Stochastik

Andreas Eberle

4. Februar 2022

Inhaltsverzeichnis

Einleitung

„Stochastik“ ist ein Oberbegriff für die Bereiche „Wahrscheinlichkeitstheorie” und „Statistik”. Inhalt dieses Teils der Vorlesung ist eine erste Einführung in grundlegende Strukturen und Aussagen der Stochastik, wobei wir uns zunächst auf Zufallsvariablen mit *diskretem*, d.h. endlichem oder abzählbar unendlichem Wertebereich beschränken. Bevor wir die Grundbegriffe der Wahrscheinlichkeitstheorie einführen, wollen wir kurz darüber nachdenken, wie Methoden der Stochastik bei der mathematischen Modellierung von Anwendungsproblemen eingesetzt werden. Dabei wird sich zeigen, dass stochastische Modelle häufig auch dann sinnvoll eingesetzt werden können, wenn das zu beschreibende Phänomen gar nicht zufällig ist.

Zufall und mathematische Modelle

Beschäftigt man sich mit Grundlagen der Stochastik, dann kommt einem vermutlich die Frage „Was ist Zufall?“ in den Sinn. Diese Frage können und wollen wir hier natürlich nicht beantworten. Wir können aus ihr aber eine andere, viel konkretere Frage ableiten: „Welche Objekte, Phänomene oder Vorgänge können wir sinnvoll unter Verwendung von Methoden der Wahrscheinlichkeitstheorie untersuchen?”. Hier fallen uns auf Anhieb eine ganze Reihe entsprechender „Zufallsvorgänge” ein, die aber gar nicht immer wirklich zufällig sind:

Zufallszahlengenerator. Ein Zufallszahlengenerator ist ein Algorithmus, der eine Folge $u_0, u_1, u_2, \ldots$ von *Pseudozufallszahlen* im Intervall $[0, 1]$ erzeugt. Beispielsweise generiert der von Marsaglia 1972 eingeführte lineare Kongruenzgenerator Binärzahlen zwischen 0 und 1 mit 32 Nachkommastellen auf folgende Weise: Wir setzen $m = 2^{32}$ und wählen einen Startwert (“seed”) $x_0 \in \{0, \ldots, m-1\}$. Dann wird eine Folge $x_0, x_1, x_2, \ldots$ von ganzen Zahlen zwischen 0 und $m-1$ induktiv durch die folgende Rekursion definiert:

$$x_{n+1} = (69069 \cdot x_n + 1) \mod m,$$

und man setzt schließlich $u_n := x_n \cdot 2^{-32}$. Offensichtlich ist sowohl die Folge $(x_n)_{n\in\mathbb{N}}$ von Zahlen zwischen 0 und 2^{32}, als auch die Folge $(u_n)_{n\in\mathbb{N}}$ von Pseudozufallszahlen zwischen 0 und 1 rein deterministisch. Trotzdem verhält sich $(u_n)_{n\in\mathbb{N}}$ in vielerlei Hinsicht wie eine echte Zufallsfolge: Durch eine ganze Reihe statistischer Tests kann man die Folge (u_n) nicht von einer echten Zufallsfolge unterscheiden, und in den meisten Simulationen erhält man bei Verwendung von (u_n) Ergebnisse, die denen für eine echte Zufallsfolge nahezu entsprechen.

Würfelsequenz. Eine Folge von Augenzahlen beim Würfeln ist ein Standardbeispiel einer Zufallsfolge. Tatsächlich ist diese Folge aber auch nicht wirklich zufällig, denn die Endposition des Würfels könnte man im Prinzip aus den Gesetzen der klassischen Mechanik berechnen, wenn man die Bewegung der Hand des Spielers genau beschreiben könnte. Da diese Bewegung zu kompliziert ist, verwendet man ein elementares stochastisches Modell, das in der Regel die Folge der Augenzahlen sehr gut beschreibt.

Bewegung von Gasmolekülen. Lässt man quantenmechanische Effekte außer acht, dann bewegen sich auch die Moleküle in einem Gas bei einer gewissen Temperatur nach einem deterministischen Bewegungsgesetz. Da schon ein Mol mehr als 10^{23} Moleküle enthält, ist eine deterministische Modellierung auf der mikroskopischen Ebene für viele Zwecke zu aufwändig. In der statistischen Physik beschreibt man daher die Zustände der Moleküle durch Zufallsvariablen, und leitet daraus die Gesetze der Thermodynamik her.

In den bisher genannten Beispielen setzt man ein stochastisches Modell an, da eine deterministische Beschreibung zu aufwändig ist. In den meisten praktischen Situationen fehlen uns auch einfach Informationen über das zu beschreibende Objekt:

Unbekanntes Objekt. Wenn wir eine bestimmte Größe, eine Beobachtungssequenz, einen Text oder ein Bild, einen Stammbaum etc. nicht genau kennen, sondern nur indirekte Informationen vorliegen haben (z.B. aus einem verrauschten Signal oder einer DNA-Analyse), dann ist eine stochastische Modellierung des gesuchten Objekts häufig angemessen. Das gewählte Modell oder zumindest die Modellparameter hängen dabei von der uns vorliegenden Information ab !

Aktienkurs. Bei der Modellierung eines Aktienkurses kommen mehrere der bisher genannten Aspekte zusammen: Es gibt sehr viele Einflussfaktoren, den zugrundeliegenden Mechanismus kennen wir nicht (oder nur einen sehr begrenzten Teil davon), und das gewählte stochastische Modell hängt stark von unserem Vorwissen ab.

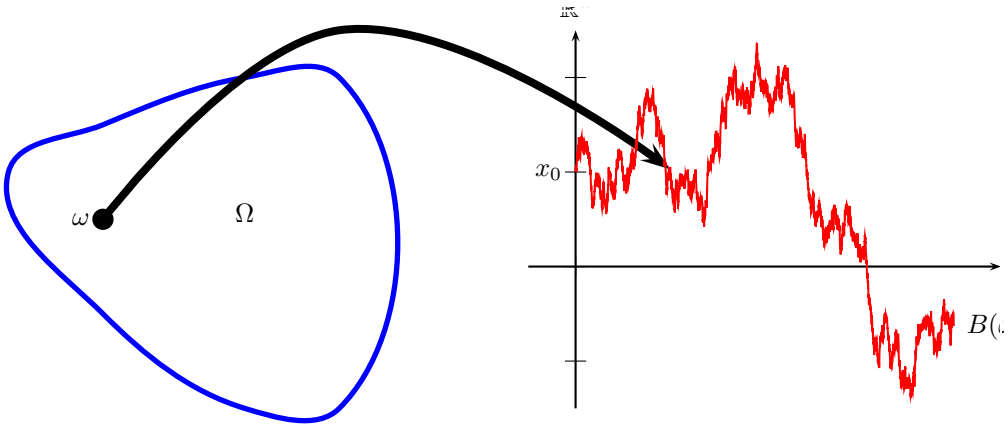


Abbildung 1: $B : \Omega \to C([0,\infty), \mathbb{R}^d), \quad B(\omega) = (B_t(\omega))_{t\geq 0}$.

Beobachtungsvorgang in der Quantenphysik.. In der Quantenmechanik sind die Zustände nicht mehr deterministisch, sondern werden durch eine Wahrscheinlichkeitsdichte beschrieben. Der beobachtete Wert eines Zustands ist daher echt zufällig. Unter `www.randomnumbers.info` kann man eine Liste mit Zufallszahlen herunterladen, die mithilfe von quantenphysikalischen Effekten erzeugt worden sind.

Wie wir sehen, werden stochastische Modelle nicht nur bei „echtem Zufall“ eingesetzt, sondern immer dann, wenn *viele Einflussfaktoren* beteiligt sind oder *unzureichende Informationen* über das zugrunde liegende System vorhanden sind. Für die Modellierung ist es nicht unbedingt nötig zu wissen, ob tatsächlich Zufall im Spiel ist. Ob ein mathematisches Modell ein Anwendungsproblem angemessen beschreibt, kann nur empirisch entschieden werden. Dabei geht man folgendermaßen vor:

- Aus dem Anwendungsproblem gewinnt man durch Abstraktion und Idealisierungen ein stochastisches Modell, das in der Sprache der Wahrscheinlichkeitstheorie formuliert ist.
- Ist das Modell festgelegt, dann können mit den mathematischen Methoden der Wahrscheinlichkeitstheorie Folgerungen aus den Grundannahmen hergeleitet werden.
- Diese Folgerungen liefern dann Vorhersagen für das Anwendungsproblem.

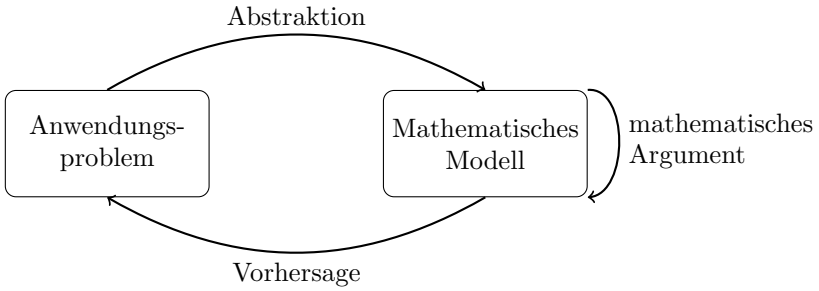


Abbildung 2: Mathematische Modellierung.

- Schließlich überprüft man, ob die Vorhersagen mit den tatsächlichen Beobachtungen übereinstimmen. Falls nicht, versucht man ggf. das Modell zu korrigieren.

In dieser Vorlesung beschränken wir uns meist auf den zweiten Schritt, in einigen einfachen Situationen werden wir aber auch kurz auf den ersten Schritt eingehen. Wichtig ist, dass die Folgerungen im zweiten Schritt *streng logisch* aus den Grundannahmen hergeleitet werden. Das Anwendungsproblem liefert zwar häufig sehr nützliche *Intuition* für mögliche Aussagen oder sogar Beweisverfahren. Der Beweis selbst erfolgt aber inner-mathematisch unter ausschließlicher Verwendung der formal klar spezifizierten Modellannahmen! Die Anwendungsebene und heuristische Argumentationen sollten wir nicht verdrängen, aber es ist wichtig, dass wir klar zwischen Intuition bzw. Heuristik und formalen Beweisen trennen.

Die Idealisierung im mathematischen Modell ermöglicht die Beschreibung einer Vielzahl ganz unterschiedlicher Anwendungssituationen mit ähnlichen mathematischen Methoden und Modellen. Beispielsweise hat sich die Theorie der stochastischen Prozesse in den letzten 100 Jahren ausgehend von Problemen der Physik und der Finanzmathematik sowie innermathematischen Fragestellungen rasant entwickelt. Heute spielen stochastische Prozesse eine zentrale Rolle in diesen Bereichen, aber auch in vielen anderen Gebieten, zum Beispiel in der mathematischen Biologie oder in der Informatik. Das oben beschriebene Schema der stochastischen Modellierung wird manchmal sogar bei rein mathematischen Problemen wie der Verteilung von Primzahlen verwendet.

Wir wollen uns abschließend Aspekte des beschriebenen Modellierungsprozesses noch einmal in einem Beispiel ansehen. In diesem Fall ist das mathematische Modell vorgegeben, und es soll untersucht werden, welcher von mehreren Datensätzen am besten zu dem Modell passt.

Beispiel (0-1 Zufallsfolgen). Wir betrachten fünf Datensätze, die jeweils aus 120 Nullen oder Einsen bestehen:

`tb` 000101100101000110100001101101001000110001000110011100011010011100110010111000010110011000110100111011010110101101010

`pa` 011001000111110000000000000000101010101111111111010101010101111100000000000000100001100010101010010000000000011010010010010

`pb` 111101010100000010101010000000101010100000000010111111010100011001100001110000111100001111000001111100010111101010000010101111010100

`ta` 000001100010110111010011101110011111100100001001011011001101000100100101001111101111100100001000000101100110000010110101

`fa` 101010100010011101010010111001011100001001011001010101110101010100010101101010111000011010100011001101001010001001001000100

Eine dieser 0-1 Folgen wurde mit einem modernen Zufallszahlengenerator erzeugt und ist praktisch nicht von echten Zufallszahlen zu unterscheiden. Die anderen Folgen wurden von verschiedenen Personen

von Hand erzeugt, die gebeten wurden, eine möglichst zufällige 0-1 Folge $x_1, x_2, \ldots, x_{120}$ zu erstellen. Das übliche mathematische Modell für eine solche Zufallsfolge sieht folgendermaßen aus:

> Die Werte $x_1, x_2, \ldots$ sind Realisierungen einer Folge $X_1, X_2, \ldots$ von unabhängigen, auf $\{0, 1\}$ gleichverteilten Zufallsvariablen. (m)

Obwohl Vokabeln wie „Zufallsvariable" oder „unabhängig" der Anschauung entlehnt sind, haben diese Begriffe eine eindeutig spezifizierte mathematische Bedeutung, siehe unten. Daher können wir nun mathematische Folgerungen aus (m) herleiten.

