsilon, \overline{X}_n + \varepsilon)$. (Man spricht daher auch von einem *95%-Konfidenzintervall* für p.)

4 BEDINGTE WAHRSCHEINLICHKEITEN UND UNABHÄNGIGKEIT

Eine zentrale Frage in der Stochastik ist, ob und wie verschiedene Ereignisse (und, in Folge, Zufallsvariablen) von einander abhängen. Diese Information ist zentral in der Statistik und kann helfen, um Wahrscheinlichkeitsverteilungen von mehrstufigen Zufallsexperimenten mit bekannter Abhängigkeit zu bestimmen (wie wir es schon naiv für die Urnenmodelle betrachtet haben). Ist bekannt, dass die einzelnen Ereignisse *nicht* voneinander abhängen, so sind stärkere Aussagen möglich.

4.1 BEDINGTE WAHRSCHEINLICHKEITEN

Seien $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{A}$ zwei verschiedene Ereignisse mit Wahrscheinlichkeit $\mathbb{P}[A]$ bzw. $\mathbb{P}[B]$. Angenommen, wir wissen bereits, dass das Ereignis B eintritt. Wie beeinflusst diese Zusatzinformation unsere Einschätzung, ob auch A eintritt? Anschaulich bedeutet dass, das wir nur die Ergebnisse $\omega \in B$ betrachten müssen, für die auch $\omega \in A$ und damit $\omega \in A \cap B$ gilt. Dies motiviert die folgende Definition.

Definition 4.1 (Bedingte Wahrscheinlichkeit). Seien $A, B \in \mathcal{A}$ mit $\mathbb{P}[B] \neq 0$. Dann heißt

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

die *bedingte Wahrscheinlichkeit von A gegeben B*.

Beachten Sie, dass bedingte Wahrscheinlichkeiten gegeben unmögliche Ereignisse nicht wohl-definiert sind. Da Zähler und Nenner immer nichtnegativ sind, ist auch $\mathbb{P}[A \mid B] \geq 0$. Aus Satz 1.8 (iii) folgt wegen $A \cap B \subset B$ weiter $\mathbb{P}[A \mid B] \leq 1$. Außerdem gilt $\mathbb{P}[A \mid B] = 0$, falls entweder $A = \emptyset$ oder $A \cap B = \emptyset$ ist. Von fundamentaler Wichtigkeit für die korrekte Anwendung in der Statistik ist auch im Kopf zu behalten, dass eine hohe bedingte Wahrscheinlichkeit keinerlei Aussage über einen kausalen Zusammenhang der beiden Ergebnisse erlaubt!

Beispiel 4.2 (Gleichverteilung). Angenommen, Ω ist endlich und $\mathbb{P} \sim \mathcal{U}(\Omega)$ die Gleichverteilung auf $\mathcal{P}(\Omega)$. Dann ist für $A, B \subset \Omega$, $B \neq \emptyset$

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\frac{|A \cap B|}{|\Omega|}}{\frac{|B|}{|\Omega|}} = \frac{|A \cap B|}{|B|}.$$

Dies entspricht genau der Motivation für unsere Definition.

Beispiel 4.3. Wir betrachten den einfachen Würfelwurf, d. h. die Gleichverteilung auf $\mathcal{P}(\{1, 2, 3, 4, 5, 6\})$. Angenommen, wir wissen, das die geworfene Augenzahl höchstens 4 ist. Was ist dann die Wahrscheinlichkeit, dass die Augenzahl 3 ist? Dies entspricht der bedingten Wahrscheinlichkeit

$$\mathbb{P}\left[\{3\} \mid \{1, 2, 3, 4\}\right] = \frac{|\{3\}|}{|\{1, 2, 3, 4\}|} = \frac{1}{4}$$

nach Beispiel 4.2. Anschaulich sind also alle unter der Zusatzannahme verbliebenen Ergebnisse gleich wahrscheinlich.

Das folgende Beispiel illustriert die Schwierigkeit im Umgang mit bedingten Wahrscheinlichkeiten in der Anwendung: Es ist nicht immer offensichtlich, *welche* Zusatzinformation genau durch eine Beobachtung eingebracht wird.

Beispiel 4.4. Angenommen, eine Familie hat zwei Kinder. Wir wissen, dass mindestens eins der Kinder ein Mädchen ist. Wie hoch ist die Wahrscheinlichkeit, dass beide Mädchen sind? Der Ergebnisraum ist hier (beachte, dass wir die beiden Kinder unterscheiden müssen)

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

und wir nehmen der Einfachheit halber an, dass jedes Ergebnis gleich wahrscheinlich ist. Die gesuchte bedingte Wahrscheinlichkeit ist nach Beispiel 4.2 dann

$$\mathbb{P}\left[\{MM\} \mid \{MM, MJ, JM\}\right] = \frac{1}{3} > \frac{1}{4} = \mathbb{P}\left[\{MM\}\right].$$

Wir modifizieren die Situation nun leicht. Anstatt nur durch Hörensagen zu wissen, dass mindestens ein Kind ein Mädchen ist, treffen wir eins davon auf der Straße und sehen, dass es ein Mädchen ist. Was ist dann die Wahrscheinlichkeit, dass beide Kinder Mädchen sind? Hier wird nun wichtig, dass die Kinder unterscheidbar sind – wir haben bereits eines getroffen; nehmen wir der Einfachheit an, dass es das erste ist. Das Ergebnis $\{JM\}$ kann also ebenfalls ausgeschlossen werden; damit ist die bedingte Wahrscheinlichkeit

$$\mathbb{P}\left[\{MM\} \mid \{MM, MJ\}\right] = \frac{1}{2}.$$

Dies entspricht genau der Wahrscheinlichkeit für das *zweite* Kind, ein Mädchen zu sein. Die Annahme, dass wir das erste Kind treffen, kann einschränkend wirken. Stattdessen können wir das Modell erweitern, indem wir im Ergebnisraum festhalten, welches Kind wir getroffen haben; das Ereignis „wir treffen das erste von beiden Mädchen“ ist dann $\{M^*M\}$ (der Stern zeigt an, welches Kind wir getroffen haben) usw. Wir erhalten dann den erweiterten Ergebnisraum

$$\Omega = \{M^*M, MM^*, M^*J, MJ^*, J^*M, JM^*, J^*J, JJ^*\}.$$

Nehmen wir wieder an, dass alle Ergebnisse gleich wahrscheinlich sind (d. h. wir jedes Kind mit der gleichen Wahrscheinlichkeit treffen). Dann ist die bedingte Wahrscheinlichkeit wieder

$$\mathbb{P}\left[\{M^*M, MM^*\} \mid \{M^*M, MM^*, M^*J, JM^*\}\right] = \frac{2}{4} = \frac{1}{2}.$$

DER SATZ VON DER TOTALEN WAHRSCHEINLICHKEIT

Bedingte Wahrscheinlichkeiten können sehr nützlich für die Berechnung von konkreten Wahrscheinlichkeiten sein.

Satz 4.5 (totale Wahrscheinlichkeit). *Sei $\{H_i\}_{i\in I} \subset \mathcal{A}$ eine höchstens abzählbare disjunkte Zerlegung (d. h. $\bigcup_{i\in I} H_i = \Omega$ und $H_i \cap H_j = \emptyset$ für alle $i \neq j \in I$, und I ist höchstens abzählbar) mit $\mathbb{P}[H_i] > 0$ für alle $i \in I$. Dann gilt*

$$\mathbb{P}[A] = \sum_{i\in I} \mathbb{P}[A \mid H_i] \cdot \mathbb{P}[H_i] \qquad \text{für alle } A \in \mathcal{A}.$$

Beweis. Sei $A \in \mathcal{A}$ beliebig. Da die H_i, $i \in I$, disjunkt sind, sind auch $A \cap H_i$, $i \in I$, disjunkt. Weiter folgt aus der Zerlegungseigenschaft

$$A = A \cap \Omega = A \cap \Big(\bigcup_{i\in I} H_i\Big) = \bigcup_{i\in I} (A \cap H_i).$$

Also ist $\{A \cap H_i\}_{i\in I}$ eine höchstens abzählbare disjunkte Zerlegung von A. Aus der σ-Additivität von $\mathbb{P}$ und $\mathbb{P}[H_i] > 0$ folgt dann mit der Definition der bedingten Wahrscheinlichkeit

$$\sum_{i\in I} \mathbb{P}[H_i]\, \mathbb{P}[A \mid H_i] = \sum_{i\in I} \mathbb{P}[A \cap H_i] = \mathbb{P}[A]. \qquad \square$$

Die H_i werden – besonders im Kontext der Statistik – auch als *Hypothesen* bezeichnet.

Um den Satz von der totalen Wahrscheinlichkeit anwenden zu können, braucht man eine vollständige(!) Zerlegung, für die man die bedingten Wahrscheinlichkeiten gut angeben kann; sind diese nicht durch die Modellierung vorgegeben, ist das Auffinden oft nicht trivial.

Beispiel 4.6. Wir betrachten eine Urne mit zwei weißen und drei schwarzen Kugeln. Wie groß ist die Wahrscheinlichkeit, bei zweimaligem Ziehen *ohne* Zurückliegen zwei Kugeln der selben Farbe zu erhalten, wenn jede Kugel mit der gleichen Wahrscheinlichkeit gezogen wird? Wir wählen als Hypothesen

- H_1: „die erste Kugel ist weiß“ mit Wahrscheinlichkeit (unter der angenommenen Gleichverteilung) $\mathbb{P}[H_1] = \frac{2}{5}$;
- H_2: „die erste Kugel ist schwarz“ mit Wahrscheinlichkeit $\mathbb{P}[H_2] = \frac{3}{5}$.

Sei A das Ereignis „beide Kugeln haben die selbe Farbe“. Da wir bedingt auf den ersten Zug den Zustand der Urne genau kennen, können wir analog die bedingten Wahrscheinlichkeiten berechnen:

- Unter Hypothese H_1 enthält die Urne noch eine weiße und drei schwarze Kugeln; um zweimal die selbe Farbe zu erhalten, müssen wir also die verbleibende *weiße* Kugel ziehen, was mit Wahrscheinlichkeit $\mathbb{P}[A \mid H_1] = \frac{1}{4}$ passiert;
- unter der Hypothese H_2 enthält die Urne noch zwei weiße und zwei schwarze Kugeln; diesmal müssen wir eine der verbleibenden schwarzen Kugel ziehen, was mit Wahrscheinlichkeit $\mathbb{P}[A \mid H_2] = \frac{2}{4} = \frac{1}{2}$ passiert.

Nach Satz 4.5 ist daher

$$\mathbb{P}[A] = \mathbb{P}[A \mid H_1] \cdot \mathbb{P}[H_1] + \mathbb{P}[A \mid H_2] \cdot \mathbb{P}[H_2] = \frac{1}{4} \cdot \frac{2}{5} + \frac{1}{2} \cdot \frac{3}{5} = \frac{2}{5}.$$

Zur Probe berechnen wir die gleiche Wahrscheinlichkeit mit Hilfe der hypergeometrischen Verteilung: Betrachten wir die weißen Kugeln als *markiert*, so entspricht A dem Ereignis „entweder zwei markierte oder keine markierten Kugeln werden gezogen“, d. h. der Wert der Zufallsvariable X, die die Anzahl der gezogenen Kugeln ohne Zurücklegen beschreibt, ist 2 oder 0. Also ist

$$\begin{aligned}\mathbb{P}[A] = \mathbb{P}[X = 2] + \mathbb{P}[X = 0] = h_{5,2,2}(2) + h_{5,2,2}(0) &= \frac{\binom{2}{2}\binom{5-2}{2-2}}{\binom{5}{2}} + \frac{\binom{2}{0}\binom{5-2}{2-0}}{\binom{5}{2}} \\ &= \frac{1}{10} + \frac{3}{10} = \frac{2}{5}.\end{aligned}$$

Ein interessanter (und unintuitiver) Effekt ist, dass bei einem Wechsel der grundlegenden Wahrscheinlichkeitsverteilungen die totale Wahrscheinlichkeit abnehmen kann, selbst wenn alle bedingten Wahrscheinlichkeiten zunehmen.

Beispiel 4.7 (Simpson-Paradoxon). Dieses Beispiel basiert auf einer tatsächlichen medizinischen Studie [Charig u. a. 1986], in der zwei Behandlungen von Nierensteinen

verglichen werden:

- Behandlung A hat bei einer Probandengruppe eine (empirische) Erfolgswahrscheinlichkeit von $\mathbb{P}_A[\{\text{Behandlung erfolgreich}\}] \approx 0.78$.
- Behandlung B hat bei einer (anderen!) Probandengruppe eine Erfolgswahrscheinlichkeit von $\mathbb{P}_B[\{\text{Behandlung erfolgreich}\}] \approx 0.83$.

Also scheint Behandlung B den besseren Erfolg zu haben.
Bei genauerer Betrachtung ist jedoch das Gegenteil der Fall: Unterteilt man die Probanden weiter in eine Gruppe, bei der der behandelte Nierenstein groß war und eine, bei der der Nierenstein klein war, so entnimmt man der Studie folgende Wahrscheinlichkeiten für $E = \{\text{Behandlung erfolgreich}\}$, $G = \{\text{großer Stein}\}$, $K = \{\text{kleiner Stein}\}$:

- $\mathbb{P}_A[E \mid K] \approx 0.93, \qquad \mathbb{P}_A[E \mid G] \approx 0.73;$
- $\mathbb{P}_B[E \mid K] \approx 0.87, \qquad \mathbb{P}_B[E \mid G] \approx 0.69.$

Behandlung A ist also in beiden Fällen besser!
Die Erklärung liegt darin, dass die beiden Hypothesen in den verschiedenen Probandengruppen sehr unterschiedliche Wahrscheinlichkeiten haben: $\mathbb{P}_A[G] \approx 0.75$, aber $\mathbb{P}_B[G] \approx 0.22$ – und die Erfolgswahrscheinlichkeit hängt stärker von der Größe des Steins ab als von der angewendeten Behandlung: die schlechtere Behandlung hat für kleine Steine eine höhere Erfolgswahrscheinlichkeit als die bessere Behandlung für große Steine. (Dies ist kein Zufall; die Ärzte, die die Studie ausgeführt haben, haben bei schwererer Krankheit natürlich eher die vielversprechendere Behandlung angewandt.)
Dieses Beispiel unterstreicht, wie wichtig eine sorgfältige und unvoreingenommene Modellierung ist, um sinnvolle statistische Resultate zu erhalten.

DER SATZ VON BAYES

Direkt aus der Definition der bedingten Wahrscheinlichkeit erhält man die folgende nützliche Formel.

Satz 4.8 (Bayes). *Seien $A, B \in \mathcal{A}$ mit $\mathbb{P}[A] > 0$. Dann gilt*

$$\mathbb{P}[B \mid A] = \frac{\mathbb{P}[A \mid B] \cdot \mathbb{P}[B]}{\mathbb{P}[A]}.$$

Beweis. Nach Definition der bedingten Wahrscheinlichkeiten $\mathbb{P}[B \mid A]$ und $\mathbb{P}[A \mid B]$ ist

$$\mathbb{P}[B \mid A] = \frac{\mathbb{P}[B \cap A]}{\mathbb{P}[A]} = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]} = \frac{\mathbb{P}[A \mid B] \cdot \mathbb{P}[B]}{\mathbb{P}[A]},$$

wobei die zweite Gleichung für $\mathbb{P}[B] > 0$ aus der Definition von $\mathbb{P}[A \mid B]$ folgt und für $\mathbb{P}[B] = 0$ trivialerweise (als $0 = 0$) gilt. □

Der Nenner wird dabei oft über den Satz 4.5 der totalen Wahrscheinlichkeit berechnet.

Folgerung 4.9. *Sei $\{H_i\}_{i\in I} \subset \mathcal{A}$ eine höchstens abzählbare disjunkte Zerlegung mit $\mathbb{P}[H_i] > 0$ für alle $i \in I$. Sei $A \in \mathcal{A}$ beliebig. Dann gilt für alle $i \in I$*

$$\mathbb{P}[H_i \mid A] = \frac{\mathbb{P}[A \mid H_i] \cdot \mathbb{P}[H_i]}{\sum_i \mathbb{P}[A \mid H_i] \cdot \mathbb{P}[H_i]}. \tag{4.1}$$

Diese Formel ist unscheinbar, bildet aber die Grundlage für einen großen Teil der Statistik: dem Entscheiden zwischen verschiedener Hypothesen aufgrund von beobachteter bedingter Wahrscheinlichkeiten; man spricht auch von *Inferenz*. Das folgende Beispiel – dessen praktische Relevanz mittlerweile leidlich bekannt sein dürfte – soll die Rolle der verschiedenen vorkommenden Wahrscheinlichkeiten illustrieren.

Beispiel 4.10. Wir betrachten einen diagnostischen Test für eine Krankheit, der die folgenden Eigenschaften hat:

- kranke Personen werden mit 95% Wahrscheinlichkeit (korrekt) als infiziert erkannt (*Sensitivität* von 95%);
- gesunde Personen werden mit 10% Wahrscheinlichkeit (fälschlicherweise) als infiziert erkannt (*Spezifizität* 90%).

Angenommen, ein Test ist positiv. Was ist die Wahrscheinlichkeit, tatsächlich erkrankt zu sein? Kennen wir die generelle Wahrscheinlichkeit, dass eine beliebige Person in der Bevölkerung infiziert ist (*Prävalenz*) – und nur dann! – so können wir diese mit Hilfe der Bayes-Formel (4.1) berechnen.
Sei H_1 die Hypothese „infiziert“ und H_2 entsprechend „nicht infiziert“. Nehmen wir eine Prävalenz von 2% an, so entspricht dies

$$\mathbb{P}[H_1] = 0.02, \qquad \mathbb{P}[H_2] = 0.98.$$

Die bedingten Wahrscheinlichkeiten sind gegeben durch die Testgüte: Ist T das Ereignis „Test positiv“, so gilt

$$\mathbb{P}[T \mid H_1] = 0.95, \qquad \mathbb{P}[T \mid H_2] = 0.1.$$

Nach der Bayes-Formel ist dann die Wahrscheinlichkeit, infiziert zu sein gegeben dass der Test positiv ist („positive Korrektheit“),

$$\begin{aligned}\mathbb{P}[H_1 \mid T] &= \frac{\mathbb{P}[T \mid H_1] \cdot \mathbb{P}[H_1]}{\mathbb{P}[T \mid H_1] \cdot \mathbb{P}[H_1] + \mathbb{P}[T \mid H_2] \cdot \mathbb{P}[H_2]} = \frac{0.95 \cdot 0.02}{0.95 \cdot 0.02 + 0.1 \cdot 0.98} \\ &\approx 0.16,\end{aligned}$$

also gerade einmal 16%! Dagegen ist wegen $\mathbb{P}[T^c \mid H_i] = 1 - \mathbb{P}[T \mid H_i]$

$$\begin{aligned}\mathbb{P}[H_2 \mid T^c] &= \frac{\mathbb{P}[T^c \mid H_2] \cdot \mathbb{P}[H_2]}{\mathbb{P}[T^c \mid H_1] \cdot \mathbb{P}[H_1] + \mathbb{P}[T^c \mid H_2] \cdot \mathbb{P}[H_2]} = \frac{0.9 \cdot 0.09}{0.05 \cdot 0.02 + 0.9 \cdot 0.98} \\ &\approx 0.99,\end{aligned}$$

d. h. die Wahrscheinlichkeit, *nicht* infiziert zu sein, wenn der Test negativ ist („negative Korrektheit"), ist 99%. Dieser Test ist offensichtlich nur nützlich, um eine Infektion *auszuschließen*, nicht um sie nachzuweisen!
Der Grund, warum die positive Korrektheit trotz hoher Sensitivität so klein ist, liegt in der niedrigen Prävalenz der Krankheit – der Test ist sensitiv, wir versuchen damit aber „aus dem Rauschen zu schätzen". Bei einer Prävalenz von 10% läge die positive Korrektheit zum Beispiel schon über 50%.
Beachten Sie, dass die gewählte Hypothesenwahrscheinlichkeit einer Gleichverteilung auf der Gesamtbevölkerung entspricht („die Wahrscheinlichkeit, dass ein zufällig gewählter Mensch krank ist"); dies mag im Rahmen einer Routinevorsorgeuntersuchung angemessen sein. Die Situation ändert sich aber grundlegend, wenn der Test wegen bereits vorliegender Symptome gemacht wird! Dann ist ist nicht mehr die Prävalenz in der Gesamtbevölkerung entscheidend, sondern die *relative* Prävalenz der Krankheit unter allen Personen, die diese Symptome haben – und diese kann (muss aber nicht!) signifikant höher sein, wodurch sich auch die positive Korrektheit deutlich erhöht. (Die Frage ist dann nicht mehr „Bin ich krank?" – das ist aufgrund der Symptome schon klar – sondern nur noch „Habe ich genau diese Krankheit?") Dies illustriert noch einmal die fundamentale Bedeutung einer sorgfältigen Modellierung der Hypothesen und ihrer Wahrscheinlichkeiten.

Der Satz von Bayes ist also kein Wunderwerkzeug, um verborgene Wahrheiten offenzulegen, sondern erlaubt lediglich, *bekanntes* Wissen durch neue Zusatzinformationen zu verbessern – und je präziser die Zusatzinformationen, desto aussagekräftiger das Ergebnis.

Beispiel 4.11 (Monty Hall-Problem, Ziegenproblem). Die folgende Knobelaufgabe (die auf der realen Spielshow *„Let's Make a Deal"* basiert, dessen Moderator der Aufgabe ihren Namen im englischen Sprachgebrauch spendiert hat und die für das deutsche Fernsehen als „Geh aufs Ganze!" adaptiert wurde) hat in den 1990er Jahren für handfestem Streit gesorgt. Die Situation ist folgende: Angenommen, man hat in einer Spielshow drei Türen zur Auswahl: hinter einer Tür ist ein Preis (üblicherweise ein Auto), hinter den anderen beiden eine Niete (üblicherweise Ziegen; daher der deutsche Name). Der Spieler wählt nun eine Tür, sagen wir Tür 1, ohne sie zu öffnen. Daraufhin öffnet der Moderator eine andere Tür (sagen wir Tür 3), hinter der sich eine Niete befindet. Nun bietet er dem Spieler an, seine Wahl zu ändern (naheliegenderweise auf Tür 2). Soll er dies machen? Schnell kamen zwei verschiedene Antworten auf:

- Die erste Antwort basiert auf der Argumentation, dass nach dem Öffnen der Tür 3 noch zwei Türen zur Auswahl stehen; da hinter jeder der beiden Türen der Preis sein könnte, haben beide die Wahrscheinlichkeit $\frac{1}{2}$, und damit lohnt sich ein Wechsel nicht.
- Die zweite Antwort bezieht sich darauf, dass der Moderator sicher nicht die Tür mit dem Preis öffnen wird, und ebenfalls nicht die vom Spieler gewählte Tür (er

öffnet ja „eine andere“). Es verbleiben also nur die beiden Möglichkeiten, dass der Preis hinter Tür 1 oder hinter Tür 2 liegt. Da wir für Tür 1 die Wahrscheinlichkeit $\frac{1}{3}$ angesetzt haben, muss die Wahrscheinlichkeit für Tür 2 also $\frac{2}{3}$ sein. Wir sollten also wechseln.

Welche Antwort ist nun die richtige? Tatsächlich widersprechen sich die Antworten nicht, sondern basieren einfach auf unterschiedlichen Modellannahmen (für das Verhalten des Moderators). Wir sind uns über zwei Dinge unsicher, wenn wir selber spielen:

(i) hinter welcher Tür sich der Preis befindet;

(ii) welche Tür der Moderator öffnet, gegeben welche Tür wir wählen und hinter welcher Tür der Preis ist (was der Moderator ja weiß).

Beides beschreiben wir durch Zufallsvariablen mit Werten in $\{1, 2, 3\}$. Da wir keine Informationen haben, wo sich der Preis befindet, nehmen wir für die Zufallsvariable P („Preis hinter Tür k“) eine Gleichverteilung an:

$$\mathbb{P}[P = 1] = \mathbb{P}[P = 2] = \mathbb{P}[P = 3] = \frac{1}{3}.$$

Wir wählen nun enstprechend dieser Verteilung eine zufällige Tür, die wir (notfalls durch Umnummerieren, was die Wahrscheinlichkeiten ja nicht ändert) als Tür 1 annehmen können. Daraufhin öffnet der Moderator eine Tür, die wir analog (nach eventuellem Umnummerieren) als Tür 3 annehmen können. Wie ändert sich dadurch unsere Einschätzung, hinter welcher Tür der Preis sein könnte – in anderen Worten, was ist $\mathbb{P}[P = k \mid M = 3]$ für $k = 1, 2, 3$?
Dafür verwenden wir die Bayes-Formel: Nach Folgerung 4.9 ist für $k = 1, 2, 3$

$$\begin{aligned}\mathbb{P}[P = k \mid M = 3] &= \frac{\mathbb{P}[M = 3 \mid P = k] \cdot \mathbb{P}[P = k]}{\sum_{i=1}^{3} \mathbb{P}[M = 3 \mid P = i] \cdot \mathbb{P}[P = i]} \\ &= \frac{\mathbb{P}[M = 3 \mid P = k]}{\mathbb{P}[M = 3 \mid P = 1] + \mathbb{P}[M = 3 \mid P = 2] + \mathbb{P}[M = 3 \mid P = 3]}\end{aligned}$$

wegen $\mathbb{P}[P = k] = \frac{1}{3}$ nach Grundannahme. Es bleiben also die bedingten Wahrscheinlichkeiten zu bestimmen, die die Strategie des Moderators beschreiben und in denen sich auch die wesentliche Modellierung (und damit der Unterschied in den Antworten) versteckt. Offensichtlich öffnet der Moderator nicht die Tür mit dem Preis (sonst ist das Spiel langweilig); also ist $\mathbb{P}[M = 3 \mid P = 3] = 0$ und damit auch $\mathbb{P}[P = 3 \mid M = 3] = 0$. Interessanter sind die anderen beiden Fälle:

- Ist der Preis hinter Tür 1, so hat der Moderator zwei Türen mit Nieten zur Auswahl; ohne weiteres Wissen nehmen wir für die Wahl der geöffneten Tür ebenfalls eine Gleichverteilung an, d. h. insbesondere

$$\mathbb{P}[M = 3 \mid P = 1] = \frac{1}{2}.$$

- Ist der Preis hinter Tür 2, kommen die zwei Varianten ins Spiel, die einer unterschiedlichen Auslegung der Aufgabenstellung entsprechen:
  (i) Die Aufgabenstellung beschreibt (nur) eine Spiel*situation*: Der Moderator *kann* eine der beiden Türen Türen mit Nieten wählen – darunter auch die von uns gewählte Tür – hat dies aber *in diesem Fall* nicht gemacht; ohne weiteres Wissen nehmen wir an, dass beide gleich wahrscheinlich sind mit insbesondere
  $$\mathbb{P}[M = 3 \mid P = 2] = \frac{1}{2}.$$
  (ii) Die Aufgabenstellung beschreibt eine Spiel*regel*: Der Moderator *darf* nur die von uns nicht gewählte Tür 3 öffnen (da Tür 2 wegen dem Preis dahinter ausgeschlossen ist), so dass gilt
  $$\mathbb{P}[M = 3 \mid P = 2] = 1.$$

(Beachten Sie: Bedingte Wahrscheinlichkeiten summieren sich über eine disjunkte Zerlegung *nicht* zu 1, siehe Satz 4.5.) In Variante (i) gilt also

$$\mathbb{P}[P = 1 \mid M = 3] = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{2} + 0} = \frac{1}{2} = \mathbb{P}[P = 2 \mid M = 3],$$

es lohnt sich also nicht zu wechseln. (Dies ist die genau erste Antwort). Dagegen ist in Variante (ii)

$$\mathbb{P}[P = 1 \mid M = 3] = \frac{\frac{1}{2}}{\frac{1}{2} + 1 + 0} = \frac{1}{3},$$
$$\mathbb{P}[P = 2 \mid M = 3] = \frac{1}{\frac{1}{2} + 1 + 0} = \frac{2}{3},$$

so dass sich das Wechseln sehr wohl lohnt. (Dies ist genau die zweite Antwort.)
Der Unterschied mag unintuitiv sein, lässt sich aber wie folgt plausibel machen: In Variante (ii) wissen wir schon im Voraus (bzw. glauben zu wissen), dass der Moderator niemals Tür 1 öffnet; wir erhalten durch das Öffnen von Tür 3 keine neue Information darüber, was hinter Tür 1 liegt. Also bleibt die Preiswahrscheinlichkeit bei $\frac{1}{3}$, und die Restwahrscheinlichkeit von $\frac{2}{3}$ muss sich auf Tür 2 konzentrieren (da ein Preis hinter Tür 3 nun ausgeschlossen ist.) In Variante (i) *hätte* der Moderator prinzipiell Tür 1 öffnen können, hat dies aber nicht gemacht – was auch daran liegen könnte, dass dahinter der Preis ist. Also ist die (subjektive) Preiswahrscheinlichkeit für Tür 1 etwas gestiegen.

Dieses Problem lässt sich beliebig abwandeln, um es noch interessanter zu gestalten. Was ist zum Beispiel, wenn wir keine Gleichverteilung für P annehmen (da wir als langjähriger Fan beobachtet haben, dass das Auto etwas öfter hinter Tür 3 liegt, und

vielleicht auch Tür 2 öfter vorkommt als Tür 1 – vermutlich, weil es dort einfacher zu parken ist)? Was ist, wenn der Moderator eine Präferenz für eine der Nietentüren hat (z. B. die mit kleinerer Zahl, weil er da weniger laufen muss um sie zu öffnen)? Was ist, wenn es auch sein kann, dass hinter *keiner* Tür ein Preis ist? Was ist, wenn wir uns bezüglich der Strategie des Moderators (Variante (i) oder (ii)) unsicher sind und diese ebenfalls durch eine Zufallsvariable beschreiben möchten?

4.2 BEDINGTER ERWARTUNGSWERT

Wir haben bislang bedingte Wahrscheinlichkeiten von einzelnen Ereignissen betrachtet. Solange diese wohldefiniert ist, können wir damit eine neue *bedingte Wahrscheinlichkeitsverteilung* definieren.

Satz 4.12. *Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Ist $B \in \mathcal{A}$ mit $\mathbb{P}[B] > 0$, so definiert*

$$\mathbb{P}_B : \mathcal{A} \to [0,1], \qquad \mathbb{P}_B[A] := \mathbb{P}[A \mid B],$$

eine Wahrscheinlichkeitsverteilung auf $\mathcal{A}$.

Beweis. Wegen $\mathbb{P}[B] > 0$ ist auch $\mathbb{P}[A \mid B] \geq 0$ für alle $A \in \mathcal{A}$. Weiter ist natürlich

$$\mathbb{P}_B[\Omega] = \frac{\mathbb{P}[\Omega \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B]}{\mathbb{P}[B]} = 1.$$

Sind schließlich $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{A}$ paarweise disjunkt, dann sind es auch $\{A_n \cap B\}_{n\in\mathbb{N}} \subset \mathcal{A}$ (vergleiche Beweis von Satz 4.5), und aus der σ-Additivität von $\mathbb{P}$ folgt

$$\begin{aligned}\mathbb{P}_B\left[\bigcup_{n\in\mathbb{N}} A_n\right] &= \frac{\mathbb{P}\left[\left(\bigcup_{n\in\mathbb{N}} A_n\right) \cap B\right]}{\mathbb{P}[B]} = \frac{\mathbb{P}\left[\bigcup_{n\in\mathbb{N}} (A_n \cap B)\right]}{\mathbb{P}[B]} = \frac{\sum_{n\in\mathbb{N}} \mathbb{P}[A_n \cap B]}{\mathbb{P}[B]} \\ &= \sum_{n\in\mathbb{N}} \mathbb{P}_B[A_n].\end{aligned}$$

□

Dies erlaubt es, bedingte Wahrscheinlichkeiten bezüglich bedingter Wahrscheinlichkeiten zu bilden. Gilt zum Beispiel $\mathbb{P}_B[C] > 0$ für ein $C \in \mathcal{A}$ (wann gilt dies?), so ist die bedingte Wahrscheinlichkeit von $A \in \mathcal{A}$ gegeben C, gegeben B

$$\mathbb{P}_B[A \mid C] = \frac{\mathbb{P}_B[A \cap C]}{\mathbb{P}_B[C]} = \frac{\frac{\mathbb{P}[(A\cap C)\cap B]}{\mathbb{P}[B]}}{\frac{\mathbb{P}[C\cap B]}{\mathbb{P}[B]}} = \frac{\mathbb{P}[A \cap B \cap C]}{\mathbb{P}[B \cap C]} = \mathbb{P}[A \mid B \cap C].$$

Sei nun $X : \Omega \to S \subset \mathbb{R}$ eine diskrete Zufallsvariable. Dann können wir – wieder für $B \in \mathcal{A}$ mit $\mathcal{P}[B] > 0$ – die bedingte Wahrscheinlichkeit von Ereignissen der Form $A = \{X = s\}$,

$s \in S$, gegeben B betrachten. Insbesondere können wir den *bedingten Erwartungswert von X gegeben B* berechnen als

$$\mathbb{E}[X \mid B] = \sum_{s \in S} s\, \mathbb{P}_B[X = s] =: \sum_{s \in S} s\, \mathbb{P}[X = s \mid B],$$

falls diese Summe wohldefiniert ist. Intuitiv ist der bedingte Erwartungswert die beste Vorhersage für X, die wir machen können, falls wir bereits wissen dass das Ereignis B eintritt.

Da die bedingte Wahrscheinlichkeit $\mathbb{P}_B$ nach Satz 4.12 eine Wahrscheinlichkeitsverteilung ist, hat der bedingte Erwartungswert alle Eigenschaften eines Erwartungswerts, insbesondere Linearität und Monotonie. Aus Folgerung 3.10 folgt weiter, dass für abzählbare Ω gilt

$$\mathbb{E}[X \mid B] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}_B[\{\omega\}] = \frac{1}{\mathbb{P}[B]} \sum_{\omega \in B} X(\omega) \cdot \mathbb{P}[\{\omega\}] \tag{4.2}$$

wegen $\mathbb{P}[\{\omega\} \cap B] = 0$ für $\omega \notin B$.

Beispiel 4.13. Wir betrachten den zweifachen fairen Münzwurf, d h. $\Omega = \{1,2,3,4,5,6\}^2$ mit der Gleichverteilung. Was ist der bedingte Erwartungswert des ersten Wurfs, wenn die Summe der Augenzahlen höchstens 5 ist? Wir beschreiben dafür den Ausgang der beiden Würfe jeweils durch eine Zufallsvariable $X_{1,2} : \Omega \to \{1,2,3,4,5,6\}$. Das gegebene Ereignis ist dann

$$B = \{X_1 + X_2 \leq 5\} = \{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(3,1),(3,2),(4,1)\}.$$

Aufgrund der Gleichverteilung ist $\mathbb{P}[B] = \frac{|B|}{|\Omega|} = \frac{10}{36} > 0$. Also gilt

$$\mathbb{E}[X \mid B] = \frac{\frac{1}{36}(1+1+1+1+2+2+2+3+3+4)}{\frac{10}{36}} = 2.$$

Analog zu Satz 4.5 erhalten wir daraus auch eine Formel zur Berechnung von Erwartungswerten über bedingte Erwartungswerte.