Wenn wir uns die Zahlenfolgen genauer ansehen, stellen wir fest, dass diese sich zum Teil sehr deutlich in den Längen der auftretenden Blöcke von aufeinanderfolgenden Nullen bzw. Einsen unterscheiden. Einen solchen Block nennt man einen **"Run"**. Jede 0-1 Folge lässt sich eindeutig in Runs maximaler Länge zerlegen. Sei R_n die Länge des n-ten Runs in der Zufallsfolge $X_1, X_2, \ldots$. Mit Wahrscheinlichkeit $1/2$ folgt auf eine Null eine Eins bzw. umgekehrt, das heißt der Run endet im nächsten Schritt. Daraus folgt, daß die Länge R_n eines Runs mit Wahrscheinlichkeit $1/2$ gleich 1, mit Wahrscheinlichkeit $1/4 = (1/2)^2$ gleich 2, und allgemein mit Wahrscheinlichkeit 2^{-n} gleich n ist. Zudem kann man beweisen, dass die Zufallsvariablen $R_1, R_2, \ldots$ wieder unabhängig sind. Die durchschnittliche Länge eines Runs ist 2. Daher erwarten wir bei 120 Zeichen ca. 60 Runs, darunter ca. 30 Runs der Länge 1, ca. 30 Runs der Länge ≥ 2, ca. 15 Runs der Länge ≥ 3, ca. 7,5 Runs der Länge ≥ 4, ca. 3,75 Runs der Länge ≥ 5, ca. 1,875 Runs der Länge ≥ 6, und ca. 0,9375 Runs der Länge ≥ 7.

Tatsächlich finden sich in den Datensätzen `tb` und `fa` nur jeweils zwei Runs mit Länge 4 und kein einziger Run mit Länge ≥ 5. Daher würden wir nicht erwarten, dass diese Folgen von einem guten Zufallszahlengenerator erzeugt worden sind, obwohl prinzipiell ein solcher Ausgang natürlich möglich ist. In der Tat kann man beweisen, dass im Modell (m) die Wahrscheinlichkeit dafür, dass es keinen Run der Länge ≥ 5 gibt, sehr klein ist. Umgekehrt finden sich im Datensatz `pa` Runs mit Längen 13 und 15. Erneut ist die Wahrscheinlichkeit dafür äußerst gering, wenn wir das Modell (m) annehmen.

Zusammenfassend ist (m) kein geeignetes mathematisches Modell zur Beschreibung der Datensätze `tb`,`fa` und `pa`. Für die Datensätze `pb` und insbesondere `ta` liegen die Anzahlen der Runs verschiedener Länge näher bei den im Mittel erwarteten Werten, sodass (m) ein geeignetes Modell zur Beschreibung dieser Folgen sein könnte. Möglicherweise zeigen aber auch noch weitergehende Tests, dass das Modell doch nicht geeignet ist. Tatsächlich stammt nur die Folge ta von einem Zufallszahlengenerator, und die anderen Folgen wurden von Hand erzeugt.

Abschließend sei noch bemerkt, dass die Unbrauchbarkeit des Modells (m) für die Folgen `tb`, `fa` und `pa` eine stochastische Modellierung natürlich nicht ausschließt. Zum Beispiel könnte man versuchen die Datensätze `tb` und `fa` durch eine Folge von Zufallsvariablen mit negativen Korrelationen, und den Datensatz `pa` durch eine Folge von Zufallsvariablen mit positiven Korrelationen zu beschreiben.

1 Diskrete Zufallsvariablen

Grundlegende Objekte im axiomatischen Aufbau der Wahrscheinlichkeitstheorie nach Kolmogorov sind die Menge Ω der in einem Modell in Betracht gezogenen *Fälle* ω, die Kollektion $\mathcal{A}$ der betrachteten *Ereignisse* A, sowie die *Wahrscheinlichkeitsverteilung* P, die jedem Ereignis A eine Wahrscheinlichkeit $P[A]$ zwischen 0 und 1 zuordnet. Dabei sind Ereignisse Teilmengen von Ω, und eine Wahrscheinlichkeitsverteilung ist eine Abbildung von $\mathcal{A}$ nach $[0,1]$. Zudem sind *Zufallsvariablen* X von zentralem Interesse, die jedem Fall ω einen Wert $X(\omega)$ zuweisen. Zur Illustration betrachten wir drei elementare Beispiele bevor wir die genannten Objekte formal definieren.

Beispiel (Würfeln und Münzwürfe).

a) Einmal Würfeln:
Die Menge der möglichen *Fälle* ist $\Omega = \{1,2,3,4,5,6\}$. Die Elemente $\omega \in \Omega$ bezeichnet man auch als *Elementarereignisse* und identifiziert sie mit den einelementigen Mengen $\{\omega\}$. Allgemeine *Ereignisse* werden durch Teilmengen von Ω beschrieben, zum Beispiel:

„Augenzahl ist 3“	$\{3\}$
„Augenzahl ist gerade“	$\{2,4,6\}$
„Augenzahl ist *nicht* gerade“	$\{1,3,5\} = \{2,4,6\}^C$
„Augenzahl ist größer als 3“	$\{4,5,6\}$
„Augenzahl ist gerade *und* größer als 3“	$\{4,6\} = \{2,4,6\} \cap \{4,5,6\}$
„Augenzahl gerade *oder* größer als 3“	$\{2,4,5,6\} = \{2,4,6\} \cup \{4,5,6\}$

Hierbei schreiben wir A^C für das Komplement $\Omega \setminus A$ der Menge A in der vorgegebenen Grundmenge Ω. für die Wahrscheinlichkeiten sollte im Falle eines „fairen“ Würfels gelten:

$$P[\text{„3“}] = \frac{1}{6},$$
$$P[\text{„Augenzahl gerade“}] = \frac{\text{Anzahl günstige Fälle}}{\text{Anzahl mögliche Fälle}} = \frac{|\{2,4,6\}|}{|\{1,2,3,4,5,6\}|} = \frac{3}{6} = \frac{1}{2},$$
$$P[\text{„Augenzahl gerade oder größer als 3“}] = \frac{4}{6} = \frac{2}{3}.$$

Beispiele für *Zufallsvariablen* sind

$$X(\omega) = \omega, \qquad \text{„Augenzahl des Wurfs“, oder}$$
$$G(\omega) = \begin{cases} 1 & \text{falls } \omega \in \{1,2,3,4,5\}, \\ -5 & \text{falls } \omega = 6, \end{cases} \qquad \text{„Gewinn bei einem fairen Spiel“.}$$

In einem anderen (detaillierteren) Modell hätte man die Menge Ω auch anders wählen können, z.B. könnte Ω alle möglichen stabilen Anordnungen des Würfels auf dem Tisch beinhalten. Wir werden später sehen, dass die konkrete Wahl der Menge Ω oft gar nicht wesentlich ist - wichtig sind vielmehr die Wahrscheinlichkeiten, mit denen die relevanten Zufallsvariablen Werte in bestimmten Bereichen annehmen.

b) Endlich viele faire Münzwürfe:
Es ist naheliegend, als Menge der möglichen Fälle

$$\Omega \;=\; \{\omega = (x_1, \ldots, x_n) \mid x_i \in \{0,1\}\} \;=\; \{0,1\}^n$$

zu betrachten, wobei n die Anzahl der Münzwürfe ist, und 0 für „Kopf“ sowie 1 für „Zahl“ steht. Alle Ausgänge sind genau dann gleich wahrscheinlich, wenn $P[\{\omega\}] = 2^{-n}$ für alle $\omega \in \Omega$ gilt. Dies wird im folgenden angenommen. Zufallsvariablen von Interesse sind beispielsweise das Ergebnis des i-ten Wurfs

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

oder die Häufigkeit

$$S_n(\omega) \;=\; \sum_{i=1}^{n} X_i(\omega)$$

von Zahl bei n Münzwürfen. Das Ereignis „i-ter Wurf ist Kopf“ wird durch die Menge

$$A_i \;=\; \{\omega \in \Omega \mid X_i(\omega) = 0\} \;=\; X_i^{-1}(0)$$

beschrieben. Diese Menge bezeichnen wir in intuitiver Kurznotation auch mit $\{X_i = 0\}$. Es gilt

$$P[X_i = 0] \;:=\; P[\{X_i = 0\}] \;=\; P[A_i] \;=\; \frac{1}{2}.$$

Das Ereignis „genau k-mal Zahl“ wird entsprechend durch die Menge

$$A \;= \{\omega \in \Omega \mid S_n(\omega) = k\} \;=\; \{S_n = k\}$$

beschrieben und hat die Wahrscheinlichkeit

$$P[S_n = k] \;=\; \binom{n}{k} 2^{-n}.$$

c) Unendlich viele Münzwürfe:
Hier kann man als Menge der möglichen Fälle den Raum

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

aller binären Folgen ansetzen. Diese Menge ist überabzählbar, da die durch die Dualdarstellung reeller Zahlen definierte Abbildung

$$(x_1, x_2, \ldots) \;\mapsto\; \sum_{i=1}^{\infty} x_i \cdot 2^{-i}$$

von Ω nach $[0,1]$ surjektiv ist. Dies hat zur Folge, dass es nicht möglich ist, *jeder* Teilmenge von Ω in konsistenter Weise eine Wahrscheinlichkeit zuzuordnen. Die formale Definition von Ereignissen und Wahrscheinlichkeiten ist daher in diesem Fall aufwändiger, und wird erst in der Vorlesung Einführung in die Wahrscheinlichkeitstheorie systematisch behandelt.

1.1 Ereignisse und ihre Wahrscheinlichkeit

Wir werden nun die Kolmogorovsche Definition eines Wahrscheinlichkeitsraums motivieren und formulieren, erste einfache Folgerungen daraus ableiten, und elementare Beispiele betrachten. Ein Wahrscheinlichkeitsraum besteht aus einer nichtleeren Menge Ω, die bis auf weiteres fest gewählt sei, einer Kollektion $\mathcal{A}$ von Teilmengen von Ω (den Ereignissen) und einer Abbildung $P : \Omega \to [0,1]$, die bestimmte Axiome erfüllen.

Ereignisse als Mengen

Seien A, B, und A_i, $i \in I$, Ereignisse, d.h. Teilmengen von Ω. Hierbei ist I eine beliebige Indexmenge. Anschaulich stellen wir uns vor, dass ein Element $\omega \in \Omega$ zufällig ausgewählt wird, und das Ereignis A eintritt, falls ω in A enthalten ist. „Zufällig“ bedeutet dabei nicht unbedingt, dass alle Fälle gleich wahrscheinlich sind ! Wir werden manchmal auch die folgenden Notationen für die Menge A verwenden:

$$A = \{\omega \in \Omega \mid \omega \in A\} = \{\omega \in A\} = \{\text{ „}A\text{ tritt ein“ }\}.$$

Da Ereignisse durch Mengen beschrieben werden, können wir mengentheoretische Operationen benutzen, um mehrere Ereignisse zu kombinieren. Wir wollen uns überlegen, was Ereignisse wie A^C, $A \cup B$, $\bigcap_{i \in I} A_i$

usw. anschaulich bedeuten. Um dies herauszufinden, betrachtet man einen möglichen Fall ω und untersucht, wann dieser eintritt. Beispielsweise gilt

$$\omega \in A \cup B \qquad \Leftrightarrow \qquad \omega \in A \text{ oder } \omega \in B,$$

also in anschaulicher Sprechweise:

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

Entsprechend gilt

$$\omega \in \bigcup_{i \in I} A_i \qquad \Leftrightarrow \qquad \text{es gibt ein } i \in I \text{ mit } \omega \in A_i,$$

also

$$\text{„}\textstyle\bigcup_{i\in I} A_i \text{ tritt ein“} \qquad \Leftrightarrow \qquad \text{„mindestens eines der Ereignisse } A_i \text{ tritt ein“}.$$

Auf analoge Weise überlegen wir uns die Bedeutungen der folgenden Mengenoperationen:

$A \cap B$	„A und B treten ein“,
$\bigcap_{i\in I} A_i$	„jedes der A_i tritt ein“,
$A^C = \Omega \setminus A$	„A tritt nicht ein“,
$A = \emptyset$	„unmögliches Ereignis“ (tritt nie ein),
$A = \Omega$	„sicheres Ereignis“ (tritt immer ein),
$A = \{\omega\}$	„Elementarereignis“ (tritt nur im Fall ω ein).

Die Kollektion $\mathcal{A}$ aller im Modell zugelassenen bzw. in Betracht gezogenen Ereignisse besteht aus Teilmengen von Ω, d.h $\mathcal{A}$ ist eine Teilmenge der *Potenzmenge*

$$\mathcal{P}(\Omega) \;=\; \{A \mid A \subseteq \Omega\}$$

Die Kollektion $\mathcal{A}$ sollte unter den oben betrachteten Mengenoperationen (Vereinigungen, Durchschnitte, Komplementbildung) abgeschlossen sein. Genauer fordern wir die Abgeschlossenheit nur unter abzählbaren Vereinigungen und Durchschnitten, da $\mathcal{A}$ andernfalls immer gleich der Potenzmenge sein müsste sobald alle einelementigen Mengen enthalten sind. Eine effiziente Formulierung der Abgeschlossenheit unter abzählbaren Mengenoperationen führt auf die folgende Definition:

Definition 1.1. Eine Kollektion $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ von Teilmengen von Ω heißt **σ-Algebra**, falls gilt:

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

(ii) für alle $A \in \mathcal{A}$ gilt: $A^C \in \mathcal{A}$,

(iii) für $A_1, A_2, \ldots \in \mathcal{A}$ gilt: $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Bemerkung. Aus der Definition folgt bereits, dass eine σ-Algebra $\mathcal{A}$ unter allen oben betrachteten endlichen und abzählbar unendlichen Mengenoperationen abgeschlossen ist, denn:

(a) Nach (i) und (ii) ist $\emptyset = \Omega^C \in \mathcal{A}$.

(b) Sind $A_1, A_2, \ldots \in \mathcal{A}$, dann folgt nach (ii) und (iii): $\bigcap_{i=1}^{\infty} A_i = (\bigcup_{i=1}^{\infty} A_i^C)^C \in \mathcal{A}$.

(c) Sind $A, B \in \mathcal{A}$, dann folgt nach (iii) und (a): $A \cup B = A \cup B \cup \emptyset \cup \emptyset \cup \ldots \in \mathcal{A}$.

(d) Entsprechend folgt $A \cap B \in \mathcal{A}$ aus (b) und (i).

Beispiele. a) Potenzmenge.
Die Potenzmenge $\mathcal{A} = \mathcal{P}(\Omega)$ ist stets eine σ-Algebra. In diskreten Modellen, in denen Ω abzählbar ist, werden wir diese σ-Algebra häufig verwenden. Bei nichtdiskreten Modellen kann man dagegen *nicht* jede Wahrscheinlichkeitsverteilung P auf einer σ-Algebra $\mathcal{A} \subset \mathcal{P}(\Omega)$ zu einer Wahrscheinlichkeitsverteilung auf $\mathcal{P}(\Omega)$ erweitern, siehe Beispiel c).

b) Partielle Information.
Wir betrachten das Modell für n Münzwürfe mit

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

Sei $k \leq n$. Dann ist die Kollektion $\mathcal{F}_k$ aller Mengen $A \subseteq \Omega$, die sich in der Form

$$A \;=\; \{(x_1, \ldots, x_n) \in \Omega \mid (x_1, \ldots, x_k) \in B\} \;=\; B \times \{0,1\}^{n-k}$$

mit $B \subseteq \{0,1\}^k$ darstellen lassen, eine σ-Algebra. Die Ereignisse in der σ-Algebra $\mathcal{F}_k$ sind genau diejenigen, von denen wir schon wissen ob sie eintreten oder nicht, wenn wir nur den Ausgang der ersten k Münzwürfe kennen. Die σ-Algebra $\mathcal{F}_k$ beschreibt also die *Information aus den ersten k Münzwürfen*.

c) Borelsche σ-Algebra. Man kann zeigen, dass es auf der Potenzmenge des reellen Intervalls $\Omega = [0,1]$ keine Wahrscheinlichkeitsverteilung P gibt, die jedem Teilintervall (a,b) die Länge als Wahrscheinlichkeit zuordnet. Andererseits gibt es eine kleinste σ-Algebra $\mathcal{B}$, die alle Teilintervalle enthält. Auf der σ-Algebra $\mathcal{B}$ existiert eine *kontinuierliche Gleichverteilung* mit der gerade beschriebenen Eigenschaft, siehe Analysis III. Sie enthält zwar alle offenen und alle abgeschlossenen Teilmengen von $[0,1]$, ist aber echt kleiner als die Potenzmenge $\mathcal{P}([0,1])$.

Wahrscheinlichkeitsverteilungen

Sei Ω eine nichtleere Menge und $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ eine σ-Algebra. Wir wollen nun die Abbildung P einführen, die jedem Ereigniss $A \in \mathcal{A}$ eine Wahrscheinlichkeit $P[A]$ zuordnet. Welche Bedingungen (Axiome) sollten wir von P fordern ? Sind $A, B \in \mathcal{A}$ Ereignisse, dann ist $A \cup B$ ein Ereignis, welches genau dann eintritt, wenn A eintritt oder B eintritt. Angenommen, die beiden Ereignisse A und B *treten nicht gleichzeitig ein*, d.h. die Mengen A und B sind *disjunkt*. Dann sollte die Wahrscheinlichkeit von $A \cup B$ die Summe der Wahrscheinlichkeiten von A und B sein:

$$A \cap B = \emptyset \qquad \Rightarrow \qquad P[A \cup B] \;=\; P[A] + P[B],$$

d.h. die Abbildung P ist *additiv*. Wir fordern etwas mehr, nämlich dass eine entsprechende Eigenschaft sogar für *abzählbar* unendliche Vereinigungen von disjunkten Mengen gilt. Dies wird sich als wichtig erweisen, um zu einer leistungsfähigen Theorie zu gelangen, die zum Beispiel Konvergenzaussagen für Folgen von Zufallsvariablen liefert.

Definition 1.2 (Axiome von Kolmogorov). Eine Abbildung $P : \mathcal{A} \to [0,\infty]$, $A \mapsto P[A]$, heißt **Wahrscheinlichkeitsverteilung** auf $(\Omega, \mathcal{A})$, falls gilt:

(i) P ist *„normiert“*, d.h.

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

(ii) P ist *„σ-additiv“*, d.h. für Ereignisse $A_1, A_2, \ldots \in \mathcal{A}$ mit $A_i \cap A_j = \emptyset$ für $i \neq j$ gilt:

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

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

Bemerkung (Maße). Gilt nur Eigenschaft (ii) und $P[\emptyset] = 0$, dann heißt P ein *Maß*. Eine Wahrscheinlichkeitsverteilung ist ein normiertes Maß, und wird daher auch äquivalent als **Wahrscheinlichkeitsmaß** bezeichnet. Maße spielen auch in der Analysis eine große Rolle, und werden in der Vorlesung Analysis III systematisch behandelt.

Man kann sich fragen, weshalb wir die Additivität nicht für beliebige Vereinigungen fordern. Würden wir dies tun, dann gäbe es nicht viele interessante Wahrscheinlichkeitsverteilungen auf kontinuierlichen Räumen. Beispielsweise sollte unter der Gleichverteilung auf dem Intervall $[0, 1]$ jede Menge, die nur aus einem Punkt besteht, die Wahrscheinlichkeit 0 haben, da sie in beliebig kleinen Intervallen enthalten ist. Würde Additivität für beliebige Vereinigungen gelten, dann müsste auch das ganze Intervall $[0, 1]$ Wahrscheinlichkeit 0 haben, da es die Vereinigung seiner einelementigen Teilmengen ist. Die Forderung der σ-Additivität liefert also einen angemessenen Kompromiss, der genügend viele interessante Modelle zulässt und es gleichzeitig ermöglicht, sehr weitreichende Aussagen herzuleiten.

Der folgende Satz zeigt, dass Wahrscheinlichkeitsverteilungen einige elementare Eigenschaften besitzen, die wir von der Anschauung her erwarten würden:

Satz 1.3 (Elementare Eigenschaften und erste Rechenregeln).
Ist $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, dann gelten die folgenden Aussagen:

(i) $P[\emptyset] = 0$,

(ii) für $A, B \in \mathcal{A}$ mit $A \cap B = \emptyset$ gilt

$$P[A \cup B] = P[A] + P[B] \qquad \text{„endliche Additivität“}.$$

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

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

Insbesondere folgt

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

(iv) für beliebige Ereignisse $A, B \in \mathcal{A}$ gilt

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

Beweis. (i) Wegen der σ-Additivität von P gilt

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

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

(ii) für disjunkte Ereignisse $A, B \in \mathcal{A}$ folgt aus der σ-Additivität und mit (i)

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

(iii) Gilt $A \subseteq B$, dann ist $B = A \cup (B \backslash A)$. Da diese Vereinigung disjunkt ist, folgt mit (ii)

$$P[B] \;=\; P[A] + P[B \backslash A] \;\geq\; P[A].$$

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

(iv) für beliebige Ereignisse $A, B \in \mathcal{A}$ gilt nach (iii) gilt:

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

■

Aussage (iv) des Satzes lässt sich für endlich viele Ereignisse verallgemeinern. Beispielsweise folgt durch mehrfache Anwendung von (iv) für die Vereinigung von drei Ereignissen

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

Mit vollständiger Induktion ergibt sich eine Formel für die Wahrscheinlichkeit der Vereinigung einer beliebigen endlichen Anzahl von Ereignissen:

Korollar 1.4 (Einschluss-/Ausschlussprinzip). Für $n \in \mathbb{N}$ und Ereignisse $A_1, \ldots, A_n \in \mathcal{A}$ gilt:

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

Das Einschluss-/Ausschlussprinzip werden wir auf eine elegantere Weise am Ende dieses Kapitels beweisen (siehe Satz 1.18).

Diskrete Wahrscheinlichkeitsverteilungen

Ein ganz einfaches Beispiel für eine diskrete Wahrscheinlichkeitsverteilung ist das Grundmodell für einen Münzwurf oder ein allgemeineres 0-1-Experiment mit Erfolgswahrscheinlichkeit $p \in [0,1]$. Hier ist $\Omega = \{0,1\}$, $\mathcal{A} = \mathcal{P}(\Omega) = \{\emptyset, \{0\}, \{1\}, \Omega\}$, und P ist gegeben durch

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

Die Verteilung P nennt man auch eine *(einstufige) Bernoulliverteilung* mit Parameter p.

Auf analoge Weise erhalten wir Wahrscheinlichkeitsverteilungen auf endlichen oder abzählbar unendlichen Mengen Ω. In diesem Fall können wir die Potenzmenge $\mathcal{P}[\Omega]$ als σ-Algebra verwenden, und Wahrscheinlichkeiten von beliebigen Ereignissen aus den Wahrscheinlichkeiten der Elementarereignisse berechnen.

Satz 1.5. (i) Sei $0 \leq p(\omega) \leq 1$, $\sum_{\omega \in \Omega} p(\omega) = 1$, eine Gewichtung der möglichen Fälle. Dann ist durch

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

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

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

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

Definition 1.6. Die Funktion $p : \Omega \to [0,1]$ heißt **Massenfunktion** („probability mass function“) der diskreten Wahrscheinlichkeitsverteilung P.