Satz 4.14. *Sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum und sei $\{H_i\}_{i \in I} \subset \mathcal{P}(\Omega)$ eine disjunkte Zerlegung mit $\mathbb{P}[H_i] > 0$ für alle $i \in I$. Dann gilt für jede Zufallsvariable $X \in \ell^1$*

$$\mathbb{E}[X] = \sum_{i \in I} \mathbb{E}[X \mid H_i] \cdot \mathbb{P}[H_i].$$

Beweis. Da Ω (und damit auch die Zerlegung) höchstens abzählbar ist und alle Reihen nach Annahme absolut konvergieren, folgt aus Folgerung 3.10 und (4.2) für $B = H_i$

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}[\{\omega\}] = \sum_{i \in I} \left(\sum_{\omega \in H_i} X(\omega) \cdot \mathbb{P}[\{\omega\}] \right) = \sum_{i \in I} \mathbb{E}[X \mid H_i] \cdot \mathbb{P}[H_i]. \qquad \square$$

Ist $Z : \Omega \to T \subset \mathbb{R}$ eine diskrete Zufallsvariable, so können wir insbesondere auch für die Bedingung Ereignisse der Form $B = \{Z = t\}$ betrachten. In einem weiteren Schritt können wir nun die Abhängigkeit von t betrachten; da der Wert von Z zufällig ist, gilt dies auch für den durch sie bedingten Erwartungswert! Die so definierte Zufallsvariable nennt man *bedingte Erwartung*

$$\mathbb{E}[X \mid Z] : \Omega \to \mathbb{R}, \qquad \mathbb{E}[X \mid Z](\omega) := \begin{cases} \mathbb{E}[X \mid Z = Z(\omega)] & \text{falls } \mathbb{P}[Z = Z(\omega)] > 0, \\ 0 & \text{sonst.} \end{cases}$$

Von dieser Zufallsvariable kann man nun wieder die Wahrscheinlichkeitsverteilung untersuchen. Speziell enthält man aus Satz 4.14 mit $H_t := \{Z = t\}$ für $t \in T$ (nach Annahme abzählbar) mit $\mathbb{P}[Z = t] > 0$ zusammen mit der Definition des Erwartungswerts die folgende eingängige Formel:

Folgerung 4.15. *Sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum und $Z : \Omega \to T \subset \mathbb{R}$ eine diskrete Zufallsvariable. Dann gilt für jede Zufallsvariable $X \in \ell^1$*

$$\mathbb{E}[X] = \sum_{t \in T} \mathbb{E}[X \mid Z = t] \cdot \mathbb{P}[Z = t] = \mathbb{E}\left[\mathbb{E}[X \mid Z]\right].$$

Schließlich kann man ganz analog die bedingte Varianz einer Zufallsvariablen (gegeben ein Ereignis oder eine andere Zufallsvariable) definieren, was wir uns hier aber sparen.

4.3 UNABHÄNGIGKEIT

Wir können uns nun fragen, was es bedeutet, wenn uns das Wissen, dass ein Ereignis B eintritt, *keine* neuen Informationen über ein Ereignis A liefert – d. h. dass gilt

$$\mathbb{P}[A \mid B] = \mathbb{P}[A].$$

Umgekehrt kann uns Eintreten von A keine neuen Informationen über B liefern, d. h.

$$\mathbb{P}[B \mid A] = \mathbb{P}[B].$$

Nach Definition der bedingten Wahrscheinlichkeit sind beide Fälle äquivalent zu

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot \mathbb{P}[B],$$

was auch für $\mathbb{P}[A] = 0$ oder $\mathbb{P}[B] = 0$ sinnvoll ist. Ereignisse, für die (4.3) gilt, nennt man *(stochastisch) unabhängig*; wollen wir die Wahrscheinlichkeitsverteilung betonen, sagen wir *unabhängig bezüglich* $\mathbb{P}$. Beachten Sie, dass dies nur eine Aussage über gemeinsame Informationen (oder proportionale Überschneidung von Wahrscheinlichkeiten) bzw. deren Fehlen ist, nicht über kausale Zusammenhänge!

Beispiel 4.16. Wir betrachten wieder den fairen doppelten Würfelwurf, d. h. die Gleichverteilung auf $\mathcal{P}(\{1,2,3,4,5,6\}^2)$ sowie die Ereignisse

$$A := \{\text{Augensumme ist } 7\} = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\},$$
$$B := \{\text{erster Wurf zeigt } 6\} = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\}.$$

Dann ist offensichtlich $A \cap B = \{(1,6)\}$ und damit

$$\mathbb{P}[A \cap B] = \frac{1}{36} = \frac{6}{36} \cdot \frac{6}{36} = \mathbb{P}[A] \cdot \mathbb{P}[B].$$

Also sind die beiden Ereignisse stochastisch unabhängig, obwohl die Augensumme kausual vom ersten Wurf abhängt.
(Dies ist nur zufällig der Fall, da wir 7 als Augensumme gewählt haben.)

Die stochastische Unabhängigkeit ist also nichts weiter als eine Modellannahme, die uns bequeme Rechnungen ermöglicht, da wir nur die Wahrscheinlichkeiten der einzelnen Ereignisse kennen müssen. (In der Tat hat die falsche Annahme der Unabhängigkeit aus Bequemlichkeit schon viel Unheil angerichtet, darunter – grob vereinfacht – die globale Finanzkrise von 2007.)

Wir können auch Unabhängigkeit von mehr als zwei Ereignissen definieren.

Definition 4.17 (unabhängige Ereignisse). Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Eine Familie $\{A_i\}_{i\in I} \subset \mathcal{A}$ für eine nichtleere Menge I heißt *unabhängig* bezüglich $\mathbb{P}$, falls für jede nichtleere endliche Teilmenge $J \subset I$ gilt

$$\mathbb{P}\left[\bigcap_{i\in J} A_i\right] = \prod_{i\in J} \mathbb{P}[A_i].$$

(Im Gegensatz zur – immer geltenden – σ-Additivität fordern wir also eine *endliche* Multiplikativität der Wahrscheinlichkeiten.) Beachten Sie, dass eine rein paarweise oder totale Unabhängigkeit nicht ausreicht!

Beispiel 4.18. Wir betrachten den doppelten fairen Münzwurf, d. h. die Gleichverteilung auf $\mathcal{P}(\{0,1\}^2)$ sowie die Ereignisse

$$A_1 := \{\text{erster Wurf ergibt Zahl}\} = \{(1,0),(1,1)\},$$
$$A_2 := \{\text{zweiter Wurf ergibt Zahl}\} = \{(0,1),(1,1)\},$$
$$A_3 := \{\text{beide Würfe haben das gleiche Ergebnis}\} = \{(0,0),(1,1)\}.$$

Dann ist $A_i \cap A_j = \{(1,1)\}$ für alle $i, j = 1,2,3$ mit $i \neq j$ und daher

$$\mathbb{P}[A_i \cap A_j] = \frac{1}{4} = \frac{2}{4} \cdot \frac{2}{4} = \mathbb{P}[A_i] \cdot \mathbb{P}[A_j],$$

d. h. die Ereignisse sind paarweise unabhängig. Dagegen sind alle drei Ereignisse zusammen *nicht* unabhängig, denn es gilt

$$\mathbb{P}[A_1 \cap A_2 \cap A_3] = \frac{1}{4} \neq \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{2}{4} = \mathbb{P}[A_1] \cdot \mathbb{P}[A_2] \cdot \mathbb{P}[A_3].$$

Umgekehrt gilt für den dreifachen Münzwurf und

$$\begin{aligned} B_1 &:= \{(1,0,0), (1,1,0), (1,1,1), (1,0,1)\}, \\ B_2 &:= \{(1,0,0), (1,1,0), (1,1,1), (1,0,1)\}, \\ B_3 &:= \{(1,0,0), (0,0,0), (0,0,1), (0,1,0)\} \end{aligned}$$

wegen $B_1 \cap B_2 \cap B_3 = \{(1,0,0)\}$

$$\mathbb{P}[B_1 \cap B_2 \cap B_3] = \frac{1}{8} = \mathbb{P}[B_1] \cdot \mathbb{P}[B_2] \cdot \mathbb{P}[B_3],$$

aber

$$\mathbb{P}[B_1 \cap B_2] = \frac{1}{2} \neq \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}[B_1] \cdot \mathbb{P}[B_2].$$

Auch diese drei Ereignisse sind also nicht unabhängig.

Sind A und B unabhängige Ereignisse, so sind es auch ihre Komplemente: Zum Beispiel folgt aus der disjunkten Zerlegung $A = (A \cap B) \cup (A \cap B^c)$, dass gilt

$$\begin{aligned} \mathbb{P}[A \cap B^c] &= \mathbb{P}[A] - \mathbb{P}[A \cap B] = \mathbb{P}[A] - \mathbb{P}[A]\,\mathbb{P}[B] = \mathbb{P}[A](1 - \mathbb{P}[B]) \\ &= \mathbb{P}[A]\,\mathbb{P}[B^c]. \end{aligned}$$

Allgemein kann man analog zum Einschluss-Ausschluss-Prinzip (Folgerung 3.14) das folgende Stabilitätsresultat beweisen.

Lemma 4.19. *Seien die Ereignisse $A_1, \dots A_n \in \mathcal{A}$ unabhängig und gelte $B_j \in \{A_j, A_j^c\}$ für $j = 1, \dots, n$. Dann sind auch $B_1, \dots, B_n$ unabhängig.*

Wir übertragen nun auf bereits bekannte Weise den Begriff der Unabhängigkeit auf diskrete Zufallsvariablen: Zwei diskrete Zufallsvariablen $X : \Omega \to S$ und $Y : \Omega \to T$ heißen *unabhängig*, wenn die Ereignisse $\{X = s\}$ und $\{Y = t\}$ unabhängig sind für alle $s \in S$, $t \in T$. Im allgemeinen Fall erhält man daraus die folgende Definition.

Definition 4.20 (unabhängige Zufallsvariablen). Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Eine Familie $\{X_i\}_{i \in I}$ von diskreten Zufallsvariablen $X_i : \Omega \to S_i$ für eine nichtleere Menge I heißt *unabhängig* bezüglich $\mathbb{P}$, falls für jede nichtleere endliche Teilmenge $J \subset I$ gilt

$$\mathbb{P}\left[\bigcap_{i \in J} \{X_i = s_i\}\right] = \prod_{i \in J} \mathbb{P}[X_i = s_i] \qquad \text{für alle } s_i \in S_i.$$

Unabhängige Zufallsvariablen sind insbesondere paarweise unkorrelliert.

Satz 4.21. *Seien $\{X_i\}_{i\in I} \subset \ell^2$ unabhängige diskrete Zufallsvariablen. Dann gilt* $\operatorname{Cov}[X_i, X_j] = 0$ *für alle $i \neq j$ und damit*

$$\begin{aligned}\mathbb{E}[X_iX_j] &= \mathbb{E}[X_i] \cdot \mathbb{E}[X_j],\\ \mathbb{V}[X_i + X_j] &= \mathbb{V}[X_i] + \mathbb{V}[X_j].\end{aligned}$$

Beweis. Nach Folgerung 3.26 und Satz 3.30 müssen wir nur die erste Gleichheit zeigen. Seien $X_i : \Omega \to S_i$ und $X_j : \Omega \to S_j$ mit $S_i, S_j \subset \mathbb{R}$ abzählbar. Dann folgt aus dem Transformationssatz 3.8 zusammen mit der Unabhängigkeit

$$\begin{aligned}\mathbb{E}[X_iX_j] &= \sum_{s_i\in S_i}\sum_{s_j\in S_j} s_is_j\,\mathbb{P}[X_i = s_i, X_j = s_j]\\ &= \sum_{s_i\in S_i}\sum_{s_j\in S_j} s_is_j\,\mathbb{P}[X_i = s_i]\,\mathbb{P}[X_j = s_j]\\ &= \sum_{s_i\in S_i} s_i\,\mathbb{P}[X_j = s_j] \cdot \sum_{s_j\in S_j} s_j\,\mathbb{P}[X_i = s_i]\\ &= \mathbb{E}[X_i] \cdot \mathbb{E}[X_j].\end{aligned}$$

□

Dagegen sind unkorrelierte Zufallsvariablen nicht notwendigerweise unabhängig!

Beispiel 4.22. Sei X gleichverteilt auf $S = \{-1, 0, 1\}$ und $Y = X^2$. Dann sind X und Y unkorrelliert wegen

$$\mathbb{E}[XY] = \frac{1}{3}(-1 \cdot 1 + 0 \cdot 0 + 1 \cdot 1) = 0 = 0 \cdot \frac{2}{3} = \mathbb{E}[X]\,\mathbb{E}[Y],$$

aber

$$\mathbb{P}[X = 0, Y = 0] = \mathbb{P}[X = 0] = \frac{1}{3} \neq \frac{1}{3} \cdot \frac{1}{3} = \mathbb{P}[X = 0]\,\mathbb{P}[Y = 0]$$

und damit sind X und Y nicht unabhängig.

Mit dieser Definition können wir die Konstruktion aus Beispiel 3.15 und Beispiel 3.31 sauber begründen.

Beispiel 4.23 (Binomial-Verteilung). Wir betrachten die Situation, dass aus einer Urne mit m Kugeln, von denen r markiert sind, n Kugeln mit Zurücklegen gezogen werden. Wir definieren wieder die Zufallsvariable $X_i : \Omega \to \{0, 1\}$, die angibt, dass im i-ten Zug eine markierte Kugel gezogen wurde. Nach Annahme, dass jede Kugel gleichwahrscheinlich gezogen wird, ist $\mathbb{P}[X_i = 1] = p := \frac{r}{m}$ für alle $i = 1, \dots n$. Weiter nehmen wir an(!), dass die Züge (stochastisch) unabhängig voneinander sind. Dann gilt für die Zufallsvariable

$$X = X_1 + \cdots + X_n : \Omega \to \{0, \dots, n\},$$

die die Anzahl der markierten Kugeln, die gezogen wurden, angibt, für alle $k = 0, \dots, n$

$$\begin{aligned}
\mathbb{P}[X = k] &= \mathbb{P}\left[\bigcup_{|I|=k} (\{X_i = 1, i \in I\} \cap \{X_i = 0, i \notin I\})\right] \\
&= \sum_{|I|=k} \mathbb{P}\left[\bigcap_{i\in I}\{X_i = 1\} \cap \bigcap_{i\notin I}\{X_i = 0\}\right] \\
&= \sum_{|I|=k} \prod_{i\in I} \mathbb{P}[X_i = 1] \prod_{i\notin I} \mathbb{P}[X_i = 0] \\
&= \sum_{|I|=k} p^k (1-p)^{n-k} \\
&= \binom{n}{k} p^k (1-p)^{n-k} = b_{n,p}(k),
\end{aligned}$$

denn die Anzahl der Summanden entspricht den $\binom{n}{k}$ Möglichkeiten, Teilmengen I von k Elementen aus der Menge $\{1, \dots, n\}$ mit n Elementen zu wählen. Also ist X in der Tat binomialverteilt.

Im Allgemeinen können wir auf ähnliche Weise für unabhängige(!) Zufallsvariablen die Verteilung ihrer Summe berechnen: Sind $X : \Omega \to S \subset \mathbb{R}$ und $Y : \Omega \to T \subset \mathbb{R}$ unabhängige diskrete Zufallsvariablen, dann ist für alle k im Wertebereich von $X + Y$

$$\begin{aligned}
\mathbb{P}[X + Y = k] &= \sum_{s\in S, t\in T, s+t=k} \mathbb{P}[X = s, Y = t] \\
&= \sum_{s\in S, k-s\in T} \mathbb{P}[X = s, Y = k - s] \\
&= \sum_{s\in S, k-s\in T} \mathbb{P}[X = s]\, \mathbb{P}[Y = k - s].
\end{aligned}$$

Hat X die Zähldichte p_X und Y die Zähldichte p_Y (jeweils mit 0 fortgesetzt auf $\mathbb{R}$), dann können wir also schreiben

$$p_{X+Y}(k) = \sum_s p_X(s \in S) p_Y(k - s)$$

(wobei nach Annahme höchstens abzählbar viele Summanden von Null verschieden sind). Die Reihe auf der rechten Seite nennt man *Faltung* von p_X und p_Y.

Beispiel 4.24. Seien X, Y unabhängig und Poisson-verteilt mit Parameter λ_X und λ_Y.

Dann gilt für alle $k \geq 0$

$$\begin{aligned}
\mathbb{P}[X + Y = k] &= \sum_{m=0}^{k} p_{\lambda_X}(m) p_{\lambda_Y}(k-m) \\
&= \sum_{m=0}^{k} \frac{\lambda_X^m}{m!} e^{-\lambda_X} \frac{\lambda_Y^{k-m}}{(k-m)!} e^{-\lambda_Y} \\
&= \left(\sum_{m=0}^{k} \frac{1}{m!(k-m)!} \lambda_X^m \lambda_Y^{k-m} \right) e^{-(\lambda_X+\lambda_Y)} \\
&= \frac{(\lambda_X + \lambda_Y)^k}{k!} e^{-(\lambda_X+\lambda_Y)} \\
&= p_{\lambda_X+\lambda_Y}(k)
\end{aligned}$$

nach der binomischen Formel $(a+b)^k = \sum_{m=0}^{k} \frac{k!}{m!(k-m)!} a^m b^{k-m}$. Also ist $X + Y$ ebenfalls Poisson-verteilt mit Parameter $\lambda_X + \lambda_Y$.

Im Allgemeinen hat jedoch die Summe selbst von unabhängigen identisch verteilten Zufallsvariablen *nicht* die gleiche Verteilung; betrachte zum Beispiel die Summe der gleichverteilten Augenzahlen zweier Würfel, die bekanntermaßen nicht gleichverteilt ist.