Für den Beweis des Satzes brauchen wir einige Vorbereitungen. Wir bemerken zunächst, dass für eine abzählbare Menge A die Summe der Gewichte $p(\omega)$ über $\omega \in A$ definiert ist durch

$$\sum_{\omega \in A} p(\omega) \;:=\; \sum_{i=1}^{\infty} p(\omega_i), \tag{1.1}$$

wobei $\omega_1, \omega_2, \ldots$ eine beliebige Abzählung von A ist. Da die Gewichte nichtnegativ sind, existiert die Summe auf der rechten Seite (wobei der Wert $+\infty$ zugelassen ist). Der erste Teil des folgenden Lemmas zeigt, dass die Summe über $\omega \in A$ durch (1.1) wohldefiniert ist:

Lemma 1.7. *(i) Unabhängig von der gewählten Abzählung gilt*

$$\sum_{\omega \in A} p(\omega) \;=\; \sup_{\substack{F \subseteq A \\ |F| < \infty}} \sum_{\omega \in F} p(\omega). \tag{1.2}$$

Insbesondere hängt die Summe monoton *von A ab, d.h. für $A \subseteq B$ gilt*

$$\sum_{\omega \in A} p(\omega) \le \sum_{\omega \in B} p(\omega). \tag{1.3}$$

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

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

Beweis. (i) Sei $\omega_1, \omega_2, \ldots$ eine beliebige Abzählung von A. Aus $p(\omega_i) \ge 0$ für alle $i \in \mathbb{N}$ folgt, dass die Folge der Partialsummen $\sum_{i=1}^{n} p(\omega_i)$ monoton wachsend ist. Somit gilt

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

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

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

Wir zeigen zunächst „$\le$“, und Anschließend „$\ge$“:

„≤“: für alle $n \in \mathbb{N}$ gilt:

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

da das Supremum auch über $F = \{\omega_1, \ldots, \omega_n\}$ gebildet wird. Damit folgt „≤“.

„≥“: Ist F eine endliche Teilmenge von A, dann gibt es ein $n \in \mathbb{N}$, so dass $F \subseteq \{\omega_1, \ldots, \omega_n\}$. Daher gilt

$$\sum_{\omega \in F} p(\omega) \;\leq\; \sum_{i=1}^{n} p(\omega_i) \;\leq\; \sup_{n \in \mathbb{N}} \sum_{i=1}^{n} p(\omega_i),$$

und es folgt „≥“.

(ii) Falls A endlich ist, dann gilt $A_i \neq \emptyset$ nur für endlich viele $i \in \mathbb{N}$ und alle A_i sind endlich. Die Behauptung folgt dann aus dem Kommutativ- und dem Assoziativgesetz. Wir nehmen nun an, dass A abzählbar unendlich ist. In diesem Fall können wir die Aussage aus der Aussage für endliche A unter Verwendung von (i) herleiten. Wir zeigen erneut „≤“ und „≥“ separat:

„≤“: Ist F eine endliche Teilmenge von A, so ist $F = \bigcup_{i=1}^{\infty}(F \cap A_i)$. Da diese Vereinigung wieder disjunkt ist, folgt mit σ-Additivität und Gleichung (1.3):

$$\sum_{\omega \in F} p(\omega) \;=\; \sum_{i=1}^{\infty} \sum_{\omega \in F \cap A_i} p(\omega) \;\leq\; \sum_{i=1}^{\infty} \sum_{\omega \in A_i} p(\omega).$$

Also folgt nach (i) auch:

$$\sum_{\omega \in A} p(\omega) \;=\; \sup_{\substack{F \subseteq A \\ |F|<\infty}} \sum_{\omega \in F} p(\omega) \;\leq\; \sum_{i=1}^{\infty} \sum_{\omega \in A_i} p(\omega).$$

„≥“: Seien $F_i \subseteq A_i$ endlich. Da die F_i wieder disjunkt sind, folgt mit σ-Additivität und Gleichung (1.3) für alle $n \in \mathbb{N}$:

$$\sum_{i=1}^{n} \sum_{\omega \in F_i} p(\omega) \;=\; \sum_{\omega \in \bigcup_{i=1}^{n} F_i} p(\omega) \;\leq\; \sum_{\omega \in A} p(\omega).$$

Nach (i) folgt dann auch

$$\sum_{i=1}^{n} \sum_{\omega \in A_i} p(\omega) \;\leq\; \sum_{\omega \in A} p(\omega),$$

und damit die Behauptung für $n \to \infty$. ■

Beweis (Beweis von Satz 1.5). (i) Nach Voraussetzung gilt $P[A] \geq 0$ für alle $A \subseteq \Omega$ und $P[\Omega] = \sum_{\omega \in \Omega} p(\omega) = 1$. Seien nun A_i ($i \in \mathbb{N}$) disjunkt. Dann folgt aus Lemma 1.7.(ii):

$$P\Big[\bigcup_{i=1}^{\infty} A_i\Big] \;=\; \sum_{\omega \in \bigcup A_i} p(\omega) \;=\; \sum_{i=1}^{\infty} \sum_{\omega \in A_i} p(\omega) \;=\; \sum_{i=1}^{\infty} P[A_i],$$

also die σ-Additivität von P.

(ii) Umgekehrt folgt aus der σ-Additivität von P für $A \subseteq \Omega$ sofort

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

Gleichverteilungen (Laplace-Modelle)

Ist Ω endlich, dann existiert auf $\mathcal{A} = \mathcal{P}(\Omega)$ eine eindeutige Wahrscheinlichkeitsverteilung P mit konstanter Massenfunktion

$$p(\omega) \;=\; \frac{1}{|\Omega|} \qquad \text{für alle } \omega \in \Omega.$$

Als Wahrscheinlichkeit eines Ereignisses $A \subseteq \Omega$ ergibt sich

$$P[A] \;=\; \sum_{\omega \in A} \frac{1}{|\Omega|} \;=\; \frac{|A|}{|\Omega|} \;=\; \frac{\text{Anzahl „günstiger“ Fälle}}{\text{Anzahl aller Fälle}}. \tag{1.4}$$

Die Verteilung P heißt *Gleichverteilung* auf Ω und wird auch mit $\mathrm{Unif}(\Omega)$ bezeichnet. Laplace (1814) benutzte (1.4) als Definition von Wahrscheinlichkeiten. Dabei ist zu beachten, dass die Gleichverteilung nicht erhalten bleibt, wenn man zum Beispiel mehrere Fälle zu einem zusammenfasst. Der Laplacesche Ansatz setzt also voraus, dass man eine Zerlegung in gleich wahrscheinliche Fälle finden kann.

Beispiele. a) n FAIRE MÜNZWÜRFE:
Die Gleichverteilung $\mathrm{Unif}(\Omega)$ auf $\Omega = \{0,1\}^n$ hat die Massenfunktion

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

Die gleich wahrscheinlichen Fälle sind hier die 2^n möglichen Münzwurfsequenzen.

b) ZUFÄLLIGE PERMUTATIONEN:
Sei $\Omega = \mathcal{S}_n$ die Menge aller Bijektionen $\omega : \{1,2,\dots,n\} \to \{1,2,\dots,n\}$. Der 1 können n verschiedene Zahlen zugeordnet geordnet werden, der 2 die verbleibenden $n-1$, usw. Somit gibt es insgesamt $n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 1$ dieser Permutationen. Bezüglich der Gleichverteilung auf $\mathcal{S}_n$ gilt also

$$P[A] = \frac{|A|}{n!} \qquad \text{für alle } A \subseteq \mathcal{S}_n.$$

Anschauliche Beispiele für zufällige Permutationen sind die Anordnung eines gemischten Kartenspiels, oder das zufällige Vertauschen von n Hüten oder Schlüsseln. In letzterem Beispiel gilt:

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

Wie groß ist die Wahrscheinlichkeit, dass einer der Schlüssel sofort passt? Das Ereignis „Schlüssel i passt“ wird beschrieben durch die Menge

$$A_i \;=\; \{\omega \mid \omega(i) = i\} \;=\; \{\text{„}i \text{ ist Fixpunkt von } \omega\text{“ }\}.$$

Die Wahrscheinlichkeit für das Ereignis „ein Schlüssel passt“ lässt sich dann nach dem Einschluss-/Ausschlussprinzip (Satz 1.18) berechnen:

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

Hierbei haben wir benutzt, dass es $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ k-elementige Teilmengen $\{i_1,\dots,i_k\}$ von $\{1,\dots,n\}$ gibt. für das Gegenereignis erhalten wir:

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

Die letzte Summe konvergiert für $n \to \infty$ gegen e^{-1}. Der Grenzwert existiert also und ist weder 0 noch 1! Somit hängt die Wahrscheinlichkeit, dass keiner der Schlüssel passt, für große n nur wenig von n ab.

Empirische Verteilungen

Sei $x_1, x_2, \ldots \in \Omega$ eine Liste von Beobachtungsdaten oder Merkmalsausprägungen, zum Beispiel das Alter aller Einwohner von Bonn. Für $k \in \mathbb{N}$ ist

$$N_k[A] := |\{i \in \{1,\ldots,k\} \mid x_i \in A\}| \qquad \text{die Häufigkeit der Werte in } A \text{ unter } x_1,\ldots,x_k, \qquad \text{und}$$
$$P_k[A] := N_k[A]/k, \qquad \text{die entsprechende relative Häufigkeit von Werten in } A.$$

Für jedes feste k ist P_k eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{P}(\Omega))$, deren Massenfunktion

$$p_k(\omega) = \frac{N_k[\{\omega\}]}{k}$$

durch die relativen Häufigkeit der möglichen Merkmalsausprägungen unter $x_1,\ldots,x_k$ gegeben ist. Die Wahrscheinlichkeitsverteilung P_k heißt *empirische Verteilung* der Werte $x_1,\ldots,x_k$. In der beschreibenden Statistik analysiert man empirische Verteilungen mithilfe verschiedener Kenngrößen.

Beispiele. a) Abzählung aller möglichen Fälle:
Sei $x_1,\ldots,x_k$ eine Abzählung der Elemente in Ω. Dann stimmt die empirische Verteilung P_k mit der Gleichverteilung auf Ω überein.

b) Empirische Verteilung von n Zufallszahlen aus $\{1,2,3,4,5,6\}$:
Das **empirische Gesetz der großen Zahlen** besagt, dass sich die empirischen Verteilungen für

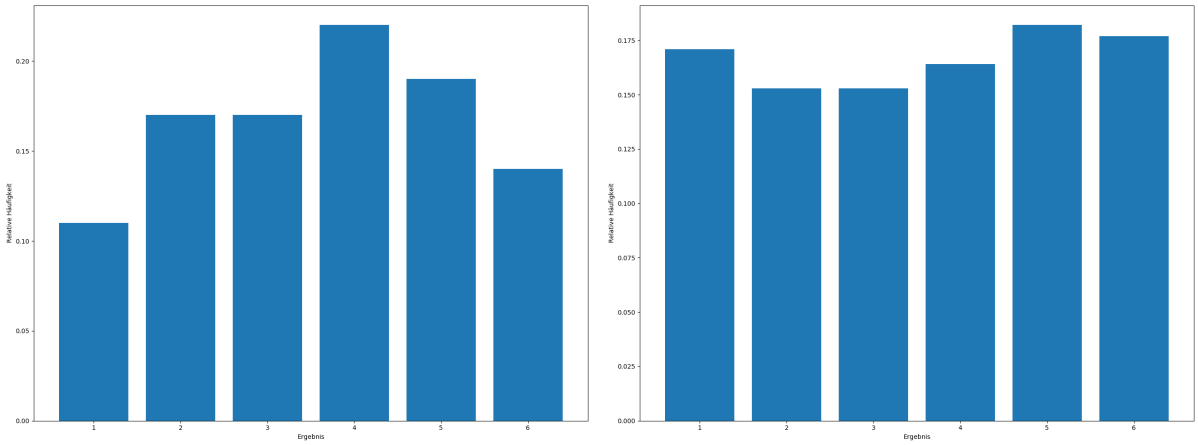


Abbildung 1.1: Empirische Verteilung von 100 bzw. 1000 Würfen eines fairen Würfels.