5 REELLE ZUFALLSVARIABLEN

Bislang haben wir uns auf diskrete Zufallsvariablen konzentriert, die höchstens abzählbar viele verschiedene Werte annehmen. In der Praxis tauchen jedoch oft Situationen auf, in der prinzipiell beliebige reelle Werte auftreten können (z. B. der Zeitpunkt des ersten Telefonanrufs an einem Tag). Dafür benötigt man *reellwertige* oder auch kurz *reelle Zufallsvariablen*. Da der Wertebereich nun überabzählbar ist, braucht man dafür einen technisch anspruchsvolleren Ansatz. Wir beschränken uns hier darauf, diesen (sauber) darzustellen, verzichten dafür aber auf Beweise. (Dies kann auch als willkommene Auffrischung der Inhalte der ersten vier Kapitel dienen.)

5.1 WAHRSCHEINLICHKEITSRAUM

Wir beginnen zunächst mit dem grundlegenden Wahrscheinlichkeitsraum. Naheliegenderweise wählen wir als Ergebnisraum $\Omega = \mathbb{R}$ (den Fall $\Omega \subsetneq \mathbb{R}$ betrachten wir in Kürze). Da $\mathbb{R}$ überabzählbar ist, können wir als Ereignisraum nicht mehr die Potenzmenge wählen; stattdessen wählen wir die *Borel-Algebra* aus Beispiel 1.6 (ii).

Definition 5.1 (Borel-Algebra). Sei

$$\mathcal{E} := \{(a, b) \mid a, b \in \mathbb{R},\ a < b\}\,.$$

Dann ist die *Borel-Algebra* auf $\mathbb{R}$ definiert als

$$\mathcal{B} := \sigma(\mathcal{E}) := \bigcap_{\mathcal{E} \subset \mathcal{A}\ \sigma\text{-Algebra}} \mathcal{A},$$

d. h. die von $\mathcal{E}$ erzeugte σ-Algebra (die kleinste σ-Algebra, die $\mathcal{E}$ enthält. Eine Menge $A \in \mathcal{B}$ nennt man *Borel-Menge*.

Man verifiziert leicht anhand der Definition, dass $\mathcal{B}$ als Schnitt von σ-Algebren wieder eine σ-Algebra ist. Diese enthält

- alle offenen Mengen (als abzählbare Vereinigung von offenen Intervallen);
- alle abgeschlossenen Mengen (als Komplement von offenen Mengen);

- praktisch alle explizit angebbaren Mengen (als beliebig oft wiederholte Vereinigung von offenen und abgeschlossenen Mengen).

Trotzdem existieren Mengen – die man nicht explizit angeben kann – die keine Borel-Mengen sind. Für unsere Zwecke ist wichtig, dass man die Borel-Algebra auch auf andere Weise erzeugen kann.

Lemma 5.2. *Sei*

$$\mathcal{E}' := \{(-\infty, c] \mid c \in \mathbb{R}\}.$$

Dann gilt $\mathcal{B} = \sigma(\mathcal{E}')$.

Wir werden sehen, dass diese halboffenen Intervalle für reelle Zufallsvariablen die selbe Rolle spielen wie die Elementarereignisse für diskrete Zufallsvariablen.

Ist $\Omega \subsetneq \mathbb{R}$ nichtleer, dann erhalten wir durch

$$\mathcal{B}(\Omega) := \{A \cap \Omega \mid A \in \mathcal{B}\}$$

eine σ-Algebra auf Ω, ebenfalls genannt *Borel-Algebra* auf Ω. Offensichtlich gilt $\mathcal{B}(\Omega) \subset \mathcal{B}$ für $\Omega \in \mathcal{B}$.

Wir kommen nun zum wesentlichen – und schwierigen – Punkt: Der Definition von Wahrscheinlichkeitsverteilungen, die nur mit Mitteln der Maß- und Integrationstheorie möglich ist. Wir skizzieren dies wegen ihrer Wichtigkeit kurz; Details findet man zum Beispiel in [Brokate & Kersting 2019]. Das zentrale Resultat ist das folgende.

Satz 5.3. *Es gibt genau eine σ-additive Abbildung $\lambda : \mathcal{B} \to [0, \infty]$, genannt* Lebesgue-Maß, *mit*

$$\lambda((a, b)) = b - a \qquad \text{für alle } a < b \in \mathbb{R}.$$

Die Einschränkung auf Borel-Mengen ist hier wesentlich! Der schwierige Teil ist hier die Existenz; da sowohl $\mathcal{E}$ als auch $\mathcal{E}'$ schnittstabil sind (d. h. $A, B \in \mathcal{E}'$ impliziert $A \cap B \in \mathcal{E}'$), folgt die Eindeutigkeit aus Satz 1.15. Durch Normierung erhält man daraus sofort die Existenz einer Gleichverteilung auf Borel-Mengen (und nur diesen!)

Folgerung 5.4 (reelle Gleichverteilung). *Sei $\Omega \in \mathcal{B}$ mit $\lambda(\Omega) \in (0, \infty)$. Dann wird durch*

$$\mathbb{P}[(a, b)] = \frac{b - a}{\lambda(\Omega)} \qquad \text{für alle } a, b \in \mathbb{R} \text{ mit } a < b$$

eine eindeutige Wahrscheinlichkeitsverteilung auf $\mathcal{B}(\Omega)$ definiert. Die so definierte Verteilung

$$\mathbb{P} : \mathcal{B}(\Omega) \to [0, \infty), \qquad \mathbb{P}[A] := \frac{\lambda(A)}{\lambda(\Omega)} \quad \text{für alle } A \in \mathcal{B}(\Omega),$$

wird (reelle) Gleichverteilung *auf $\mathcal{B}(\Omega)$ genannt und mit $\mathcal{U}(\Omega)$ bezeichnet.*

Die Existenz des Lebesgue-Maß kann genutzt werden, um analog zu Satz 1.11 weitere Wahrscheinlichkeitsverteilungen zu konstruieren. Da Ω nicht mehr abzählbar ist, kommen wir dafür nicht mehr mit einer Summe bzw. Reihe aus, sondern benötigen einen passenden Integralbegriff: das *Lebesgue-Integral*. Ansatz ist hier – anders als im Riemann-Integral – eine Zerlegung des *Wertebereichs* einer Funktion. Sei zunächst $f : \mathbb{R} \to [0, \infty)$ eine Funktion, die auf $\Omega \subset \mathbb{R}$ höchstens abzählbar viele verschiedene Werte annimmt, d. h.

$$f(x) = \sum_c c \cdot \mathbb{1}_{\{f=c\}}(x) \qquad \text{für alle } x \in \Omega,$$

wobei die Summe über höchstens abzählbar viele $c \in \mathbb{R}$ geht. Gilt nun $\Omega \in \mathcal{B}$ und $\{f = c\} := \{x \in \Omega \mid f(x) = c\} \in \mathcal{B}(\Omega)$ für alle c, so definiert man das Lebesgue-Integral von f als

$$\int_\Omega f(x)\, d\lambda(x) := \sum_c c \cdot \lambda(\{f = c\}).$$

Durch Approximation und Grenzübergang erhält man daraus den folgenden allgemeinen Integralbegriff, zunächst nur für nichtnegative Funktionen.

Satz 5.5 (Lebesgue-Integral). *Sei $\Omega \in \mathcal{B}$ und $f : \Omega \to [0, \infty)$ so, dass gilt*

$$\{x \in \Omega \mid f(x) \le c\} \in \mathcal{B}(\Omega) \qquad \text{für alle } c > 0. \tag{5.1}$$

Dann existiert das Lebesgue-Integral $\int_\Omega f(x)\, d\lambda(x) \in [0, \infty]$ *und es gilt*

(i) ist f Riemann-integrierbar, so ist $\int_\Omega f(x)\, d\lambda(x) = \int_\Omega f(x)\, dx$;

(ii) erfüllen $f_n : \Omega \to [0, \infty)$, $n \in \mathbb{N}$ die Bedingung (5.1), so ist

$$\int_\Omega \sum_n f_n(x)\, d\lambda(x) = \sum_n \int_\Omega f_n(x)\, d\lambda(x).$$

Die Bedingung (5.1) wird auch als *Meßbarkeit* bezeichnet; sie ist insbesondere für alle stetigen Funktionen erfüllt. Eigenschaft (i) erlaubt, für konkrete Berechnungen mit „typischen Funktionen“ auf die Techniken aus Analysis 1 zurückzugreifen; die σ-Additivität in Eigenschaft (ii) ist aber wesentlich, um das Analogon zu Satz 1.15 beweisen zu können.

Satz 5.6 (Wahrscheinlichkeitsdichte). *Sei $\Omega \in \mathcal{B}$ und $\rho : \Omega \to [0, \infty)$ so, dass gilt*

(i) $\{x \in \Omega \mid \rho(x) \le c\} \in \mathcal{B}(\Omega)$ für alle $c > 0$;

(ii) $\int_\Omega \rho(x)\, d\lambda(x) = 1$.

Dann wird durch

$$\mathbb{P}[A] := \int_A \rho(x)\, d\lambda(x), \qquad A \in \mathcal{B}(\Omega),$$

eine eindeutige Wahrscheinlichkeitsverteilung auf $\mathcal{B}(\Omega)$ definiert. In diesem Fall nennt man ρ Wahrscheinlichkeitsdichte *oder* Dichtefunktion *von $\mathbb{P}$.*

Die Gleichverteilung entspricht hier genau dem Fall $\rho(x) = \lambda(\Omega)^{-1}$. Allerdings hat anders als im diskreten Fall nicht jede Wahrscheinlichkeitsverteilung eine Dichtefunktion; auch entspricht $\rho(x)$ *nicht* mehr unbedingt der „Wahrscheinlichkeit von x“; tatsächlich gilt oft $\mathbb{P}[\{x\}] = 0$ auch für $\rho(x) > 0$. Sinnvolle Aussagen sind also in der Regel nur über Ereignisse möglich, die durch „echte“ Intervalle beschrieben werden.

Eine besondere Rolle (nicht nur) in der Wahrscheinlichkeitstheorie spielt die folgende Wahrscheinlichkeitsdichte.

Definition 5.7 (Normalverteilung). Sei $m, \sigma \in \mathbb{R}$. Die durch die Wahrscheinlichkeitsdichte

$$\rho_{m,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

definierte Wahrscheinlichkeitsverteilung wird *Normalverteilung* mit Mittelwert m und Varianz σ^2 auf $\mathbb{R}$ genannt und mit $\mathcal{N}_{m,\sigma}$ bezeichnet. Ist $m = 0$ und $\sigma = 1$, so spricht man von der *Standardnormalverteilung*.

(Dass $\rho_{m,\sigma}$ eine Wahrscheinlichkeitsdichte ist, folgt aus der offensichtlichen Stetigkeit sowie der – nicht-trivialen – Berechnung von $\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\, dx = \frac{1}{\sqrt{2\pi}}$ zusammen mit der Substitutionsformel.)

Weitere durch Dichtefunktionen definierte Wahrscheinlichkeitsverteilungen, die man manchmal in freier Wildbahn antrifft, sind die folgenden:

Beispiel 5.8 (Wahrscheinlichkeitsdichten).

(i) die *Cauchy-Verteilung* mit Parametern $s > 0$ und $t \in \mathbb{R}$, definiert durch

$$\rho_{s,t}(x) = \frac{s}{\pi(s^2 + (x-t)^2)};$$

(ii) die *Gamma-Verteilung* mit Parametern $a \in \mathbb{N}$ und $b > 0$, definiert durch

$$\rho_{a,b}(x) = \begin{cases} \frac{b^a x^{a-1}}{a!} e^{-bx} & \text{falls } x > 0, \\ 0 & \text{sonst.} \end{cases}$$

Den allgemeinen Fall $f : \Omega \to \mathbb{R}$ können wir durch Zerlegung von $f = f^+ - f^-$ mit $f^+, f^- \geq 0$ behandeln. Ist $\int_\Omega f^+(x)\, d\lambda(x) < \infty$ und $\int_\Omega f^-(x)\, d\lambda(x) < \infty$, so nennt man f *(Lebesgue-)integrierbar* und setzt

$$\int_\Omega f(x)\, d\lambda(x) := \int_\Omega f^+(x)\, d\lambda(x) - \int_\Omega f^-(x)\, d\lambda(x).$$

Aus dieser Konstruktion folgen sofort die vom Riemann-Integral bekannten Eigenschaften wie Linearität und Monotonie.

5.2 ZUFALLSVARIABLEN UND VERTEILUNGSFUNKTIONEN

Wir haben nun alles zur Hand, um reelle Zufallsvariablen zu definieren.

Definition 5.9 (reelle Zufallsvariable). Seien $(\Omega, \mathcal{A})$ ein Ereignisraum. Eine Funktion

$$X : \Omega \to \mathbb{R}$$

heißt *reelle Zufallsvariable* von $(\Omega, \mathcal{A})$ nach $(\mathbb{R}, \mathcal{B})$, wenn gilt

$$\{X \leq c\} = \{\omega \in \Omega \mid X(\omega) \in (-\infty, c]\} \in \mathcal{A} \qquad \text{für alle } c \in \mathbb{R}.$$

Hier haben wir wieder Beispiel 1.6 verwendet, um die allgemeine Definition 2.1 auf die schnittstabile Erzeugendenmenge $\mathcal{E}'$ einschränken zu dürfen. (Vergleiche auch (5.1)!) Nach Satz 2.3 (der für allgemeine Zufallsvariablen gilt) definiert eine reelle Zufallsvariable X eine Wahrscheinlichkeitsverteilung $\mathbb{P}_X$ auf $\mathcal{B}$ durch

$$\mathbb{P}_X : \mathcal{B} \to [0, \infty], \qquad \mathbb{P}_X\left[(-\infty, c]\right] := \mathbb{P}[X \leq c] \qquad \text{für alle } c \in \mathbb{R}.$$

Die Wahrscheinlichkeitsverteilung $\mathbb{P}_X$ wird also eindeutig charakterisiert durch die Werte $F_X(c) := \mathbb{P}_X[(-\infty, c]]$ für alle $c \in \mathbb{R}$. Dies legt die folgende Definition nahe.

Definition 5.10 (Verteilungsfunktion). Sei $X : \Omega \to \mathbb{R}$ eine reelle Zufallsvariable auf $(\Omega, \mathcal{A})$. Dann heißt

$$F_X : \mathbb{R} \to [0, 1], \qquad F_X(c) := \mathbb{P}[X \leq c],$$

die *(kumulative) Verteilungsfunktion* von X.

Direkt aus den Eigenschaften von Wahrscheinlichkeitsverteilungen erhält man die folgenden Eigenschaften.

Lemma 5.11. *Sei $X : \Omega \to \mathbb{R}$ eine reelle Zufallsvariable auf $(\Omega, \mathcal{A})$ mit Verteilungsfunktion F_X. Dann gilt*

(i) *F_X ist monoton wachsend;*

(ii) *$\lim_{c \to -\infty} F_X(c) = 0$ und $\lim_{c \to \infty} F_X(c) = 1$;*

(iii) *$F_X(b) - F_X(a) = \mathbb{P}[a < X \leq b]$ für alle $a, b \in \mathbb{R}$ mit $a < b$;*

(iv) *F_X ist* rechtsseitig stetig, *d. h.* *$\lim_{t \to c+} F_X(t) = F_X(c)$ für alle $c \in \mathbb{R}$.*

Dagegen gilt für den linksseitigen Grenzwert nur

$$\lim_{t \to c^-} F_X(t) = F_X(c) - \mathbb{P}[X = c],$$

d. h. F_X ist stetig in c genau dann, wenn $\mathbb{P}[X = c] = 0$ ist.

Wir kommen nun zum versprochenen Analogon der Zähldichte für reelle Zufallsvariablen.

Definition 5.12. Sei $X : \Omega \to \mathbb{R}$ eine reelle Zufallsvariable auf $(\Omega, \mathcal{A})$. Existiert eine integrierbare Funktion $\rho_X : \mathbb{R} \to [0, \infty)$ mit

$$F_X(c) = \int_{-\infty}^{c} \rho_X(x)\, d\lambda(x),$$

so nennt man X *(absolut)stetig* und ρ_X *Dichtefunktion* von X.

Nach Satz 1.15 und Definition der Verteilungsfunktion gilt für eine absolutstetige Zufallsvariable stets

$$\mathbb{P}[X \in B] = \mathbb{P}_X[B] = \int_B \rho_X(x)\, d\lambda(x) \qquad \text{für alle } B \in \mathcal{B}.$$

In Gegensatz zu diskreten Zufallsvariablen hat aber nicht jede reelle Zufallsvariable eine Dichtefunktion; reelle Zufallsvariablen ohne Dichtefunktion sind aber schwierig zu konstruieren und in der Praxis selten anzutreffen. Eine Dichtefunktion existiert insbesondere immer, wenn die Verteilungsfunktion stetig differenzierbar ist; nach den Hauptsatz der Differential- und Integralrechnung und Lemma 5.11 (ii) gilt dann nämlich

$$F_X(c) = F_X(c) - \lim_{t \to -\infty} F_X(t) = \int_{-\infty}^{c} F_X'(x)\, dx$$

d. h. F_X' ist Dichtefunktion von F_X.

Beispiel 5.13 (Dichtefunktionen).

(i) Ist X eine gleichverteilte reelle Zufallsvariable auf (a, b), so ist X absolutstetig mit Dichtefunktion

$$\rho_X(x) = \begin{cases} \frac{1}{b-a} & \text{falls } x \in (a, b), \\ 0 & \text{sonst,} \end{cases}$$

und Verteilungsfunktion

$$F_X(c) = \begin{cases} 0 & \text{für } c \le a, \\ \frac{c-a}{b-a} & \text{für } a < c < b, \\ 1 & \text{für } c \ge b. \end{cases}$$

(ii) Ist X normalverteilt mit Mittelwert m und Varianz σ^2, so ist X nach Definition absolutstetig mit Dichtefunktion

$$\rho_X(x) = \rho_{m,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

Die Verteilungsfunktion hat keine bekannte geschlossene Darstellung und kann nur näherungsweise (z. B. numerisch) ausgewertet werden; Statistik-Bücher enthalten oft seitenweise Tabellen für diesen Zweck.

Beispiel 5.14 (Exponentialverteilung). Die *Exponentialverteilung* beschreibt die Wartezeit auf seltene Vorfälle (z. B. den nächsten Telefonanruf). Für die Herleitung gehen wir vor wie in Kapitel 2 und unterteilen für beliebiges $t > 0$ das Intervall $(0, t]$ in n gleich lange Teilintervalle $(\frac{(k-1)t}{n}, \frac{kt}{n}]$, $k = 1, \dots, n$. Sei A_k das Ereignis, dass der Vorfall im k-ten Intervall eintritt. Wieder machen wir die Modellannahme, dass die Ereignisse stochastisch unabhängig sind und ihre Wahrscheinlichkeit nur von der Intervalllänge abhängt, d. h. dass $\mathbb{P}[A_k] = \lambda \frac{t}{n}$ für ein $\lambda > 0$ gilt. Sei nun $T : \Omega \to (0, \infty)$ eine reelle Zufallsvariable, die den Eintrittszeitpunkt angibt. Dann folgt aus der angenommenen Unabhängigkeit und Lemma 4.19

$$\begin{aligned}
\mathbb{P}[T > t] &= \mathbb{P}\left[\{T \in (0,t]\}^c\right] = \mathbb{P}\left[(A_1 \cup \dots \cup A_n)^c\right] = \mathbb{P}\left[A_1^c \cap \dots \cap A_n^c\right] \\
&= \mathbb{P}\left[A_1^c\right] \cdots \mathbb{P}\left[A_n^c\right] = (1 - \mathbb{P}[A_1]) \cdots (1 - \mathbb{P}[A_n]) \\
&= \left(1 - \frac{\lambda t}{n}\right)^n \to e^{-\lambda t}.
\end{aligned}$$

Also hat T die Verteilungsfunktion

$$F_T(t) = \mathbb{P}[T \le t] = \begin{cases} 1 - e^{-\lambda t} & \text{falls } t > 0, \\ 0 & \text{sonst.} \end{cases}$$

Diese Funktion ist stetig und sowohl auf $(-\infty, 0]$ als auch auf $(0, \infty)$ stetig differenzierbar, und daher hat T die Dichtefunktion

$$\rho_T(t) = \rho_\lambda(t) := F_T'(t) = \begin{cases} \lambda e^{-\lambda t} & \text{falls } t > 0, \\ 0 & \text{sonst.} \end{cases}$$

Eine reelle Zufallsvariable mit Dichtefunktion ρ_λ heißt *exponentialverteilt* mit Parameter $\lambda > 0$, geschrieben $T \sim \mathrm{Exp}(\lambda)$. (Dies ist ein Spezialfall der Gamma-Verteilung für $a = 1$ und $b = \lambda$.)

5.3 ERWARTUNGSWERT UND VARIANZ

Für reelle Zufallsvariablen mit Dichtefunktion kann man den Erwartungswert analog zum diskreten Fall definieren.

Definition 5.15 (Erwartungswert für absolutstetige Zufallsvariablen). Sei X eine reelle Zufallsvariable auf $(\Omega, \mathcal{B}(\Omega))$ mit Dichtefunktion ρ_X. Ist das Lebesgue-Integral

$$\mathbb{E}[X] := \int_\Omega x\rho_X(x)\, d\lambda(x)$$

wohldefiniert, so nennt man $\mathbb{E}[X]$ *Erwartungswert* von X (bezüglich dem Lebesgue-Maß). Dies ist insbesondere dann der Fall, wenn gilt

$$\int_\Omega |x|\rho_X(x)\, d\lambda(x) < \infty;$$

wir schreiben dafür kurz $X \in L^1(\Omega)$.

Wieder sieht man aus der Definition sofort, dass der Erwartungswert (nur) verteilungsabhängig ist. (Tatsächlich kann man den Erwartungswert noch allgemeiner definieren, um sowohl diskrete als auch reelle Zufallsvariablen gemeinsam behandeln zu können; dies ist aber endgültig zu technisch für den Rahmen dieser Vorlesung.)

Beispiel 5.16 (Erwartungswerte).

(i) Ist X gleichverteilt auf $\Omega = (a, b)$, dann gilt

$$\mathbb{E}[X] = \int_a^b x\frac{1}{b-a}\, dx = \frac{1}{b-a}\left(\frac{1}{2}b^2 - \frac{1}{2}a^2\right) = \frac{1}{2}(b+a).$$

(ii) Ist X normalverteilt auf $\Omega = \mathbb{R}$ mit Mittelwert 0 und Varianz 1, dann gilt wegen der Symmetrie des Integranden

$$\begin{aligned}\mathbb{E}[X] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-\frac{1}{2}x^2}\, dx &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} xe^{-\frac{1}{2}x^2}\, dx + \frac{1}{\sqrt{2\pi}}\int_0^{\infty} xe^{-\frac{1}{2}x^2}\, dx \\ &= \frac{1}{\sqrt{2\pi}}\int_0^{\infty} -xe^{-\frac{1}{2}(-x)^2}\, dx + \frac{1}{\sqrt{2\pi}}\int_0^{\infty} xe^{-\frac{1}{2}x^2}\, dx \\ &= 0.\end{aligned}$$

(iii) Ist X exponentialverteilt mit Parameter $\lambda > 0$, dann folgt mit partieller Integration

$$\begin{aligned}\mathbb{E}[X] &= \int_0^{\infty} x \cdot \lambda e^{-\lambda x}\, dx = \int_0^{\infty} (-x)\left(-\lambda e^{-\lambda x}\right) dx \\ &= \left[(-x)\cdot e^{-\lambda x}\right]_{x=0}^{x=\infty} - \int_0^{\infty} (-1)(e^{-\lambda x})\, dx = 0 + \left[\frac{e^{-\lambda x}}{-\lambda}\right]_{x=0}^{x=\infty} = \frac{1}{\lambda}.\end{aligned}$$

Ähnlich wie in Satz 3.8, allerdings mit erheblich höherem technischen Aufwand, erhält man einen Transformationssatz für den Erwartungswert von absolutstetigen Zufallsvariablen.

Satz 5.17 (Transformationssatz). *Sei X eine reelle Zufallsvariable nach $(\Omega, \mathcal{B}(\Omega))$ mit Dichtefunktion ρ_X und $g : \mathbb{R} \to \mathbb{R}$ eine beliebige Funktion. Dann ist auch $g(X) := g \circ X$ eine reelle Zufallsvariable mit Erwartungswert*

$$\mathbb{E}[g(X)] = \int_\Omega g(x)\rho_X(x)\, d\lambda(x),$$

falls dieses Integral wohldefiniert ist.

Daraus folgt wie im diskreten Fall die Linearität des Erwartungswerts für integrierbare Zufallsvariablen.

Satz 5.18. *Seien $X, Y \in L^1(\Omega)$ absolutstetige Zufallsvariablen. Dann gilt für alle $\alpha, \beta \in \mathbb{R}$ auch $\alpha X + \beta Y \in L^1(\Omega)$ und*

$$\mathbb{E}[\alpha X + \beta Y] = \alpha\, \mathbb{E}[X] + \beta\, \mathbb{E}[Y].$$

Beispiel 5.19. Ist X normalverteilt mit Mittelwert 0 und Varianz 1, so ist $Z := m + \sigma X$ normalverteilt mit Mittelwert m und Varianz σ^2. Aus Satz 5.18 und Beispiel 5.16 folgt dann sofort

$$\mathbb{E}[Z] = m + \sigma\, \mathbb{E}[X] = m.$$

Der Erwartungswert einer normalverteilten Zufallsvariable ist also genau ihr Mittelwert.

Aus den Eigenschaften des Lebesgue-Integrals folgt auch für absolutstetige Zufallsvariablen die Monotonie des Erwartungswerts.

Lemma 5.20. *Sei X eine nichtnegative absolutstetige Zufallsvariable. Dann gilt*

(i) $\mathbb{E}[X] \geq 0$;

(ii) $\mathbb{E}[X] = 0$ *genau dann, wenn* $\mathbb{P}[X = 0] = 1$ *gilt.*

Folgerung 5.21. *Seien $X, Y \in L^1(\Omega)$ absolutstetige Zufallsvariablen. Gilt $X \geq Y$, dann gilt auch*

$$\mathbb{E}[X] \geq \mathbb{E}[Y].$$

Genau wie für diskrete Zufallsvariablen können wir nun für absolutstetige Zufallsvariablen die Varianz über den Erwartungswert definieren:

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

falls gilt

$$\mathbb{E}[X^2] = \int_\Omega x^2 \rho_X(x)\, d\lambda(x) < \infty;$$

wir schreiben in dem Fall $X \in L^2(\Omega)$. Alle für diskrete Zufallsvariablen in ℓ^2 bewiesenen Eigenschaften wie Translationsformel, Nichtnegativität, Chebyshev-Ungleichung, Cauchy-Schwarz-Ungleichung, und Bienaymé-Formel gelten also auch für absolutstetige Zufallsvariablen in $L^2(\Omega)$.

Beispiel 5.22 (Varianzen).

(i) Ist X gleichverteilt auf $\Omega = (a, b)$, dann gilt mit Substitution

$$\begin{aligned}\mathbb{V}[X] &= \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{1}{b-a}\, dx = \frac{1}{b-a} \int_{a-\frac{1}{2}(a+b)}^{b-\frac{1}{2}(a+b)} x^2\, dx \\ &= \frac{1}{b-a}\left[\frac{1}{3}\left(\frac{b-a}{2}\right)^3 - \frac{1}{3}\left(\frac{a-b}{2}\right)^3\right] = \frac{1}{3}\frac{(b-a)^2}{8} + \frac{1}{3}\frac{(b-a)^2}{8} \\ &= \frac{(b-a)^2}{12}.\end{aligned}$$

(ii) Ist X normalverteilt auf $\Omega = \mathbb{R}$ mit Mittelwert 0 und Varianz 1, dann erhalten wir mit partieller Integration

$$\begin{aligned}\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}x^2}\, dx - 0 \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (-x)\left(-xe^{-\frac{1}{2}x^2}\right) dx \\ &= \frac{1}{\sqrt{2\pi}} \left(\left[-xe^{-\frac{1}{2}x^2}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\, dx\right) \\ &= 1\end{aligned}$$

da der erste Term für $x \to \pm\infty$ gegen Null geht und $\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\, dx = \sqrt{2\pi}$ ist.

Aus der Translationsformel folgt dann für $Z = m + \sigma X \sim \mathcal{N}_{m,\sigma}$, wie nicht anders zu erwarten,

$$\mathbb{V}[Z] = \sigma^2\, \mathbb{V}[X] = \sigma^2.$$

Eine Normalverteilung ist also durch ihren Erwartungswert und ihre Varianz eindeutig festgelegt!

(iii) Ist X exponentialverteilt mit Parameter $\lambda > 0$, dann folgt mit partieller Integration

$$\begin{aligned}\mathbb{E}[X^2] &= \int_0^\infty x^2 \cdot \lambda e^{-\lambda x}\, dx = \int_0^\infty (-x^2)\left(-\lambda e^{-\lambda x}\right) dx \\ &= \left[(-x^2)e^{\lambda x}\right]_{x=0}^{x=\infty} - \int_0^\infty (-2x)e^{-\lambda x}\, dx \\ &= \frac{2}{\lambda}\int_0^\infty x(\lambda e^{-\lambda x})\, dx = \frac{2}{\lambda}\,\mathbb{E}[X].\end{aligned}$$

Also ist

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{2}{\lambda}\cdot\frac{1}{\lambda} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

Je länger die mittlere Wartezeit ist, desto größer ist also die Unsicherheit, wie lange man warten muss.

5.4 BEDINGTE WAHRSCHEINLICHKEIT UND UNABHÄNGIGKEIT

Da wir bedingte Wahrscheinlichkeiten für allgemeine Wahrscheinlichkeitsräume definiert haben, bleiben alle Definitionen und Sätze (insbesondere der Satz 4.8 von Bayes) für reelle Wahrscheinlichkeitsräume unverändert gültig.

Beispiel 5.23 (Gedächtnislosigkeit der Exponentialverteilung). Sei T exponentialverteilt mit Parameter $\lambda > 0$ und beschreibe den Zeitpunkt eines erwarteten Telefonanrufs. Angenommen, wir haben schon $s > 0$ Stunden gewartet. Was ist die Wahrscheinlichkeit, dass wir noch weitere $t > 0$ Stunden warten müssen? Wir suchen also die bedingte Wahrscheinlichkeit

$$\begin{aligned}\mathbb{P}[T > s+t \mid T > s] &= \frac{\mathbb{P}[\{T > s+t\} \cap \{T > s\}]}{\mathbb{P}[T > s]} = \frac{\mathbb{P}[T > s+t]}{\mathbb{P}[T > s]} = \frac{1 - F_T(s+t)}{1 - F_T(s)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = 1 - F_T(t) = \mathbb{P}[T > t].\end{aligned}$$

Egal wie lange wir also bereits gewartet haben, die Wahrscheinlichkeit für das Eintreten ist nicht größer als am Anfang! (Die Exponentialverteilung „merkt sich“ also nicht, wie lange schon gewartet wurde.)

Auch die Unabhängigkeit von Ereignissen definieren wir völlig analog: Zwei Ereignisse $A_1, A_2 \in \mathcal{B}$ heißen unabhängig, wenn gilt

$$\mathbb{P}[A_1 \cap A_2] = \mathbb{P}[A_1]\,\mathbb{P}[A_2].$$

Definition 5.24 (unabhängige Zufallsvariablen). Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum.

Eine Familie $\{X_i\}_{i\in I}$ von reellen Zufallsvariablen $X_i : \Omega \to \mathbb{R}$ für eine nichtleere Menge I heißt *unabhängig* bezüglich $\mathbb{P}$, falls für jede nichtleere endliche Teilmenge $J \subset I$ gilt

$$\mathbb{P}\left[\bigcap_{i\in J}\{X_i \le c_i\}\right] = \prod_{i\in J}\mathbb{P}[X_i \le c_i] \qquad \text{für alle } c_i \in \mathbb{R}.$$

Durch Grenzübergang erhält man mit Hilfe der Rechnung aus Satz 4.21 das folgende Resultat.

Satz 5.25. *Seien $\{X_i\}_{i\in I} \subset L^2(\Omega)$ unabhängige reelle Zufallsvariablen. Dann gilt $\operatorname{Cov}[X_i, X_j] = 0$ für alle $i \neq j$ und damit*

$$\begin{aligned}\mathbb{E}[X_iX_j] &= \mathbb{E}[X_i]\cdot\mathbb{E}[X_j],\\ \mathbb{V}[X_i + X_j] &= \mathbb{V}[X_i] + \mathbb{V}[X_j].\end{aligned}$$

Damit erhält man nun ganz analog Gesetze der großen Zahl.

Satz 5.26 (schwaches Gesetz der großen Zahl). *Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $\{X_n\}_{n\in\mathbb{N}} \subset L^2(\Omega)$ eine Folge von unabhängigen reellen Zufallsvariablen mit $\mathbb{E}[X_n] = m$ und $\mathbb{V}[X_n] = M$ für alle $n \in \mathbb{N}$. Dann gilt*

$$\lim_{n\to\infty}\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}X_i - m\right| \ge \varepsilon\right] = 0 \qquad \textit{für alle } \varepsilon > 0.$$

Satz 5.27 (starkes Gesetz der großen Zahl). *Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $\{X_n\}_{n\in\mathbb{N}} \subset L^2(\Omega)$ eine Folge von unabhängigen reellen Zufallsvariablen mit $\mathbb{E}[X_n] = m$ und $\mathbb{V}[X_n] = M$ für alle $n \in \mathbb{N}$. Dann gilt*

$$\mathbb{P}\left[\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}X_i = m\right] = 1.$$

Für den Beweis der letzten Aussage – der auf allgemeinen Resultaten für Zufallsvariablen basiert – siehe wieder [Krengel 2005, Satz 12.4]. Wir können also auch für reelle Zufallsvariablen *empirische Schätzer* für Erwartungswert und Varianz verwenden.

Wir betrachten nun unabhängige reelle Zufallsvariablen genauer über ihre Dichten. Dafür brauchen wir etwas Vorarbeit. Seien $X_i : \Omega \to \mathbb{R}$, $i = 1, \dots, n$, reelle Zufallsvariablen mit Verteilung $\mathbb{P}_{X_i} : \mathcal{B} \to [0, 1]$. Dann ist auch $X := (X_1, \dots, X_n)$ eine Zufallsvariable von $(\Omega, \mathcal{A})$ nach $(\mathbb{R}^n, \mathcal{B}^n)$, wobei die *Borel-Algebra* auf $\mathbb{R}^n$ erzeugt wird von Mengen der Form

$$A := (-\infty, c_1] \times \cdots \times (-\infty, c_n], \qquad c_1, \dots, c_n \in \mathbb{R}.$$

Daraus folgt, dass X eine Zufallsvariable ist genau dann, wenn die X_i Zufallsvariablen sind. Die *gemeinsame Verteilung* der X_i ist dann definiert durch

$$\mathbb{P}_{X_1,\dots,X_n}[A] = \mathbb{P}\left[\{X_1 \le c_1\} \cap \dots \cap \{X_n \le c_n\}\right], \qquad c_1, \dots, c_n \in \mathbb{R}.$$

Die gemeinsame Verteilung heißt *absolutstetig*, wenn eine integrierbare Funktion $\rho_X : \mathbb{R}^n \to [0, \infty)$ existiert mit

$$\mathbb{P}_{X_1,\dots,X_n}[A] = \int_{-\infty}^{c_1} \cdots \int_{-\infty}^{c_n} \rho_X(x_1, \dots, x_n)\, d\lambda(x_n) \dots d\lambda(x_1).$$

Aus der Vertauschbarkeit von Integralen nach dem Satz von Fubini erhalten wir damit die folgende Charakterisierung von Unabhängigkeit absolutstetiger Zufallsvariablen.