$k \to \infty$ der zugrundeliegenden Wahrscheinlichkeitsverteilung P (hier der Gleichverteilung auf $\{1,2,\ldots,6\}$) annähern:

$$P_k[A] = \frac{|\{i \in \{1,\ldots,k\} \mid x_i \in A\}|}{k} \to P[A] \qquad \text{für } k \to \infty.$$

Diese Aussage wird auch als frequentistische „Definition“ der Wahrscheinlichkeit von A in den empirischen Wissenschaften verwendet. Wir werden die Konvergenz der empirischen Verteilungen von unabhängigen, identisch verteilten Zufallsvariablen unten aus den Kolmogorovschen Axiomen herleiten.

c) Empirische Verteilung der Buchstaben „a“ bis „z“ in dem Wort „Eisenbahnschranken-waerterhaeuschen“ und in einem englischen Wörterbuch:

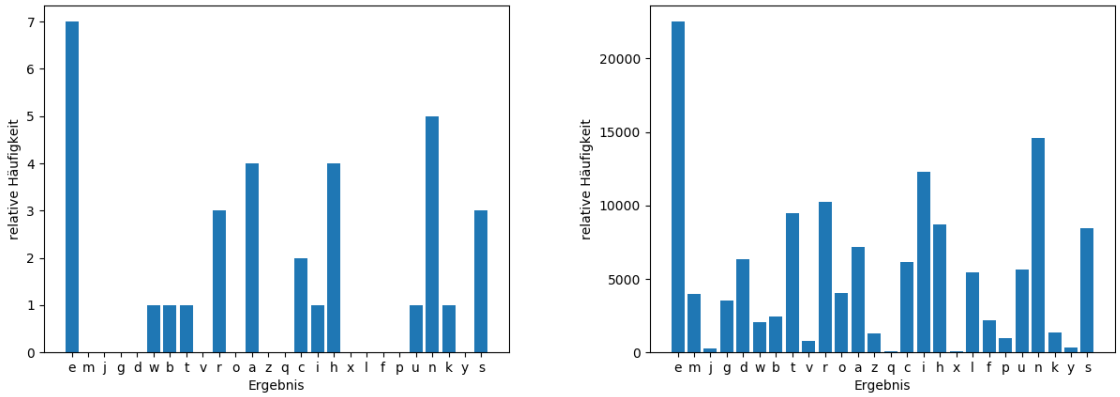
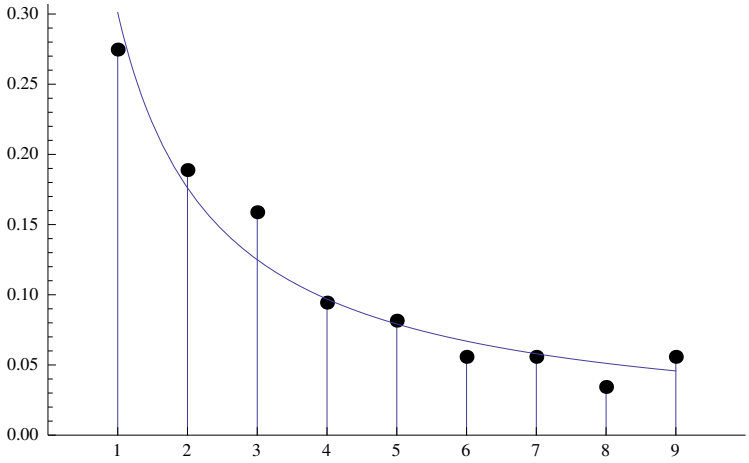


Abbildung 1.2: Empirische Verteilung der Buchstaben in dem Wort „Eisenbahnschrankenwaerterhaeuschen“ bzw. in Faust 1.

d) Benfordsches Gesetz:
Das Benfordsche Gesetz beschreibt eine Gesetzmäßigkeit in der Verteilung der Anfangsziffern von Zahlen in empirischen Datensätzen. Es lässt sich etwa in Datensätzen über Einwohnerzahlen von Städten, Geldbeträge in der Buchhaltung, Naturkonstanten etc. beobachten. Ist d die erste Ziffer einer Dezimalzahl, so tritt sie nach dem Benfordschen Gesetz in empirischen Datensätzen näherungsweise mit folgenden relativen Häufigkeiten $p(d)$ auf:

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

In der Grafik unten (`Quelle`: „Wolfram Demonstrations Project“) werden die relativen Häufigkeiten der Anfangsziffern 1 bis 9 in den Anzahlen der Telefonanschlüsse in allen Ländern der Erde mit den nach dem Benfordschen Gesetz prognostizierten relativen Häufigkeiten verglichen.



1.2 Diskrete Zufallsvariablen und ihre Verteilung

Sei $(\Omega, \mathcal{A}, P)$ ein gegebener Wahrscheinlichkeitsraum. Meistens ist man nicht so sehr an den Elementen $\omega \in \Omega$ selbst interessiert, sondern an den Werten $X(\omega)$, die bestimmte von ω (also vom Zufall) abhängende Größen

X annehmen. Entsprechende Abbildungen $\omega \to X(\omega)$ nennt man Zufallsvariablen, wenn die Ereignisse

$$\{X \in B\} \;=\; \{\omega \in \Omega : X(\omega) \in B\} \;=\; X^{-1}(B)$$

für hinreichend viele Teilmengen B des Wertebereichs von X in der zugrundeliegenden σ-Algebra $\mathcal{A}$ enthalten sind. Wir beschränken uns zunächst auf Zufallsvariablen mit abzählbarem Wertebereich.

Zufallsvariablen, Verteilung und Massenfunktion

Definition 1.8. (i) Eine **diskrete Zufallsvariable** ist eine Abbildung

$$X : \Omega \to S, \qquad S \text{ abzählbar},$$

so dass für alle $a \in S$ gilt:

$$X^{-1}(a) \;=\; \{\omega \in \Omega \mid X(\omega) = a\} \;\in\; \mathcal{A}. \tag{1.5}$$

für die Menge $X^{-1}(a)$ schreiben wir im folgenden kurz $\{X = a\}$.

(ii) Die **Verteilung** einer diskreten Zufallsvariable $X : \Omega \to S$ ist die Wahrscheinlichkeitsverteilung μ_X auf S mit Gewichten

$$p_X(a) \;=\; P[\{X = a\}] \qquad (a \in S).$$

Statt $P[\{X = a\}]$ schreiben wir auch kurz $P[X = a]$.

Bemerkung. a) Man verifiziert leicht, dass p_X tatsächlich die Massenfunktion einer Wahrscheinlichkeitsverteilung μ_X auf S ist. In der Tat gilt $p_X(a) \geq 0$ für alle $a \in S$. Da die Ereignisse $\{X = a\}$ disjunkt sind, folgt zudem:

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

für eine beliebige Teilmenge $B \subseteq S$ des Wertebereichs von X ist $\{X \in B\}$ wieder ein Ereignis in der σ-Algebra $\mathcal{A}$, denn

$$\{X \in B\} \;=\; \underbrace{\{\omega \in \Omega : X(\omega) \in B\}}_{X^{-1}(B)} \;=\; \bigcup_{a \in B} \underbrace{\{X = a\}}_{\in \mathcal{A}} \;\in\; \mathcal{A}$$

nach der Definition einer σ-Algebra. Wegen der σ-Additivität von P gilt

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

Die Verteilung μ_X gibt also an, mit welchen Wahrscheinlichkeiten die Zufallsvariable X Werte in bestimmten Teilmengen des Wertebereichs S annimmt.

b) Ist Ω selbst abzählbar und $\mathcal{A} = \mathcal{P}(\Omega)$, dann ist jede Abbildung $X : \Omega \to S$ eine Zufallsvariable.

c) Eine **reellwertige Zufallsvariable** ist eine Abbildung $X : \Omega \to \mathbb{R}$, so dass die Mengen $\{X \leq c\} = X^{-1}((-\infty, c])$ für alle $c \in \mathbb{R}$ in der σ-Algebra $\mathcal{A}$ enthalten sind. Man überzeugt sich leicht, dass diese Definition mit der Definition oben konsistent ist, wenn der Wertebereich S eine abzählbare Teilmenge von $\mathbb{R}$ ist.

Wir beginnen mit einem elementaren Beispiel:

Beispiel (Zweimal Würfeln). Sei $P = \mathrm{Unif}(\Omega)$ die Gleichverteilung auf der Menge

$$\Omega \;=\; \{\omega = (x_1, x_2) \;:\; x_i \in \{1, \ldots, 6\}\}.$$

Die Augenzahl des i-ten Wurfs ($i = 1, 2$) wird durch $X_i(\omega) := x_i$ beschrieben. Die Abbildung

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

ist eine diskrete Zufallsvariable. Die Verteilung μ_{X_i} hat die Massenfunktion

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

d.h. μ_{X_i} ist die Gleichverteilung auf S.

Die Summe der Augenzahlen bei beiden Würfen wird durch die Zufallsvariable

$$Y(\omega) \;:=\; X_1(\omega) + X_2(\omega)$$

beschrieben. Die Gewichte der Verteilung von Y sind

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

Die Zufallsvariable Y ist also nicht mehr gleichverteilt !

Das folgende Beispiel verallgemeinert die Situation aus dem letzten Beispiel:

Beispiel. Sei P die Gleichverteilung auf einer endlichen Menge $\Omega = \{\omega_1, \ldots, \omega_n\}$ mit n Elementen, und sei $X : \Omega \to S$ eine beliebige Abbildung in eine Menge S. Setzen wir $x_i := X(\omega_i)$, dann ist X eine Zufallsvariable mit Massenfunktion

$$P[X = a] \;=\; \frac{|\{\omega \in \Omega \;:\; X(\omega) = a\}|}{|\Omega|} \;=\; \frac{|\{1 \le i \le n \;:\; x_i = a\}|}{n}.$$

Die Verteilung μ_X von X unter der Gleichverteilung ist also die empirische Verteilung der Werte $x_1, \ldots, x_n$.

Binomialverteilungen

Wir wollen nun zeigen, wie man von der Gleichverteilung zu anderen fundamentalen Verteilungen der Wahrscheinlichkeitstheorie gelangt. Dazu betrachten wir zunächst eine endliche Menge (Grundgesamtheit, Zustandsraum, Population) S. In Anwendungen können die Elemente von S alles mögliche beschreiben, zum Beispiel die Kugeln in einer Urne, die Einwohner von Bonn, oder die Fledermäuse im Kottenforst. Wir wollen nun die zufällige Entnahme von n Einzelstichproben aus S mit Zurücklegen modellieren. Dazu setzen wir

$$\Omega \;=\; S^n \;=\; \{\omega = (x_1, \ldots, x_n) \;:\; x_i \in S\}.$$

Wir nehmen an, dass alle kombinierten Stichproben gleich wahrscheinlich sind, d.h. die zugrundeliegende Wahrscheinlichkeitsverteilung P sei die Gleichverteilung auf dem Produktraum Ω. Erste relevante Zufallsvariablen sind die Stichprobenwerte $X_i(\omega) = x_i$, $i = 1, \ldots, n$. Wie im ersten Beispiel oben gilt

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

d.h. die Zufallsvariablen X_i sind gleichverteilt auf S. Sei nun $E \subseteq S$ eine Teilmenge des Zustandsraums, die für eine bestimmte Merkmalsausprägung der Stichprobe steht (zum Beispiel Ziehen einer roten Kugel oder

Beobachtung einer bestimmten Fledermausart). Die Ereignisse $\{X_i \in E\}$, dass diese Merkmalsausprägung bei der i-ten Einzelstichprobe vorliegt, haben die Wahrscheinlichkeit

$$P[X_i \in E] \;=\; \mu_{X_i}[E] \;=\; |E|/|S|.$$

Wir betrachten nun die Häufigkeit von E in der gesamten Stichprobe $(X_1, \ldots, X_n)$. Diese wird durch die Zufallsvariable $N : \Omega \to \{0, 1, 2, \ldots, n\}$,

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

beschrieben. Ist $p = |E|/|S|$ die relative Häufigkeit des Merkmals E in der Population S, dann erhalten wir:

Lemma 1.9. *Für $k \in \{0, 1, \ldots, n\}$ gilt:*

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

Beweis. Es gilt

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

Hierbei gibt $\binom{n}{k}$ die Anzahl der Möglichkeiten an, k Indizes aus $\{1, \ldots, n\}$ auszuwählen (diejenigen, für die die Merkmalsausprägung E vorliegt), $|E|^k$ ist die Anzahl der Möglichkeiten für die nun festgelegten k Stichproben Werte aus E zu wählen, und $|S \backslash E|^{n-k}$ ist die Anzahl der Möglichkeiten für die verbleibenden $n - k$ Stichproben Werte aus $S \setminus E$ zu wählen. Da P die Gleichverteilung auf S^n ist, folgt

$$P[N = k] \;=\; \frac{\binom{n}{k} |E|^k \, |S \backslash E|^{n-k}}{|S|^n} \;=\; \binom{n}{k} \left(\frac{|E|}{|S|}\right)^k \left(\frac{|S \backslash E|}{|S|}\right)^{n-k} \;=\; \binom{n}{k} p^k \, (1-p)^{n-k}. \qquad \blacksquare$$

Definition 1.10. Sei $n \in \mathbb{N}$ und $p \in [0, 1]$. Die Wahrscheinlichkeitsverteilung auf $\{0, 1, \ldots, n\}$ mit Massenfunktion

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

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

Bemerkung. Dass $b_{n,p}$ die Massenfunktion einer Wahrscheinlichkeitsverteilung ist, kann man mit der allgemeinen binomischen Formel nachrechnen. Dies ist aber gar nicht notwendig, da sich diese Eigenschaft bereits aus Lemma 1.9 ergibt !