Satz 5.28. *Seien $X_i : \Omega \to \mathbb{R}$, $i = 1, \dots, n$ absolutstetige Zufallsvariablen mit Dichtefunktion $\rho_{X_i} : \mathbb{R} \to [0, 1]$. Dann sind die X_i unabhängig genau dann, wenn die gemeinsame Verteilung $\mathbb{P}_{X_1,\dots,X_n}$ absolutstetig ist mit Dichtefunktion*

$$\rho_X(x_1, \dots, x_n) = \rho_{X_1}(x_1) \cdots \rho_{X_n}(x_n), \qquad x_1, \dots, x_n \in \mathbb{R}.$$

Wir wenden dies nun auf die Summe von absolutstetigen Zufallsvariablen an. Wieder aus dem Satz von Fubini erhält man die folgende Darstellung.

Satz 5.29. *Seien X und Y unabhängige absolutstetige Zufallsvariablen mit Dichtefunktion ρ_X bzw. ρ_Y. Dann ist auch $X + Y$ absolutstetig mit Dichtefunktion*

$$\rho_{X+Y}(x) = \int_{-\infty}^{\infty} \rho_X(z)\rho_Y(x - z)\, d\lambda(z.$$

Beweis. Sei $c \in \mathbb{R}$ beliebig. Dann gilt $X+Y \le c$ genau dann, wenn $(X, Y) \in \left\{(x, y) \in \mathbb{R}^2 \,\middle|\, y \le c - x\right\}$ ist. Wegen der Unabhängigkeit von X und Y und der Vertauschbarkeit von Integralen erhalten wir daraus mit Hilfe der Substitution $z = x + y$

$$\begin{aligned}
\mathbb{P}[X + Y \le c] &= \int_{-\infty}^{\infty} \int_{-\infty}^{c-x} \rho_{(X,Y)}(x, y)\, d\lambda(y)\, d\lambda(x) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{c-x} \rho_X(x)\rho_Y(y)\, d\lambda(y)\, d\lambda(x) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{c} \rho_X(x)\rho_Y(z - x)\, d\lambda(z)\, d\lambda(x) \\
&= \int_{-\infty}^{c} \int_{-\infty}^{\infty} \rho_X(x)\rho_Y(z - x)\, d\lambda(x)\, d\lambda(z).
\end{aligned}$$

Vertauschen von x und z im letzten Integral ergibt die Behauptung. □

Analog zum diskreten Fall nennt man das Integral auf der rechten Seite *Faltung* von ρ_X und ρ_Y.

Beispiel 5.30 (Faltung von Standardverteilungen). Seien X und Y unabhängig gleichverteilt auf $(0,1)$ mit Dichtefunktion $\rho_X(x) = \rho_Y(x) = \mathbb{1}_{(0,1)}(x)$. Durch Fallunterscheidung erhält man daraus

$$\rho_{X+Y}(x) = \int_{-\infty}^{\infty} \mathbb{1}_{(0,1)}(z)\mathbb{1}_{(0,1)}(x-z)\,dz = \int_{\max\{0,x-1\}}^{\min\{x,1\}} 1\,dz = \begin{cases} x & x \in (0,1], \\ 2-x & x \in (1,2], \\ 0 & \text{sonst.} \end{cases}$$

Mit etwas mehr Aufwand rechnet man auf ähnliche Weise nach, dass für unabhängige normalverteilte $X_1 \sim \mathcal{N}(m_1, \sigma_1^2)$ und $X_2 \sim \mathcal{N}(m_2, \sigma_2^2)$ gilt

$$\rho_{X_1+X_2}(x) = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} e^{-\frac{(x-(m_1+m_2))^2}{2(\sigma_1^2+\sigma_2^2)}}\,dx,$$

d. h $X + Y \sim \mathcal{N}(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$.

Zusammen mit den Rechenregeln für die Normalverteilung (aus der Substitutionsformel) folgt daraus, dass für unabhängige $X_1, \ldots, X_n \sim \mathcal{N}(m, \sigma^2)$ gilt

$$S_n := \frac{X_1 + \cdots + X_n - nm}{\sqrt{n}} \sim \mathcal{N}(0, \sigma^2).$$

Seien nun $X_1, \ldots X_n$ *beliebige* unabhängige reelle Zufallsvariablen mit $\mathbb{E}[X_i] = m$ und $\mathbb{V}[X_i] = \sigma^2$ für alle $i = 1, \ldots, n$. Dann folgt aus Satz 5.25 ebenfalls

$$\mathbb{E}[S_n] = 0, \qquad \mathbb{V}[S_n] = \sigma^2.$$

Dies sagt natürlich noch nichts über die gesamte Verteilung von S_n aus. Mit Hilfe der Faltungsformel, Taylor-Entwicklung, und einer sorgfältigen Abschätzung der Rest-Terme kann man aber zeigen, dass S_n in einem geeigneten Sinn tatsächlich gegen eine Normalverteilung konvergiert. Das folgende Resultat ist einer der zentralen Sätze in der Wahrscheinlichkeitstheorie und erklärt, warum die Normalverteilung so häufig anzutreffen ist (nämlich immer dann, wenn sehr viele unabhängige, identische, Prozesse irgendwie „gemittelt“ werden).

Satz 5.31 (zentraler Grenzwertsatz). *Sei $\{X_n\}_{n\in\mathbb{N}}$ eine Folge von unabhängigen und identisch verteilten Zufallsvariablen mit $\mathbb{E}[X_n] = m$ und $\mathbb{V}[X_n] = \sigma^2$ für alle $n \in \mathbb{N}$. Dann gilt für $S_n := \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n - nm)$ und alle $a \le b$*

$$\lim_{n\to\infty} \mathbb{P}\left[a \le S_n \le b\right] = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\,dx.$$

Teil II

STATISTIK

6 SCHÄTZER

Bisher haben wir Zufallssituationen mathematisch modelliert, indem wir einen festen (hoffentlich passenden) Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ gewählt haben und, darauf aufbauend, Wahrscheinlichkeiten von Ereignissen bestimmt haben (meistens mit Hilfe geeigneter Zufallsvariablen). In der Realität ist es jedoch oft so, dass die „richtige" Wahrscheinlichkeitsverteilung nicht oder nicht vollständig bekannt ist; stattdessen hat man nur Beobachtungen von tatsächlich auftretenden Ergebnissen zur Hand und möchte daraus auf die zugrundeliegende Wahrscheinlichkeitsverteilung – oder zumindest daraus abgeleitete Informationen wie z. B. Erwartungswerte – schließen. Dies ist Aufgabe der *Statistik*.

Die typische Situation ist die folgende: Angenommen, wir haben eine Zufallsvariable X, deren Verteilung wir nicht kennen, für die wir aber n „Auswertungen" oder *Stichproben* zur Hand haben. Typische Fragestellungen – für die wir schon Beispiele gesehen haben – sind:

(i) Wir wissen, dass X Bernoulli-verteilt ist, kennen aber den Parameter $p \in (0,1)$ nicht. Da in dem Fall gilt $p = \mathbb{E}[X]$, können wir diesen Parameter durch den Erwartungswert *schätzen*, etwa durch den *empirischen Mittelwert*; siehe Beispiel 3.35. (Da der Schätzer eine einzige Zahl ist, spricht man auch von einem *Punktschätzer*.)

(ii) Wir wollen zusätzlich eine Fehlerangabe haben, etwa durch ein *Konfidenzintervall*; siehe ebenfalls Beispiel 3.35. (Hier spricht man von einem *Intervallschätzer*.)

(iii) Wir sind eigentlich nur an der Frage interessiert, ob der Münzwurf fair ist, d. h. ob $p = \frac{1}{2}$ ist oder nicht – wir wollen also zwischen der *Hypothese*, dass die Münze fair ist, und der Alternative entscheiden. Hier spricht man von einem *Hypothesentest*; vergleiche Beispiel 4.10.

Hierbei ist es wichtig zu beachten, dass die Antworten auf diese Fragen auf der – zufälligen! – *Realisierung* der Stichproben, die wir für die Rechnung verwenden, beruhen, und daher selber wieder eine Zufallsvariable mit einer (nicht vollständig bekannten) Wahrscheinlichkeitsverteilung sind. Dies erfordert einen passenden mathematischen Rahmen. Dafür sind wir bereits gut vorbereitet; da sich die Statistik in gewissen Dingen parallel zur mathematischen Wahrscheinlichkeitstheorie entwickelt hat, ist die Terminologie aber eine eigene.

Wir beginnen wieder mit den möglichen (Beobachtungs-)Ergebnissen, die den *Stichprobenraum* Ξ bilden; aus dieser wird durch eine Beobachtung ein zufälliges Element $x \in \Xi$, genannt *Stichprobe* ausgewählt. Dies könnte zum Beispiel sein

(i) die Anzahl der Würfe „Zahl“ bei n Münzwürfen – dann ist $\Xi = \{0, 1, \dots, n\}$;

(ii) die n verschiedenen, mit einem zufälligen Messfehler behafteten, wiederholten Messungen einer reellen physikalischen Größe – dann ist $\Xi = \mathbb{R}^n$.

Auf diesem Stichprobenraum definieren wir wieder eine σ-Algebra $\mathcal{A}$; dafür wählt man kanonischerweise

(i) die Potenzmenge $\mathcal{P}(\Xi)$, falls Ξ höchstens abzählbar ist;

(ii) die Borelalgebra $\mathcal{B}^n$, falls $\Xi = \mathbb{R}^n$ ist.

Schließlich zur Wahrscheinlichkeitsverteilung und damit zum wesentlichen Unterschied zur Wahrscheinlichkeitstheorie. Anstelle einer festen Wahrscheinlichkeitsverteilung $\mathbb{P}$ müssen wir hier eine ganze Familie von Verteilungen betrachten. Dafür betrachten wir eine nichtleere, mindestens zweielementige, Menge Θ, genannt *Parameterraum*, und für jeden *Parameter* $\vartheta \in \Theta$ eine Wahrscheinlichkeitsverteilung $\mathbb{P}_\vartheta : \mathcal{A} \to [0, 1]$. Beispielsweise könnten wir betrachten

(i) $\Theta = (0, 1)$ und $\mathbb{P}_\vartheta = \mathrm{Bin}(n, \vartheta)$ die Binomialverteilung für ein festes $n \in \mathbb{N}$;

(ii) $\Theta = \mathbb{R}^2$ und $\mathbb{P}_\vartheta = \mathcal{N}(\vartheta_1, \vartheta_2^2)$ für $\vartheta = (\vartheta_1, \vartheta_2)$;

(iii) $\Theta = \{1, 2\}$ und $\mathbb{P}_1 = \mathcal{U}((0, 100))$ und $\mathbb{P}_2 = \mathrm{Exp}(1)$.

Um die Abhängigkeit vom *Parameter* ϑ zu verdeutlichen, schreiben wir auch $\mathbb{E}_\vartheta$ und $\mathbb{V}_\vartheta$ für den daraus abgeleiteten Erwartungswert bzw. Varianz.

Das Tripel $(\Xi, \mathcal{A}, \{\mathbb{P}_\vartheta \mid \vartheta \in \Theta\})$ nennt man *statistisches Modell*. Ist $\Theta \subset \mathbb{R}^d$, so spricht man von einem *(d-)parametrischen Modell*. Ist Ξ höchstens abzählbar, nennt man das Modell *diskret*; in diesem Fall hat jedes $\mathbb{P}_\vartheta$ eine Zähldichte $p_\vartheta : \Xi \to [0, 1]$. Ist $\Xi \subset \mathbb{R}^n$ und hat jedes $\mathbb{P}_\vartheta$ eine Dichtefunktion $\rho_\vartheta^n : \Xi \to [0, 1]$, so spricht man analog von einem *stetigen Modell*. Gilt in letzterem Fall speziell

$$\rho_\vartheta^n(x_1, \dots, x_n) = \rho_\vartheta(x_1) \cdots \rho_\vartheta(x_n), \qquad (x_1, \dots, x_n) \in \Xi,$$

für ein $\rho_\vartheta : \mathbb{R} \to [0, 1]$, so spricht man von einem *Produktmodell*.

Das Produktmodell ist nun der geeignete Rahmen, um die eingangs beschriebene Situation der n-fachen Realisierung der unbekannten Zufallsvariable X zu beschreiben: Statt n verschiedener Realisierungen $x = (X(\omega_1), \dots X(\omega_n))$ beschreiben wir die Stichprobe als *eine* Realisierung von n Zufallsvariablen $x = (X_1(\omega), \dots X_n(\omega)) =: \mathcal{X}(\omega)$; dies hat den Vorteil, dass die *Zufallsstichprobe* $\mathcal{X}$ auf einem von n unabhängigen Wahrscheinlichkeitsraum definiert sind. Da man aber in der in der Statistik leider üblicherweise nicht angibt, auf welchen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ die unbekannte Zufallsvariable X definiert

ist, behandelt man die Zufallsstichprobe $\mathcal{X}$ als die Identität auf Ξ, d. h. $\mathcal{X}(x) = x$. Das Produktmodell entspricht dann nach Satz 5.28 genau der Situation, dass die einzelnen Zufallsstichproben $X_i \sim X$ (jetzt ebenfalls auf Ξ!) unabhängig und identisch verteilt sind mit Dichtefunktion $\rho_{\bar{\vartheta}}$ für einen (unbekannten) *wahren Parameter* $\bar{\vartheta} \in \Theta$. (Das Modell muss also die „wahre Situation" beschreiben können; sonst erhält man potentiell unsinnige Antworten!)

6.1 PUNKTSCHÄTZER

Auf Basis eines statistischen Modells betrachten wir nun die einleitenden typischen Fragestellungen, beginnend mit Punktschätzern. Sei $(\Xi, \mathcal{A}, \{\mathbb{P}_\vartheta \mid \vartheta \in \Theta\})$ ein statistisches Modell und $(\Sigma, \mathcal{S})$ ein weiterer Ereignisraum. Dann nennt man eine beliebige Zufallsvariable von $(\Xi, \mathcal{A})$ nach $(\Sigma, \mathcal{S})$ auch *Statistik*. Konkret ist man an Statistiken für Abbildungen $\tau : \Theta \to \Sigma$ interessiert, die jedem Parameter $\vartheta \in \Theta$ eine Kenngröße $\tau(\vartheta) \in \Sigma$ zuweisen. Mögliche Beispiele sind

(i) $\Sigma = \Theta$ und $\tau(\vartheta) = \vartheta$ (insbesondere für $\Theta \subset \mathbb{R}$);

(ii) $\Sigma \supset \Theta$ und $\tau(\vartheta) = \vartheta$ (z. B. für $\Theta = (0, 1)$ aber $T(\Theta) = [0, 1]$);

(iii) $\Sigma = \mathbb{R}$ und $\tau(\vartheta) = \mathbb{E}_\vartheta[X]$.

Solch eine Statistik wird *(Punkt-)Schätzer für* τ genannt, um zu verdeutlichen, dass wir uns dafür interessieren, wie verlässlich dieser Schätzer für τ ist. (Die Definition von T nimmt ja keinen Bezug auf τ.) Ist wie in (i) und (ii) τ die Identität – die gesuchte Kenngröße also der unbekannte Parameter selber – spricht man auch von einem *Parameterschätzer*. Beachten Sie, dass ein Schätzer nichts anderes als eine Zufallsvariable $T : \Xi \to \Sigma$ ist; in der Praxis wird man für eine konkrete Stichprobe $x \in \Xi$ den *konkreten Schätzwert* $T(x) \in \Sigma$ berechnen.

Wie man von der sehr abstrakten Definition erwarten kann, existieren zahlreiche verschiedene Punktschätzer mit verschiedenen Eigenschaften. Eine spezielle Klasse von Punktschätzern für $\tau(\vartheta) = \vartheta$ sind die *maximum likelihood-Schätzer*, die für eine konkrete Beobachtung $x \in \Xi$ den „plausibelsten" (Englisch: „most likely") Parameter berechnen – nämlich den, für den die Beobachtung „am wahrscheinlichsten" ist. Um das Maximum dieser Wahrscheinlichkeit zu berechnen, definiert man sich für gegebenes $x \in \Xi$ eine geeignete *Likelihood-Funktion* $L_x : \Theta \to [0, 1]$, und zwar

(i) für ein *diskretes* statistisches Modell $L_x(\vartheta) := \mathbb{P}_\vartheta[\mathcal{X} = x] = p_\vartheta(x)$;

(ii) für ein *stetiges* statistisches Modell $L_x(\vartheta) = \rho_\vartheta(x)$.

(Beachten Sie, dass für stetige Zufallsvariablen in der Regel $\mathbb{P}_\vartheta[\mathcal{X} = x] = 0$ gilt!)

Ein Schätzer $T : \Xi \to \Sigma \supset \Theta$ heißt nun *maximum likelihood-Schätzer*, falls gilt

$$L_x(T(x)) = \sup_{\vartheta \in \Theta} L_x(\vartheta) \qquad \text{für alle } x \in \Xi.$$

Beachten Sie, dass das Supremum nicht unbedingt in Θ angenommen werden muss, und dass der Schätzer nicht eindeutig sein muss.

Speziell für Produktmodelle kann es günstiger sein, die *log-Likelihood-Funktion*

$$\log \rho_\vartheta^n(x) = \log\left(\rho_\vartheta(x_1) \cdots \rho_\vartheta(x_n)\right) = \sum_{i=1}^{n} \log \rho_\vartheta(x_i)$$

zu maximieren; wegen der Monotonie des Logarithmus hat sie den selben Maximierer (wenn auch nicht das selbe Maximum) wie die Likelihood-Funktion.