Wir haben gesehen, wie sich die Binomialverteilung aus der Gleichverteilung auf einer endlichen Produktmenge ableiten lässt. Binomialverteilungen treten aber noch allgemeiner auf, nämlich als Verteilung der Häufigkeit des Eintretens unabhängiger Ereignisse mit gleichen Wahrscheinlichkeiten. Ereignisse $E_1, \ldots, E_n$ heißen ***unabhängig***, falls

$$P[E_{i_1} \cap E_{i_2} \cap \ldots \cap E_{i_k}] \;=\; P[E_{i_1}] \cdot P[E_{i_2}] \cdots P[E_{i_k}]$$

für alle $k \le n$ und $1 \le i_1 < i_2 < \ldots < i_k \le n$ gilt. Wir werden Unabhängigkeit systematisch in Abschnitt 2.3 diskutieren. Im Vorgriff darauf erwähnen wir schon die folgende wichtige Aussage:

Sind $E_1, \ldots, E_n$ unabhängige Ereignisse mit Wahrscheinlichkeit $P[E_i] = p$, dann gilt

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

d.h. die Anzahl der Ereignisse, die eintreten, ist binomialverteilt.

Der Beweis folgt in Abschnitt 2.3.

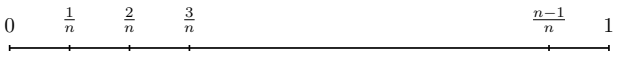
Poissonverteilungen und Poissonscher Grenzwertsatz

Aus der Binomialverteilung lässt sich eine weitere Wahrscheinlichkeitsverteilung ableiten, die die Häufigkeit von seltenen Ereignissen beschreibt. Bevor wir den entsprechenden mathematischen Grenzwertsatz formulieren und beweisen, sehen wir, wie sich in diversen Anwendungssituationen aus einigen wenigen Grundannahmen dasselbe mathematische Modell ergibt, wenn man die Anzahl der Ereignisse, die in einem bestimmten Zeitintervall eintreten, beschreiben möchte.

Beispiel (Seltene Ereignisse in stetiger Zeit). Wir betrachten eine Folge von Ereignissen, die zu zufälligen Zeitpunkten eintreten. Dies können zum Beispiel eingehende Schadensfälle bei einer Versicherung, ankommende Anrufe in einer Telefonzentrale, oder radioaktive Zerfälle sein. Wir sind hier auf der Anwendungsebene - mit „Ereignissen“ meinen wir also im Moment keine mathematischen Objekte. Uns interessiert die Anzahl N der Ereignisse, die in einem festen Zeitintervall der Länge t eintreten. Der Einfachheit halber und ohne wesentliche Beschränkung der Allgemeinheit setzen wir $t = 1$. Wir treffen nun einige Grundannahmen, die näherungsweise erfüllt sein sollten. Diese Grundannahmen sind zunächst wieder auf der Anwendungsebene, und werden erst später durch Annahmen an das mathematische Modell präzisiert. Wir formulieren die Annahmen für die radioaktiven Zerfälle - entsprechende Annahmen gelten aber näherungsweise auch in vielen anderen Situationen.

Annahme 1: „Die Zerfälle passieren ‚unabhängig‘ voneinander zu ‚zufälligen‘ Zeitpunkten“.

Um die Verteilung der Anzahl der Zerfälle pro Zeiteinheit näherungsweise bestimmen zu können, unterteilen wir das Zeitintervall $(0, 1]$ in die n Teilintervalle $((k-1)/n, k/n]$, $k = 1, 2, \ldots, n$:



Annahme 2: „Wenn n sehr groß ist, dann passiert in einer Zeitspanne der Länge $\frac{1}{n}$ ‚fast immer‘ höchstens ein Zerfall“.

In einem stochastischen Modell repräsentiere E_i das Ereignis, dass im Zeitintervall $\left(\frac{i-1}{n}, \frac{i}{n}\right]$ mindestens ein radioaktiver Zerfall stattfindet. Die Wahrscheinlichkeit von E_i sei unabhängig von i und näherungsweise proportional zu $\frac{1}{n}$, also:

Annahme 3: „Es gilt $P[E_i] \approx \lambda/n$ mit einer Konstanten $\lambda \in (0, \infty)$ (der Intensität bzw. Zerfallsrate).“

Wir gehen weiter davon aus, dass sich die erste Annahme dadurch präzisieren lässt, dass wir Unabhängigkeit der Ereignisse $E_1, \ldots, E_n$ fordern. Das ist nicht ganz offensichtlich, lässt sich aber in einem anspruchsvolleren mathematischen Modell, dass die Zeitpunkte aller Zerfälle beschreibt, rechtfertigen. Unter den Annahmen 1, 2 und 3 sollte für das Ereignis, dass genau k radioaktive Zerfälle im Zeitintervall $[0, 1]$ stattfinden, dann näherungsweise gelten, dass

$$P[N = k] \approx P[\text{„genau } k \text{ der } E_i \text{ treten ein“}] \approx b_{n,\frac{\lambda}{n}}(k),$$

wobei $b_{n,\frac{\lambda}{n}}(k)$ das Gewicht von k unter der Binomialverteilung mit Parametern n und $\frac{\lambda}{n}$ ist. Diese Näherung sollte zudem *„für große n immer genauer werden“*. Daher sollten wir die Anzahl der Zerfälle pro Zeiteinheit bei Intensität λ durch eine Zufallsvariable mit nichtnegativen ganzzahligen Werten beschreiben, deren Verteilung die Massenfunktion

$$p_\lambda(k) = \lim_{n\to\infty} b_{n,\frac{\lambda}{n}}(k)$$

hat. Der folgende Satz zeigt, dass p_λ in der Tat die Massenfunktion einer Wahrscheinlichkeitsverteilung ist, nämlich der Poissonverteilung mit Parameter λ.

Satz 1.11 (Poissonapproximation der Binomialverteilung). Sei $\lambda \in (0, \infty)$. Dann gilt:

$$\lim_{n\to\infty} b_{n,\frac{\lambda}{n}}(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad \text{für alle } k = 0, 1, 2, \ldots.$$

Beweis. Für $k \in \{0,1,2,\ldots\}$ und $n \to \infty$ gilt

$$\begin{aligned} b_{n,\lambda/n}(k) &= \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\cdot\underbrace{\frac{n\cdot(n-1)\cdot\ldots\cdot(n-k+1)}{n^k}}_{\to 1}\cdot\underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}}\cdot\underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\to 1} \longrightarrow \frac{\lambda^k}{k!}\,e^{-\lambda}. \end{aligned}$$

■

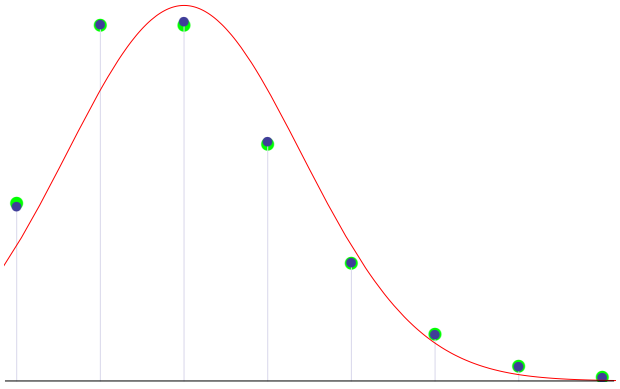
Definition 1.12. Die Wahrscheinlichkeitsverteilung auf $\{0,1,2,\ldots\}$ mit Massenfunktion

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

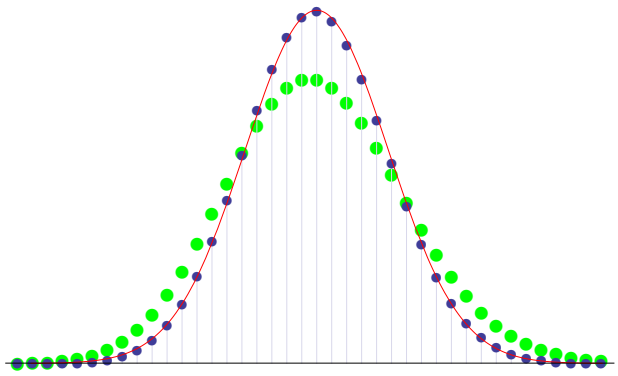
heißt **Poissonverteilung mit Parameter (Intensität)** λ.

Aufgrund des Satzes verwendet man die Poissonverteilung zur näherungsweisen Modellierung der *Häufigkeit seltener Ereignisse* (zum Beispiel Rechtschreibfehler in einer Zeitung, Programmfehler in einer Software, Lottogewinne, Unfälle oder Naturkatastrophen, Zusammenbrüche von Mobilfunknetzen, usw.), und damit zur „Approximation“ von Binomialverteilungen mit kleinen Erfolgswahrscheinlichkeiten p.
Für häufigere Ereignisse (zum Beispiel wenn die Erfolgswahrscheinlichkeit p unabhängig von n ist) verwendet man hingegen besser eine Normalverteilung zur näherungsweisen Modellierung der (geeignet reskalierten) relativen Häufigkeit $\frac{k}{n}$ des Ereignisses für große n. Definition und Eigenschaften von Normalverteilungen werden wir später kennenlernen.
Die folgenden (mit MATHEMATICA erstellten) Graphiken zeigen die Poisson- und Normalapproximation (Poissonverteilung grün, reskalierte Dichte der Normalverteilung rot) der Binomialverteilung Bin(n,p) (blau) für unterschiedliche Parameterwerte:

$n = 100,\ p = 0,02$



$n = 100,\ p = 0,35$



Hypergeometrische Verteilungen

Abschließend zeigen wir, wie sich eine weitere Klasse von Wahrscheinlichkeitsverteilungen, die hypergeometrischen Verteilungen, aus Gleichverteilungen ableiten lässt. Diese Verteilungen treten bei der Entnahme von Stichproben ohne Zurücklegen aus einer Gesamtpopulation auf.

Beispiel (Stichproben ohne Zurücklegen). Wir betrachten eine Population S mit insgesamt m Objekten, z.B. die Kugeln in einer Urne, die Wähler in einem Bundesland, oder die Bäume in einem Waldstück.