Beispiel 6.1 (Binomial-Modelle). Wir wollen den Parameter in einer Binomialverteilung schätzen, z. B. die Wahrscheinlichkeit, dass eine geworfene Münze auf „Zahl“ landet. Wir werfen dafür die Münze n Mal und beobachten die Anzahl der Würfe mit „Zahl“. Dies entspricht dem diskreten statistischen Modell

$$(\{0, 1, \ldots, n\}, \mathcal{P}(\{0, 1, \ldots, n\}), \{\mathrm{Bin}(n, \vartheta) \mid \vartheta \in (0, 1)\}).$$

Wir wählen als Likelihood-Funktion also die Zähldichte

$$p_\vartheta(x) = \binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x}$$

mit log-Likelihood-Funktion

$$\log p_\vartheta(x) = \log \binom{n}{x} + x \log \vartheta + (n - x) \log(1 - \vartheta).$$

Einen Kandidaten für den Maximierer über alle $\vartheta \in (0, 1)$ erhalten wir durch die Lösung von

$$0 = \frac{d}{d\vartheta} \log p_\vartheta(x) = \frac{x}{\vartheta} - \frac{n - x}{1 - \vartheta},$$

die gegeben ist durch $\hat{\vartheta} = \frac{x}{n} \in [0, 1]$. Man verifiziert leicht z. B. mit Hilfe der zweiten Ableitung, dass dies in der Tat ein Maximierer ist.
Mit der Zufallsstichprobe $X(x) = x$, $x \in \Xi$, ist der gesuchte maximum likelihood-Schätzer also $T = \frac{X}{n}$ mit Schätzwert $T(x) = \frac{X(x)}{n} = \frac{x}{n}$.

Beispiel 6.2 (gleichverteilte Modelle). Als etwas konstruiertes Beispiel wollen wir den Zahlenbereich eines Zufallszahlengenerators schätzen. Unsere Beobachtung dafür sind n Zufallszahlen, die auf einem Intervall $(0, a)$ mit $a > 0$ unbekannt gleichverteilt sind.

Dies entspricht dem stetigen statistischen Produktmodell

$$((0,\infty)^n, \mathcal{B}^n, \{\mathcal{U}((0,\vartheta)^n) \mid \vartheta \in (0,\infty)\}).$$

Hier wählen wir als Likelihood-Funktion die Dichtefunktion

$$\rho_\vartheta^n(x) = \begin{cases} \frac{1}{\vartheta^n} & \text{für } x = (x_1,\dots,x_n) \in (0,\vartheta)^n, \\ 0 & \text{sonst.} \end{cases}$$

Man überlegt sich leicht, dass hier das Supremum in

$$\hat{\vartheta} := \max\{x_1,\dots,x_n\} \in [0,\infty)$$

angenommen wird. Mit der Zufallsstichprobe $\mathcal{X} = (X_1,\dots,X_n)$, $\mathcal{X}(x) = x$, erhalten wir also den maximum likelihood-Schätzer $T = \max\{X_1,\dots,X_n\}$.

Beispiel 6.3 (normalverteilte Modelle). Wir messen eine physikalische Größe X, z. B. eine Temperatur, was wegen zufälliger Messfehler niemals den exakten Wert ergeben kann. Nimmt man an, dass die Messfehler durch Mittelung vieler unabhängiger Prozesse (z. B. thermisches Rauschen, Brownsche Bewegung) entstehen, kann man diese Messung als normalverteilt mit Mittelwert $m \in \mathbb{R}$ (dem wahren Wert) und Varianz $\sigma^2 > 0$ (dem unbekannten Fehlerniveau) modellieren. Wir wiederholen nun die Messung n-mal und wollen daraus den exakten Wert und das Fehlerniveau schätzen. Dies entspricht dem stetigen statistischen Produktmodell

$$(\mathbb{R}^n, \mathcal{B}^n, \{\mathcal{N}(\vartheta_1,\vartheta_2)^n \mid \vartheta_1 \in \mathbb{R}, \vartheta_2 > 0\}).$$

Hier wählen wir als Likelihood-Funktion also die Dichtefunktion

$$\rho_\vartheta^n(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\vartheta_2}} e^{-\frac{1}{2\vartheta_2}(x_i-\vartheta_1)^2} = \frac{1}{\left(\sqrt{2\pi\vartheta_2}\right)^n} e^{-\frac{1}{2\vartheta_2}\sum_{i=1}^n (x_i-\vartheta_1)^2}$$

mit der log-Likelihood-Funktion

$$\log \rho_\vartheta^n(x) = -\frac{n}{2}\left(\log\vartheta_2 + \log(2\pi)\right) - \frac{1}{2\vartheta_2}\sum_{i=1}^n (x_i-\vartheta_1)^2.$$

Glücklicherweise erlaubt uns diese Struktur, das Maximum bezüglich ϑ_1 und ϑ_2 der Reihe nach zu bestimmen. Wir betrachten zuerst die Lösung von

$$0 = \frac{d}{d\vartheta_1}\log\rho_\vartheta^n(x) = -\frac{1}{\vartheta_2}\sum_{i=1}^n (x_i-\vartheta_1).$$

Wegen $\vartheta_2 > 0$ genügt es, dass die Summe verschwindet, was unabhängig von ϑ_2 für

$$\hat{\vartheta}_1 := \frac{1}{n}\sum_{i=1}^n x_i = \hat{M}(x)$$

mit $\hat{M} := \frac{1}{n}\sum_{i=1}^{n} X_i$, d. h. den empirischen Mittelwert, der Fall ist (zweite Ableitung prüfen!)
Wir halten nun $\vartheta_1 = \hat{\vartheta}_1$ fest und maximieren nach ϑ_2. Wir nehmen zuerst an, dass $x_i \neq \hat{M}(x)$ für ein $i = 1, \dots, n$ gilt und lösen

$$0 = \frac{d}{d\vartheta_2} \log \rho_\vartheta^n(x) = -\frac{n}{2\vartheta_2} + \frac{1}{2\vartheta_2^2} \sum_{i=1}^{n} (x_i - \hat{\vartheta}_1)^2$$

Multiplikation mit $\vartheta_2 > 0$ und Auflösen ergibt dann mit $\hat{\vartheta}_1 = \hat{M}(x)$

$$\hat{\vartheta}_2 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{M}(x))^2 = \hat{V}(x)$$

mit $\hat{V} := \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{M})^2 > 0$ nach Annahme, was genau die empirische Varianz ist. Ist $x_i = \hat{M}(x)$ für alle $i = 1 \dots, n$, so ist $\hat{V}(x) = 0$, was wegen

$$\log \rho^n_{(\hat{\vartheta}_1, \vartheta_2)}(x) = -\frac{n}{2} \left(\log \vartheta_2 + \log(2\pi)\right) \longrightarrow \infty$$

für $\vartheta_2 \to 0$ auch in dem Fall das Supremum der Dichtefunktion ergibt.
Also ist unser gesuchter maximum Likelihood-Schätzer bzw. Schätzwert

$$T = (\hat{M}, \hat{V}), \qquad T(x) = (\hat{M}(x), \hat{V}(x)) \in \mathbb{R} \times [0, \infty).$$

Wir fragen uns jetzt nach der Güte von Schätzern. Eine Möglichkeit, diese zu beurteilen, ist zu prüfen, ob der Schätzer „im Mittel“ über alle möglichen Stichproben den richtigen Wert liefert. Wir nennen deshalb einen Schätzer $T : \Xi \to \mathbb{R}$ *erwartungstreu* für die reelle Kenngröße $\tau : \Theta \to \mathbb{R}$, falls gilt

$$\mathbb{E}_\vartheta[T] = \tau(\vartheta) \qquad \text{für alle } \vartheta \in \Theta.$$

Wir betrachten beispielhaft die beiden Schätzer aus Beispiel 6.3.

Beispiel 6.4. Zunächst schauen wir uns $T = \hat{M} = \frac{1}{n}\sum_{i=1}^{n} X_i$ an als Schätzer für die Kenngröße $\tau(\vartheta) = \mathbb{E}_\vartheta[X]$ für die normalverteilte Größe X an. Dann folgt aus der Linearität des Erwartungswerts und der identischen Verteilung der Stichproben für beliebiges $\vartheta \in \Theta$ sofort

$$\mathbb{E}_\vartheta[\hat{M}] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\vartheta[X_i] = \frac{1}{n}(n\, \mathbb{E}_\vartheta[X]) = \mathbb{E}_\vartheta[X].$$

Dieser Schätzer ist also erwartungstreu.
Wir betrachten nun $T = \hat{V} = \frac{1}{n}\sum_{i=1}^{N}(X_i - \hat{M})^2$ als Schätzer für $\tau(\vartheta) = \mathbb{V}_\vartheta[X]$. Aus der Linearität des Erwartungswerts, der Definition der Varianz, der Erwartungstreue von

$\hat{M}$, und der Unabhängigkeit der identisch verteilten Stichproben folgt dann

$$\begin{aligned}\mathbb{E}_\vartheta[\hat{V}] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\vartheta[(X_i-\hat{M})^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{V}_\vartheta[X_i-\hat{M}] + \left(\mathbb{E}_\vartheta[X_i-\hat{M}]\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{V}_\vartheta\left[\frac{n-1}{n}X_i + \frac{1}{n}\sum_{j\neq i}X_j\right] \\ &= \left(\frac{n-1}{n}\right)^2\mathbb{V}_\vartheta[X] + \frac{n-1}{n^2}\mathbb{V}_\vartheta[X] \\ &= \frac{n-1}{n}\mathbb{V}_\vartheta[X].\end{aligned}$$

Also ist $\hat{V}$ *nicht* erwartungstreu; dafür aber (wieder wegen der Linearität des Erwartungswerts) der *korrigierte Schätzer*

$$\hat{V}^* := \frac{n}{n-1}\hat{V} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{M})^2.$$

Ist T nicht erwartungstreu, so nennt man

$$B_T : \Theta \to \mathbb{R}, \qquad B_T(\vartheta) := \mathbb{E}_\vartheta[T] - \tau(\vartheta) = \mathbb{E}_\vartheta[T-\tau(\vartheta)],$$

die *Verzerrung* (Englisch: *bias*) von T. Dieser ist nützlich, ein weiteres Gütekriterium zu untersuchen: den *mittleren Fehler*

$$E_T : \Theta \to \mathbb{R}, \qquad E_T(\vartheta) := \mathbb{E}_\vartheta\left[(T-\tau(\vartheta))^2\right] = \mathbb{V}_\vartheta(T) + B_T(\vartheta)^2,$$

wobei die zweite Gleichung wieder aus der Definition der Varianz folgt; man spricht daher auch von einer *Bias-Varianz-Zerlegung* des mittleren Fehlers. Ein guter Schätzer hat also sowohl kleinen Bias als auch kleine Varianz für alle $\vartheta \in \Theta$, wobei eine kleine Varianz oft einen größeren Bias bedingt – in diesem Fall wird die Summe minimal, wenn beide Terme die selbe Größe haben.

Abschließend untersuchen wir die Frage nach der Abhängigkeit von der Stichprobengröße. Dazu betrachten wir für jedes $n \in \mathbb{N}$ wieder ein Produktmodell mit $\Xi = \mathbb{R}^n$ und einen entsprechenden Schätzer $T_n : \Xi \to \mathbb{R}$. Dann heißt die Folge $\{T_n\}_{n\in\mathbb{N}}$ *konsistent* für die Kenngröße $\tau : \Theta \to \mathbb{R}$, wenn gilt

$$\mathbb{P}_\vartheta\left[\lim_{n\to\infty} T_n = \tau(\vartheta)\right] = 1 \qquad \text{für alle } \vartheta \in \Theta.$$

Beispiel 6.5. Für den Schätzer $T = \hat{M}$ aus Beispiel 6.3 folgt die Konsistenz für den Erwartungswert direkt aus dem starken Gesetz der großen Zahl, Satz 5.27. Mit etwas

Basteln und der Subadditivität und der Monotonie von Wahrscheinlichkeitsverteilungen erhält man daraus auch die Konsistenz von $\hat{V}$ und $\hat{V}^*$.

6.2 BEREICHSCHÄTZER

Auch für einen erwartungstreuen und konsistenten Punktschätzer kann der einzelne Schätz*wert* beliebig schlecht sein (wir können ja immer „Pech“ mit der Stichprobe haben). Um zu wissen, wie verlässlich der Schätzer ist, benötigen wir Fehlerschranken – d. h. Intervallschätzer.

Sei $(\Xi, \mathcal{A}, \{\mathbb{P}_\vartheta \mid \vartheta \in \Theta\})$ ein statistisches Modell und $(\Sigma, \mathcal{S})$ ein weiterer Ereignisraum. Sei weiter $\tau : \Theta \to \Sigma$ eine Kenngröße und $\alpha \in (0,1)$. Eine Abbildung $C : \Xi \to \mathcal{S}$ heißt dann *Bereichschätzer* für die Kenngröße τ zum *Konfidenzniveau* $1-\alpha$, wenn gilt

$$\mathbb{P}_\vartheta[\tau(\vartheta) \in C] = \mathbb{P}_\vartheta\left[\{x \in \Xi \mid \tau(\vartheta) \in C(x)\}\right] \geq 1-\alpha \qquad \text{für alle } \vartheta \in \Theta.$$

Ist $\Sigma \subset \mathbb{R}$ und $C(x)$ ein Intervall (und damit eine Borelmenge) für alle $x \in \Xi$, so spricht man auch von einem *Intervallschätzer*. Die Menge $C(x)$ für eine konkrete Stichprobe $x \in \Xi$ nennt man *Konfidenzbereich* bzw. *Konfidenzintervall*.

Wir verlangen also, dass – unabhängig vom (unbekannten) wahren Parameterwert – die Wahrscheinlichkeit, dass die wahre Kenngröße *nicht* im Konfidenzbereich $C(x)$ liegt, kleiner ist als das *Irrtumsniveau* α. Natürlich ist nicht die Kenngröße sondern der Konfidenzbereich zufällig; es ist also besser zu sagen, dass $C(x)$ die wahre Kenngröße mit einer Wahrscheinlichkeit von mindestens $1-\alpha$ überdeckt. Man spricht daher auch von einer *Überdeckungswahrscheinlichkeit* $1-\alpha$. Typische Werte von α sind 0.1, 0.05, oder 0.01.

Auch hier ist die Definition extrem allgemein gehalten, um einen gemeinsamen Rahmen für sehr verschiedene, in der Praxis verwendete, Bereichsschätzer zu haben. Natürlich ist nicht jeder Bereichschätzer sinnvoll: z. B. definiert offensichtlich $C(x) := \Sigma$ für alle $x \in \Xi$ einen Schätzer zu jedem Konfidenzniveau. Die Herausforderung besteht darin, für möglichst großes Konfidenzniveau $1-\alpha$ einen möglichst kleinen Konfidenzbereich $C(x)$ zu konstruieren. Je mehr Informationen über das statistische Modell wir in den Schätzer stecken, desto besser wird das möglich sein (um den Preis, dass dieser Schätzer nur für dieses Modell funktioniert).

Beispiel 6.6. Wir betrachten zum Beispiel $\tau(\vartheta) = \mathbb{E}_\vartheta[X]$ für eine beliebige reelle Zufallsvariable X mit dem Produktmodell

$$(\mathbb{R}^n, \mathcal{B}^n, \{\mathbb{P}_\vartheta \mid \vartheta \in \Theta\}),$$

wobei wir annehmen, dass $\mathbb{V}_\vartheta[X] \leq M$ für ein bekanntes $M > 0$ ist. Als Schätzer für τ verwenden wir wieder den empirischen Mittelwert $\hat{M} := \frac{1}{n}\sum_{i=1}^n X_i$. Aus dem schwachen

Gesetz der großen Zahl (siehe Satz 3.32) folgt dann für jedes $\varepsilon > 0$ und beliebiges $\vartheta \in \Theta$

$$\mathbb{P}_\vartheta \left[|\hat{M} - \mathbb{E}_\vartheta[X]| \geq \varepsilon \right] \leq \frac{M}{n\varepsilon^2}.$$

Wir wählen nun für z. B. ein Konfidenzniveau $1 - \alpha = 0.95$ ein $\varepsilon > 0$ so, dass $\frac{M}{n\varepsilon^2} < \alpha = 0.05$ gilt. Dann ist

$$\mathbb{P}_\vartheta \left[\hat{M} - \varepsilon < \mathbb{E}_\vartheta[X] < \hat{M} + \varepsilon \right] = 1 - \mathbb{P}_\vartheta \left[|\hat{M}\, \mathbb{E}_\vartheta[X]| \geq \varepsilon \right] \geq 1 - 0.05 = 0.95,$$

d.h. $(\hat{M} - \varepsilon, \hat{M} + \varepsilon)$ ist ein Konfidenzintervall zum Konfidenzniveau 0.95; vergleiche Beispiel 3.35. Hier ist nur die Lage abhängig von der Stichprobe, während die Länge nur von α, M und n abhängt: je größer n und je kleiner M, desto kleiner das Intervall; umgekehrt wird das Intervall umso größer, desto kleiner α gewählt ist.

Ein allgemeines Konstruktionsprinzip basiert wieder auf der Likelihood-Funktion $L_x(\vartheta)$ z. B. der Zähldichte $p_\vartheta(x)$ für diskrete Modelle und der Dichtefunktion $\rho_\vartheta(x)$ für stetige Modelle. Für ein gegebenes Irrtumsniveau $\alpha \in (0, 1)$ geht man dafür wie folgt vor:

(i) Für jedes $\vartheta \in \Theta$ bestimme eine Menge

$$C_\vartheta := \{x \in \Xi \mid L_x(\vartheta) \geq c_\vartheta\},$$

wobei c_ϑ gerade so groß gewählt ist, dass $\mathbb{P}_\vartheta[C_\vartheta] \geq 1 - \alpha$ ist.

(ii) Für gegebenes $x \in \Xi$ wähle dann

$$C(x) := \{\vartheta \in \Theta \mid x \in C_\vartheta\}.$$

Für (hinreichend kleine) diskrete Modelle ist dies durch Erstellung einer Tabelle, für die für alle Paare (x, ϑ) die Likelihood $L_x(\vartheta)$ eingetragen wird, möglich; im ersten Schritt sammelt man für jedes ϑ so lange x-Einträge – beginnend mit dem größten – bis die Summe der entsprechenden Zähldichten größer als $1 - \alpha$ ist. Für gegebenes x kann man dann den Konfidenzbereich direkt ablesen.

Für stetige Modelle verwendet man stattdessen die Verteilungsfunktion.

Beispiel 6.7. Wir betrachten wieder das statistische Modell

$$(\mathbb{R}^n, \mathcal{B}^n, \{\mathcal{N}(\vartheta_1, \vartheta_2)^n \mid \vartheta_1 \in \mathbb{R}, \vartheta_2 > 0\}).$$

für eine normalverteilte Zufallsvariable mit unbekanntem Mittelwert und Varianz. Um speziell für den Mittelwert $m := \vartheta_1 = \tau(\vartheta)$ einen Bereichsschätzer zu definieren, gehen wir wie folgt vor. Wir nehmen zuerst an, die Varianz $\sigma^2 := \vartheta_2$ wäre bekannt. Wir betrachten dann für den (erwartungstreuen und konsistenten) Punktschätzer $\hat{M} = \frac{1}{n} \sum X_i$ für m. Aus Beispiel 5.30 zusammen mit den Rechenregeln des Erwartungswerts

und der Varianz folgt wegen der Unabhängigkeit

$$\hat{M} \sim \mathcal{N}(\hat{m}, \hat{\sigma}^2) \quad \text{für} \quad \hat{m} = \mathbb{E}[\hat{M}] = \frac{1}{n}(nm) = m, \quad \hat{\sigma}^2 = \mathbb{V}[\hat{M}] = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

Daraus folgt umgekehrt, dass für alle $m \in \mathbb{R}$

$$\hat{S}_m := \frac{\hat{M} - m}{\sqrt{n}^{-1}\sigma} \sim \mathcal{N}(0,1)$$

standardnormalverteilt (bezüglich $\mathbb{P}_m$!) ist. Mit Hilfe der Verteilungsfunktion $F_{\mathcal{N}}$ der Standardnormalverteilung können wir also für beliebiges $t > 0$ schreiben

$$\mathbb{P}_m\left[-t \le \hat{S}_m \le t\right] = F_{\mathcal{N}}(t) - F_{\mathcal{N}}(-t) = 2F_{\mathcal{N}}(t) - 1,$$

wegen der Symmetrie der Dichtefunktion. Zu vorgegebenem $\alpha \in (0,1)$ können wir also ein $t_\alpha > 0$ mit $2F_{\mathcal{N}}(t_\alpha) - 1 = 1 - \alpha$ (d. h. $t_\alpha = F_{\mathcal{N}}^{-1}(1 - \frac{\alpha}{2})$) finden so dass (nach Einsetzen der Definition von $\hat{S}_m$ und Umformen) gilt

$$\mathbb{P}_m\left[\hat{M} - t_\alpha \frac{\sigma}{\sqrt{n}} \le m \le \hat{M} + t_\alpha \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

für alle $m \in \mathbb{R}$. Unser gesuchtes Konfidenzintervall ist also

$$C(x) = \left[\hat{M}(x) - t_\alpha \frac{\sigma}{\sqrt{n}}, \hat{M}(x) + t_\alpha \frac{\sigma}{\sqrt{n}}\right].$$

Ist nun σ^2 auch unbekannt, verwenden wir stattdessen den (erwartungstreuen und konsistenten) Schätzer

$$\hat{V}^* = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{M})^2$$

und betrachten analog

$$\hat{S}_m := \frac{\hat{M} - m}{\sqrt{n}^{-1}\hat{V}^*}.$$

Man kann nun zeigen, dass auch $\hat{S}_m$ eine von m und σ^2 unabhängige Verteilung hat, die sogenannte *Studentsche t-Verteilung mit $n-1$ Freiheitsgraden*. Diese ist absolutstetig mit explizit angebbarer (wenn auch komplizierter) Dichtefunktion, so dass man wieder mit Hilfe ihrer Verteilungsfunktion zu vorgegebenem $\alpha \in (0,1)$ ein $t_\alpha > 0$ finden kann, so dass gilt

$$\mathbb{P}_m\left[\hat{M} - t_\alpha \frac{\hat{V}^*}{\sqrt{n}} \le m \le \hat{M} + t_\alpha \frac{\hat{V}^*}{\sqrt{n}}\right] = 1 - \alpha$$

für alle $m \in \mathbb{R}$, d. h. unser gesuchtes Konfidenzintervall ist

$$C(x) = \left[\hat{M}(x) - t_\alpha \frac{\hat{V}^*(x)}{\sqrt{n}}, \hat{M}(x) + t_\alpha \frac{\hat{V}^*(x)}{\sqrt{n}}\right].$$

7 HYPOTHESENTESTS

Oft ist man nicht an einer konkreten Kenngröße interessiert, sondern möchte anhand einer Stichprobe bloß entscheiden, ob diese Größe „im grünen Bereich“ ist oder eine kritische Situation vorliegt. Das kanonische Beispiel ist die Frage, ob eine geworfene Münze fair ist; aber auch andere Situationen im Handel („Ist die gelieferte Ware fehlerhaft?“), in der Medizin („Liegt eine Krankheit vor?“) oder in der Produktion („Sind alle Schrauben korrekt angezogen?“) haben diese Struktur.

Wieder betrachten wir einen sehr allgemeinen Rahmen, ausgehend von einem statistischen Modell $(\Xi, \mathcal{A}, \{\mathbb{P}_\vartheta \mid \vartheta \in \Theta\})$. Wir zerlegen nun den Parameterraum Θ in zwei disjunkte Teilmengen

- die *Nullhypothese* $\Theta_0 \subset \Theta$ (der „gute“ Fall; z. B. $\Theta_0 = \{\frac{1}{2}\}$ „die Münze ist fair“);
- die *Alternative* $\Theta_1 = \Theta \setminus \Theta_0$ (der „schlechte“ Fall; z. B. $\Theta_1 = (0,1) \setminus \{\frac{1}{2}\}$ „die Münze ist *nicht* fair“).

Besteht Θ_0 nur aus einem Element, spricht man von einer *einfachen Hypothese*. Eine Statistik $\varphi : \Xi \to [0,1]$ heißt dann *Test von* Θ_0 *gegen* Θ_1, wenn wir φ als Entscheidungsregel interpretieren: Für eine gegebene Stichprobe $x \in \Xi$

- akzeptieren wir die Nullhypothese, wenn $\varphi(x) = 0$ ist, und
- verwerfen wir die Nullhypothese, wenn $\varphi(x) = 1$ ist.

Entsprechend nennt man

- $\{x \in \Xi \mid \varphi(x) = 0\}$ den *Akzeptanzbereich* und
- $\{x \in \Xi \mid \varphi(x) = 1\}$ den *Verwerfungsbereich*.

Ist $\varphi(x) \in \{0,1\}$ für alle $x \in \Xi$, so nennen wir den Test *nichtrandomisiert* (es gibt immer eine klare Entscheidung), ansonsten *randomisiert* (für $\varphi(x) \in (0,1)$ müssen wir uns überlegen, ob das Ergebnis für das Verwerfen ausreicht). (Für einen nichtrandomisierten Test ist φ genau die charakteristische Funktion des Verwerfungsbereichs!)

Wieder ist das Ergebnis für eine einzelne Stichprobe nicht aussagekräftig; um die Qualität eines Tests zu bewerten, interessieren wir uns für die Wahrscheinlichkeit über alle Stichproben von zwei verschiedenen Fehlern:

- *Fehler 1. Art*, dass wir die Nullhypothese zu Unrecht verwerfen ($\varphi(x) = 1$ obwohl $\vartheta \in \Theta_0$);
- *Fehler 2. Art*, dass wir die Nullhypothese zu Unrecht annehmen ($\varphi(x) = 0$ obwohl $\vartheta \in \Theta_1$).

Dafür definieren wir für einen Test φ die *Gütefunktion*

$$G_\varphi : \Theta \to [0,1], \qquad G_\varphi(\vartheta) := \mathbb{E}_\vartheta[\varphi],$$

und nennen

- $\sup_{\vartheta \in \Theta_0} G_\varphi(\vartheta)$ den *Umfang* oder das *Niveau* von φ;
- $G_\varphi(\vartheta)$ die *Macht* oder *Schärfe* von φ bei $\vartheta \in \Theta_1$.

In anderen Worten, der Umfang ist die Wahrscheinlichkeit für einen Fehler 1. Art, während die Macht die Wahrscheinlichkeit angibt, die Alternative korrekt zu erkennen (bzw. $1 - G_\varphi(\vartheta)$ ist die Wahrscheinlichkeit für einen Fehler 2. Art in $\vartheta \in \Theta_1$). Offensichtlich wollen wir gleichzeitig ein niedriges Niveau und eine große Macht, wobei sich beides in der Regel widerspricht. Wir nennen einen Test φ daher einen *gleichmäßig besten Test* zum *Signifikanzniveau* $\alpha \in [0,1]$, wenn gilt

- $\sup_{\vartheta \in \Theta_0} G_\varphi(\vartheta) \leq \alpha$;
- $G_\varphi(\vartheta) \geq G_\psi(\vartheta)$ für alle $\vartheta \in \Theta_1$ und alle Tests ψ mit $\sup_{\vartheta \in \Theta_0} G_\psi(\vartheta) \leq \alpha$.

Wir geben zum Abschluss zwei Konstruktionsprinzipien für Tests mit einfachen Hypothesen, d. h. $\Theta_0 = \{\vartheta_0\}$, an. In diesem Fall besteht ein enger Zusammenhang zu Bereichsschätzern: Haben wir einen Bereichsschätzer C für ϑ, so können wir als Test wählen

$$\varphi(x) = \begin{cases} 0 & \text{falls } \vartheta_0 \in C(x), \\ 1 & \text{falls } \vartheta_0 \notin C(x), \end{cases}$$

d. h. wir verwerfen die Nullhypothese, falls ϑ_0 nicht im Konfidenzbereich liegt und akzeptieren sie sonst.

Beispiel 7.1. Wir betrachten wieder die Normalverteilung mit Mittelwert $\vartheta = m \in \mathbb{R}$ und fester Varianz σ^2. Als Nullhypothese wählen wir $\Theta_0 = \{m_0\}$ für ein $m_0 \in \mathbb{R}$ (z. B. $m_0 = 0$). Haben wir wie in Beispiel 6.7 ein Konfidenzintervall zum Konfidenzniveau $1 - \alpha$ gegeben durch

$$C(x) = \left[\hat{M}(x) - t_\alpha \frac{\hat{V}^*(x)}{\sqrt{n}}, \hat{M}(x) + t_\alpha \frac{\hat{V}^*(x)}{\sqrt{n}}\right]$$

für das empirische Mittel $\hat{M}$, dann ist

$$\varphi(x) = \begin{cases} 0 & \text{falls } |\hat{M}(x) - m_0| \leq t_\alpha \frac{\hat{V}^*(x)}{\sqrt{n}}, \\ 1 & \text{falls } |\hat{M}(x) - m_0| > t_\alpha \frac{\hat{V}^*(x)}{\sqrt{n}}, \end{cases}$$

ein Test mit Niveau α: Da $\Theta_0 = \{m_0\}$ einfach und φ die charakteristische Funktion des Verwerfungsbereichs ist, gilt nach Konstruktion von C nämlich

$$G_\varphi(m_0) = \mathbb{E}_{m_0}[\varphi] = \mathbb{P}_{m_0}[m_0 \notin C] = 1 - \mathbb{P}_{m_0}[m_0 \in C] = 1 - (1 - \alpha) = \alpha.$$

Umgekehrt ist für jeden Test mit Signifikanzniveau α der zugehörige Verwerfungsbereich ein Konfidenzbereich mit Konfidenzniveau $1 - \alpha$.

Ist auch die Alternative einfach, d. h. $\Theta_1 = \{\vartheta_1\}$, spricht man von einem *Alternativtest*. Ein klassisches Beispiel ist, zwischen zwei verschiedenen Verteilungen zu entscheiden (z. B. der Nullhypothese „Standardnormalverteilung“ und der Alternative „Gleichverteilung auf $(-10, 10)$“). In diesem Fall können wir wieder die Likelihood-Funktion verwenden, um einen Test zu konstruieren: Gegeben Likelihood-Funktionen L_{ϑ_0} für $\mathbb{P}_{\vartheta_0}$ und L_{ϑ_1} für $\mathbb{P}_{\vartheta_1}$, betrachten wir das *Likelihood-Verhältnis*

$$R(x) := \frac{L_{\vartheta_1}(x)}{L_{\vartheta_0}(x)}, \qquad x \in \Xi.$$

(Ist $L_{\vartheta_0}(x) = 0$, so setzen wir $R(x) = \infty$.) Je größer dieses Verhältnis, desto eher glauben wir, dass die Stichprobe nach $\mathbb{P}_{\vartheta_1}$ verteilt ist. Wir wählen daher als Test für einen gegebenen Schwellwert $c \in \mathbb{R}$

$$\varphi(x) = \begin{cases} 0 & \text{falls } R(x) < c, \\ 1 & \text{falls } R(x) > c. \end{cases}$$

(Ist $L_{\vartheta_0}(x) = 0$, ist $R(x) = \infty$ und wir verwerfen die Nullhypothese sofort.) Dieser Test wird *Neymann–Pearson-Test* genannt. Dabei ist der Schwellwert so zu wählen, dass ein vorgegebenes Signifikanzniveau erreicht wird.

Beispiel 7.2. Wir betrachten wieder das Produktmodell für die Normalverteilung mit unbekanntem Erwartungswert m und bekannter Varianz $\sigma^2 > 0$, wobei wir die Nullhypothese $m = m_0$ gegen die Alternative $m = m_1 > m_0$ testen wollen. Als Likelihood-

Funktionen verwenden wir die entsprechenden Dichtefunktionen und erhalten so

$$\begin{aligned} R(x) = \frac{\rho_1(x)}{\rho_0(x)} &= \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-m_1)^2}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-m_0)^2}} \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n (x_i-m_1)^2 - \sum_{i=1}^n (x_i-m_0)^2\right)\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(2\sum_{i=1}^n x_i(m_0-m_1) + n(m_1^2-m_0^2)\right)\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(2n\hat{M}(x)(m_0-m_1) + n(m_1^2-m_0^2)\right)\right). \end{aligned}$$

Also ist $R(x) > c$ genau dann, wenn gilt

$$\hat{M}(x) > \frac{n(m_0^2-m_1^2) - 2\sigma^2\log(c)}{2n(m_0-m_1)} =: C.$$

Wir suchen für gegebenes Signifikanzniveau $\alpha \in (0,1)$ die Konstante C so, dass gilt (beachte $\Theta_0 = \{m_0\}$)

$$\alpha = G_\varphi(m_0) = \mathbb{E}_{m_0}[\varphi] = \mathbb{P}_{m_0}\left[\hat{M} > C\right] = \mathbb{P}\left[\frac{\hat{M}-m_0}{\sqrt{n}^{-1}\sigma} > \frac{C-m_0}{\sqrt{n}^{-1}\sigma}\right] = 1 - F_{\mathcal{N}}\left(\frac{c-m_0}{\sqrt{n}^{-1}\sigma}\right),$$

wieder durch Transformation der Zufallsvariable $\hat{M} \sim \mathcal{N}(m_0, \sigma^2)$ auf die Standardnormalverteilung $\mathcal{N}(0,1)$ und Verwendung ihrer Verteilungsfunktion $F_{\mathcal{N}}$. Wir verwerfen also die Nullhypothese $m = m_0$ mit Signifikanzniveau α, wenn für das empirische Mittel gilt

$$\hat{M}(x) > C(\alpha) = m_0 + \frac{\sigma}{\sqrt{n}} F_{\mathcal{N}}^{-1}(1-\alpha)$$

(vergleiche den Verwerfungsbereich in Beispiel 7.1!)

Man kann zeigen, dass für jedes Signifikanzniveau ein Neymann–Pearson-Test existiert, und dass dieser ein gleichmäßig bester Test zu diesem Signifikanzniveau ist.

Literatur

M. Brokate & G. Kersting (2019), *Maß und Integral*, 2. Aufl., Math. Kompakt, Cham: Birkhäuser, DOI: 10.1007/978-3-0348-0988-7.

C. R. Charig, D. R. Webb, S. R. Payne & J. E. Wickham (1986), Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *BMJ* 292(6524), 879–882, DOI: 10.1136/bmj.292.6524.879.

A. Eberle (2022), Stochastik, Vorlesungsskript, Institut für Angewandte Mathematik, Universität Bonn.

H.-O. Georgii (2015), *Stochastik, Einführung in die Wahrscheinlichkeitstheorie und Statistik*, 5. Aufl., Berlin: De Gruyter, DOI: 10.1515/9783110359701.

N. Henze (2021), *Stochastik für Einsteiger*, 13. Aufl., Berlin: Springer Spektrum, DOI: 10.1007/978-3-662-63840-8.

M. Hutzenthaler (2015), Stochastik, Vorlesungsskript, Fakultät für Mathematik, Universität Duisburg-Essen.

G. Kersting & A. Wakolbinger (2010), *Elementare Stochastik*, 2. Aufl., Mathematik Kompakt, Basel: Birkhäuser, DOI: 10.1007/978-3-0346-0414-7.

A. Klenke (2020), *Wahrscheinlichkeitstheorie*, 4. Aufl., Masterclass, Berlin: Springer Spektrum, DOI: 10.1007/978-3-662-62089-2.

U. Krengel (2005), *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, 8. Aufl., Vieweg Stud. Wiesbaden: Vieweg, DOI: 10.1007/978-3-663-09885-0.

W. Linde (2014), *Stochastik für das Lehramt*, De Gruyter Studium, Berlin: De Gruyter, DOI: 10.1524/9783110362411.