Unter den m Objekten seien r, die eine gewisse Eigenschaft/ Merkmalsausprägung besitzen (z.B. Wähler einer bestimmten Partei), und $m-r$, die diese Eigenschaft nicht besitzen. Wir wollen die Entnahme einer Zufallsstichprobe von n Objekten aus der Population beschreiben, wobei $n \leq \min(r, m-r)$ gelte. Dazu betrachten wir den Grundraum Ω, der aus allen Teilmengen (Stichproben) $\omega \subseteq S$ der Kardinalität n besteht. Die Menge Ω enthält $\binom{m}{n}$ Elemente. Gehen wir davon aus, dass alle Stichproben gleich wahrscheinlich sind, dann wählen wir als zugrundeliegende Wahrscheinlichkeitsverteilung in unserem Modell die Gleichverteilung

$$P \;=\; \mathrm{Unif}(\Omega).$$

Sei nun $N(\omega)$ die Anzahl der Objekte in der Stichprobe ω, die die Merkmalsausprägung haben. Für die Wahrscheinlichkeit, dass genau k der n Objekte in der Stichprobe die Merkmalsausprägung haben, ergibt sich

$$P[N=k] \;=\; \frac{|\{\omega \in \Omega \,:\, N(\omega)=k\}|}{|\Omega|} \;=\; \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}} \qquad (k=0,1,\ldots,n).$$

Definition 1.13. Die Wahrscheinlichkeitsverteilung auf $\{0,1,2,\ldots,n\}$ mit Massenfunktion

$$h_{m,r,n}(k) \;=\; \binom{r}{k}\binom{m-r}{n-k}\Big/\binom{m}{n}$$

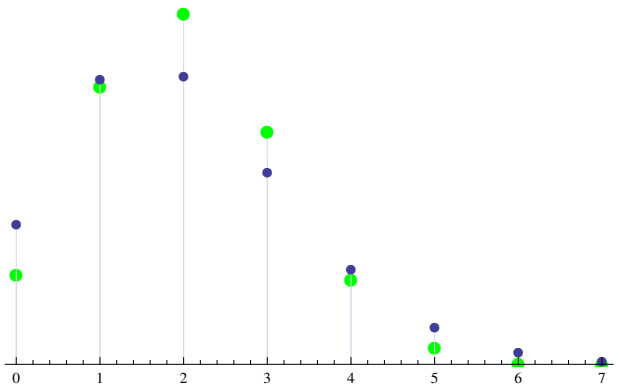
wird **hypergeometrische Verteilung mit Parametern** m**,** r **und** n genannt.

Ist die zugrundeliegende Population im Verhältnis zur Stichprobe groß, dann sollte sich kein wesentlicher Unterschied bei Ziehen mit und ohne Zurücklegen ergeben, da nur sehr selten dasselbe Objekt zweimal gezogen wird. Dies lässt sich mathematisch zum Beispiel folgendermaßen präzisieren: für ein festes $n \in \mathbb{N}$ und $m, r \longrightarrow \infty$ mit $p = r/m$ fest gilt

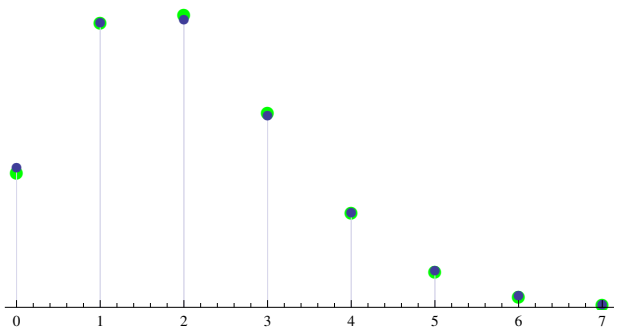
$$h_{m,r,n}(k) \;\longrightarrow\; \binom{n}{k} p^k (1-p)^k,$$

d.h. die hypergeometrische Verteilung mit Parametern m, pm und n nähert sich der Binomialverteilung $\mathrm{Bin}(n,p)$ an. Der Beweis ist eine Übungsaufgabe. Die folgenden (mit MATHEMATICA erstellten) Graphiken zeigen die Gewichte der Binomialverteilung Bin(n,p) (blau) und der hypergeometrischen Verteilung Hyp(m,pm,n) (grün) für unterschiedliche Parameterwerte:

$n = 100,\ p = 0,02,\ m = 300$



$n = 100,\ p = 0,02,\ m = 3000$



1.3 Erwartungswert

Eine erste wichtige Kenngröße reellwertiger Zufallsvariablen ist ihr Erwartungswert. Wir betrachten eine Zufallsvariable $X : \Omega \to S$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$, deren Wertebereich S eine abzählbare Teilmenge der reellen Zahlen ist. In diesem Fall können wir den Erwartungswert (Mittelwert) von X bzgl. der zugrundeliegenden Wahrscheinlichkeitsverteilung P als gewichtetes Mittel der Werte von X definieren:

Definition 1.14. Der **Erwartungswert** von X bzgl. P ist gegeben durch

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

sofern die Summe auf der rechten Seite wohldefiniert ist.

Nimmt die Zufallsvariable X nur nichtnegative Werte $X(\omega) \geq 0$ an, dann sind alle Summanden der Reihe nichtnegativ, und der Erwartungswert $E[X]$ ist wohldefiniert in $[0, \infty]$. Weiterhin ist $E[X]$ wohldefiniert und endlich, falls die Reihe absolut konvergiert. Allgemeiner können wir den Erwartungswert immer dann definieren, wenn

$$\sum_{a \in S, a<0} |a| \cdot P[X = a] \;<\; \infty \qquad \text{gilt.}$$

Der Erwartungswert $E[X]$ wird häufig als *Prognosewert* für $X(\omega)$ verwendet, wenn keine weitere Information vorliegt.

Bemerkung. Nach der Definition *hängt der Erwartungswert nur von der Verteilung μ_X der Zufallsvariablen X ab* ! Wir bezeichnen $E[X]$ daher auch als **Erwartungswert der Wahrscheinlichkeitsverteilung** μ_X auf $\mathbb{R}$.

Beispiel (Gleichverteilte Zufallsvariablen). Ist X gleichverteilt auf einer endlichen Teilmenge $S = \{a_1, \ldots, a_n\}$ von $\mathbb{R}$ mit $a_i \neq a_j$ für $i \neq j$, dann ist der Erwartungswert $E[X]$ das arithmetische Mittel der Werte von X:

$$E[X] \;=\; \frac{1}{n} \sum_{i=1}^{n} a_i.$$

Beispiel (Poissonverteilung). Für eine mit Parameter λ Poisson-verteilte Zufallsvariable N gilt

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

Beschreibt N die Häufigkeit eines Ereignisses (pro Zeiteinheit), dann können wir den Parameter λ dementsprechend als *mittlere Häufigkeit* oder *Intensität* interpretieren.

Beispiel (Erwartungswerte von Indikatorfunktionen). Die Indikatorfunktion eines Ereignisses $A \in \mathcal{A}$ ist die durch

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

definierte Zufallsvariable. Für den Erwartungswert gilt

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

Beträgt beispielsweise die Leistung in einem elementaren Versicherungskontrakt

$$Y(\omega) \;=\; \begin{cases} c & \text{falls } \omega \in A, \qquad \text{„Schadensfall“}, \\ 0 & \text{sonst}, \end{cases}$$

dann gilt $Y = c \cdot I_A$, und

$$E[Y] \;=\; c \cdot P[A].$$

Transformationssatz

Sei $X\colon \Omega \to S$ eine Zufallsvariable mit Werten in einer beliebigen abzählbaren Menge S (die nicht notwendig aus reellen Zahlen besteht). Dann können wir Erwartungswerte von Zufallsvariablen der Form

$$g(X)(\omega) \;:=\; g(X(\omega))$$

mit einer Funktion $g\colon S \to \mathbb{R}$ berechnen. Anstatt dabei über die Werte von $g(X)$ zu summieren, können wir den Erwartungswert auch direkt aus der Verteilung von X erhalten.

Satz 1.15 (Transformationssatz). Für jede reellwertige Funktion $g : S \to \mathbb{R}$ ist

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

eine diskrete Zufallsvariable. Es gilt

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

falls die Summe wohldefiniert ist (also zum Beispiel falls g nichtnegativ ist, oder die Reihe absolut konvergiert).

Beweis. Wegen $\{g(X) = b\} = \bigcup_{a \in g^{-1}(b)} \{X = a\} \in \mathcal{A}$ für alle $b \in g(S)$ ist $g(X)$ wieder eine Zufallsvariable. Da die Vereinigung disjunkt ist, erhalten wir unter Verwendung der σ-Additivität:

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

■

Beispiele. Sei $X : \Omega \to S \subset \mathbb{R}$ eine reellwertige Zufallsvariable mit abzählbarem Wertebereich S.

a) Für den Erwartungswert von $|X|$ ergibt sich

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

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

b) Die **Varianz** einer reellwertigen Zufallsvariable X mit $E[|X|] < \infty$ ist definiert als mittlere quadratische Abweichung vom Erwartungswert, d.h.,

$$\operatorname{Var}[X] \;:=\; E\left[(X - E[X])^2\right].$$

Kennen wir $E[X]$, dann berechnet sich die Varianz als

$$\operatorname{Var}[X] \;=\; \sum_{a \in S} (a - E[X])^2 \, P[X = a] \;\in\; [0, \infty].$$

Ebenso wie der Erwartungswert hängt auch die Varianz nur von der Verteilung μ_X ab.

c) Ist Ω selbst abzählbar, dann können wir den Erwartungswert auch als *gewichtetes Mittel* über $\omega \in \Omega$ darstellen. In der Tat folgt für $X\colon \Omega \to \mathbb{R}$ durch Anwenden des Transformationssatzes:

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

wobei $id_\Omega(\omega) = \omega$ die identische Abbildung auf Ω bezeichnet. Ist P die Gleichverteilung auf Ω, so ist der Erwartungswert das *arithmetische Mittel*

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

Beispiel (Sankt-Petersburg-Paradoxon). Wir betrachten ein Glücksspiel mit fairen Münzwürfen $X_1, X_2, \ldots$, wobei sich der Gewinn in jeder Runde verdoppelt bis zum ersten Mal „Kopf" fällt. Danach ist das Spiel beendet, und der Spieler erhält den Gewinn ausbezahlt. Wie hoch wäre eine faire Teilnahmegebühr für dieses Spiel?

Wir können den Gewinn beschreiben durch die Zufallsvariable

$$G(\omega) \;=\; 2^{T(\omega)}, \qquad \text{mit} \qquad T(\omega) \;=\; \min\{n \in \mathbb{N} \,:\, X_n(\omega) = 0\}.$$

Hierbei beschreibt T die Wartezeit auf den ersten „Kopf". Als Erwartungswert des Gewinns erhalten wir nach dem Transformationssatz

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

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

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

Linearität und Monotonie des Erwartungswertes

Eine fundamentale Eigenschaft des Erwartungswerts ist, dass dieser linear von der Zufallsvariable abhängt. Dies kann häufig ausgenutzt werden, um Erwartungswerte zu berechnen, siehe dazu die Beispiele unten.

Satz 1.16 (Linearität des Erwartungswerts). Seien $X : \Omega \to S_X \subseteq \mathbb{R}$ und $Y : \Omega \to S_Y \subseteq \mathbb{R}$ diskrete reellwertige Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$, für die $E[|X|]$ und $E[|Y|]$ endlich sind. Dann gilt:

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

Beweis. Wir betrachten die durch $g(x, y) = \lambda x + \mu y$ definierte Abbildung $g : S_X \times S_Y \to \mathbb{R}$. Nach dem Transformationssatz ist $g(X, Y) = \lambda X + \mu Y$ eine Zufallsvariable mit Werten in $\mathbb{R}$ und Erwartungswert

$$\begin{aligned} E[\lambda X + \mu Y] \;&=\; E[g(X,Y)] \;=\; \sum_{(a,b) \in S_X \times S_Y} g(a,b) \, P[(X,Y) = (a,b)] \qquad (1.6) \\ &= \sum_{a \in S_X} \sum_{b \in S_Y} (\lambda a + \mu b) \, P[X = a, Y = b] \\ &= \lambda \sum_{a \in S_X} a \sum_{b \in S_Y} P[X = a, Y = b] + \mu \sum_{b \in S_Y} b \sum_{a \in S_X} P[X = a, Y = b] \\ &= \lambda \sum_{a \in S_X} a \, P[X = a] + \mu \sum_{b \in S_Y} b \, P[Y = b] \\ &= \lambda \, E[X] + \mu \, E[Y]. \end{aligned}$$

Hierbei haben wir benutzt, dass die Reihe in (1.6) absolut konvergiert, da nach einer analogen Rechnung

$$\sum_{a\in S_X}\sum_{b\in S_Y} |\lambda\, a + \mu\, b|\, P[X=a, Y=b] \;\le\; |\lambda| \sum_{a\in S_X} |a|\, P[X=a] + |\mu| \sum_{b\in S_Y} |b|\, P[Y=b]$$
$$= \;|\lambda|\, E[|X|] + |\mu|\, E[|Y|]$$

gilt. Die rechte Seite ist nach Voraussetzung endlich. ■

Beispiel (Varianz). Für die Varianz einer reellwertigen Zufallsvariable X mit $E[|X|] < \infty$ gilt

$$\begin{aligned}\mathrm{Var}[X] &= E\left[(X - E[X])^2\right] \;=\; E\left[X^2 - 2X\,E[X] + E[X]^2\right] \\ &= E\left[X^2\right] - E[X]^2\,.\end{aligned}$$

Aus der Linearität folgt auch, dass der Erwartungswert monoton von der Zufallsvariablen abhängt:

Korollar 1.17 (Monotonie des Erwartungswerts). Seien die Voraussetzungen von Satz 1.16 erfüllt. Ist $X(\omega) \le Y(\omega)$ für alle $\omega \in \Omega$, dann gilt

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

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

$$0 \;\le\; E[Y - X] \;=\; E[Y] - E[X].$$

Die folgenden Beispiele demonstrieren, wie die Linearität häufig ausgenutzt werden kann, um Erwartungswerte auf einfache Weise zu berechnen:

Beispiel (Unabhängige 0-1-Experimente, Erwartungswert der Binomialverteilung).
Seien $A_1, A_2, \ldots, A_n \in \mathcal{A}$ unabhängige Ereignisse mit Wahrscheinlichkeit p, und sei $X_i = I_{A_i}$ die Indikatorfunktion des Ereignisses A_i. Die Zufallsvariablen X_i sind *Bernoulli-verteilt mit Parameter p*, d.h. es gilt

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

Damit erhalten wir

$$E[X_i] \;=\; E[I_{A_i}] \;=\; P[A_i] \;=\; p \qquad \forall\, i = 0, 1, \ldots n.$$

Die Anzahl

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

der Ereignisse, die eintreten, ist binomialverteilt mit Parametern n und p, d.h.

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

Den Erwartungswert kann man daher folgendermaßen berechnen:

$$E[S_n] \;=\; \sum_{k=0}^{n} k \cdot P[S_n = k] \;=\; \sum_{k=0}^{n} k \binom{n}{k} p^k\,(1-p)^{n-k} \;=\; \ldots \;=\; n\,p.$$

Einfacher benutzt man aber die Linearität des Erwartungswerts, und erhält direkt

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

Dies gilt sogar wenn die Ereignisse $A_1, \ldots, A_n$ *nicht unabhängig* sind !

Beispiel (Abhängige 0-1-Experimente, Erwartungswert der hypergeometrischen Verteilung). Wir betrachten eine Population aus m Objekten, darunter r, die eine gewisse Eigenschaft besitzen. Aus der Population wird eine Zufallsstichprobe aus n Objekten ohne Zurücklegen entnommen, wobei $n \leq \min(r, m-r)$ gelte. Sei A_i das Ereignis, dass das i-te Objekt in der Stichprobe die Eigenschaft besitzt, und sei $X_i = I_{A_i}$. Dann beschreibt die hypergeometrisch verteilte Zufallsvariable

$$S_n = X_1 + \cdots + X_n$$

die Anzahl der Objekte in der Stichprobe mit der Eigenschaft. Als Erwartungswert der Verteilung $\mathrm{Hyp}(m, r, n)$ erhalten wir daher analog zum letzten Beispiel:

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

Auch im nächsten Beispiel wird eine ähnliche Methode benutzt, um den Erwartungswert zu berechnen:

Beispiel (Inversionen von Zufallspermutationen und Sortieren durch Einfügen). Seien P die Gleichverteilung auf der Menge $\Omega = \mathcal{S}_n$ aller Bijektionen $\omega\colon \{1,\dots,n\} \to \{1,\dots,n\}$, und

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

die Anzahl der Inversionen einer Permutation $\omega \in \mathcal{S}_n$. Dann gilt

$$N = \sum_{1 \leq i < j \leq n} I_{A_{i,j}}, \qquad \text{wobei} \qquad A_{i,j} \;=\; \{\omega \in \mathcal{S}_n : \omega(i) > \omega(j)\}$$

das Ereignis ist, dass eine Inversion von i und j auftritt. Damit erhalten wir

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

Anwendung: Beim Sortieren durch Einfügen („Insertion Sort“) werden die Werte einer Liste $\{\omega(1), \omega(2), \dots, \omega(n)\}$ der Reihe nach an der richtigen Stelle eingefügt. Dabei wird der Wert $\omega(i)$ für $i < j$ beim Einfügen von $\omega(j)$ genau dann verschoben, wenn $\omega(j) < \omega(i)$ gilt. Ist die Anfangsanordnung eine zufällige Permutation der korrekten Anordnung, dann ist die mittlere Anzahl der Verschiebungen, die der Algorithmus vornimmt, also gleich $n\,(n-1)/4$.

Einschluss-/Ausschlussprinzip

Auch das schon oben erwähnte Einschluss-/Ausschlussprinzip lässt sich mithilfe von Indikatorfunktionen elegant beweisen. Dazu verwenden wir die elementaren Identitäten

$$I_{A \cap B} = I_A \cdot I_B \qquad \text{und} \qquad I_{A^C} = 1 - I_A.$$

Satz 1.18 (Einschluss-/Ausschlussprinzip). Für $n \in \mathbb{N}$ und Ereignisse $A_1, \dots, A_n \in \mathcal{A}$ gilt:

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

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

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

Damit folgt

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

■

2 Bedingte Wahrscheinlichkeiten und Unabhängigkeit

Um den Zusammenhang zwischen mehreren Ereignissen oder Zufallsvariablen zu beschreiben sind bedingte Wahrscheinlichkeiten von zentraler Bedeutung. In diesem Kapitel werden bedingte Wahrscheinlichkeiten eingeführt, und mehrstufige Modelle mithilfe bedingter Wahrscheinlichkeiten konstruiert. Anschließend werden wir den Begriff der Unabhängigkeit von Ereignissen und Zufallsvariablen systematisch einführen, und erste wichtige Aussagen unter Unabhängigkeitsannahmen herleiten.

2.1 Bedingte Wahrscheinlichkeiten

Sei $(\Omega, \mathcal{A}, \mathrm{P})$ ein fester Wahrscheinlichkeitsraum, und seien $A, B \in \mathcal{A}$ Ereignisse. Angenommen, wir wissen bereits, dass das Ereignis B eintritt, und wir wollen die Wahrscheinlichkeit von A unter dieser Prämisse angeben. Dann sollten wir nur noch die Fälle $\omega \in B$ in Betracht ziehen, und für diese tritt das Ereignis ein, wenn ω in $A \cap B$ enthalten ist. Damit ist die folgende Definition naheliegend:

Definition 2.1. Sei $A, B \in \mathcal{A}$ mit $P[B] \neq 0$. Dann heißt

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

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

Eine weitere Motivation für die Definition liefern relative Häufigkeiten: Ist P eine empirische Verteilung, dann sind $P[A \cap B]$ und $P[B]$ die relativen Häufigkeiten von $A \cap B$ und B, und $P[A|B]$ ist damit die relative Häufigkeit von $A \cap B$ unter Elementen aus B. Die Definition ist also auch konsistent mit einer frequentistischen Interpretation der Wahrscheinlichkeit als Grenzwert von relativen Häufigkeiten.

Bemerkung. a) Der Fall $P[B] = 0$ muss ausgeschlossen werden, da sonst sowohl Zähler als auch Nenner in dem Bruch in der Definition gleich 0 sind. Bedingte Wahrscheinlichkeiten gegeben Nullmengen sind im Allgemeinen nicht wohldefiniert.

b) Ist $P[B] \neq 0$, dann ist durch die Abbildung

$$P[\,\bullet\,|B]\colon\; A \mapsto\; P[A|B]$$

wieder eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$ gegeben, die ***bedingte Verteilung unter P gegeben B*** . Der Erwartungswert

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

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

Beispiel (Gleichverteilung). Ist P die Gleichverteilung auf einer endlichen Menge Ω, dann gilt:

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

Erste Anwendungsbeispiele

Bei der mathematischen Modellierung von Anwendungsproblemen unter Verwendung bedingter Wahrscheinlichkeiten können leicht Fehler auftreten. An dieser Stelle sollte man also sehr sorgfältig argumentieren, und ggf. zur Kontrolle verschiedene Modellvarianten verwenden. Wir betrachten einige bekannte Beispiele.

Beispiel (Mädchen oder Junge). Wie groß ist die Wahrscheinlichkeit, dass in einer Familie mit zwei Kindern beide Kinder Mädchen sind, wenn mindestens eines der Kinder ein Mädchen ist ? Hier können wir als Wahrscheinlichkeitsraum

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

ansetzen. Wir nehmen vereinfachend an, daß alle Fälle gleich wahrscheinlich sind. Dann gilt:

$$P[\text{„beide Mädchen“} \mid \text{„mindestens ein Mädchen“}] = \frac{|\{MM\}|}{|\{MM, JM, MJ\}|} = \frac{1}{3}.$$

Wir modifizieren die Fragestellung nun etwas. Angenommen, im Nachbarhaus ist heute eine neue Familie eingezogen. Alles, was wir wissen, ist, daß die Familie zwei Kinder hat. Nun sehen wir am Fenster ein Mädchen winken, und gehen davon aus, daß dies eines der beiden Kinder ist. Wie hoch ist nun die Wahrscheinlichkeit, daß beide Kinder Mädchen sind ? Die naheliegende Antwort $1/3$ ist in diesem Fall nicht richtig. Dadurch, daß eines der Kinder winkt, sind die Kinder für uns nicht mehr ununterscheidbar. Die Wahrscheinlichkeit, dass das zweite (nicht winkende) Kind ein Mädchen ist, beträgt dann $1/2$:

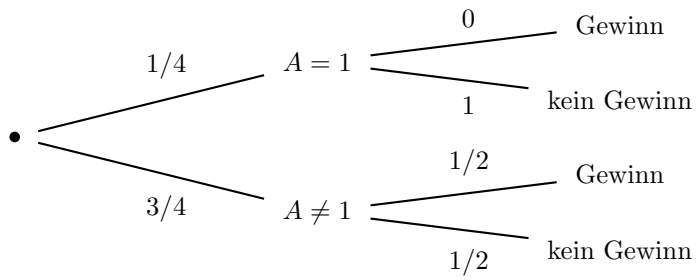
$$P[\text{„beide Mädchen“} \mid \text{„das erste ist Mädchen“}] = \frac{|\{MM\}|}{|\{MM, MJ\}|} = \frac{1}{2}.$$

Haben wir noch Zweifel an der Richtigkeit dieser Aussage, könnten wir ein präziseres Modell aufstellen. Beispielsweise könnten wir das Geschlecht des älteren und des jüngeren Kindes durch Zufallsvariablen $X_1, X_2 : \Omega \to \{M, J\}$, und die Auswahl des winkenden Kindes durch eine weitere Zufallsvariable $K : \Omega \to \{1, 2\}$ beschreiben, wobei $K = 1, 2$ bedeutet, dass das ältere bzw. jüngere Kind winkt. Nehmen wir an, dass (X_1, X_2, K) gleichverteilt auf der Menge $\{M, J\}^2 \times \{1, 2\}$ ist, dann ergibt sich

$$P[\text{„beide Mädchen“} \mid \text{„Mädchen winkt“}] = \frac{P[X_1 = X_2 = M]}{P[X_K = M]} = \frac{2/8}{4/8} = \frac{1}{2}.$$

Beispiel (Ziegenproblem). In einer leicht abgewandelten Version der Spielshow „Let's make a deal“ steht hinter einer von vier Türen ein Auto, und hinter den drei anderen Türen eine Ziege. Der Kandidat wählt zunächst eine der Türen aus (Tür 1). Anschließend öffnet der Moderator eine der verbleibenden Türen (Tür 2, 3 oder 4), wobei nie die Tür mit dem Auto geöffnet wird. Nun hat der Kandidat die Möglichkeit, die Tür nochmal zu wechseln, oder bei seiner ursprünglichen Wahl zu bleiben. Was ist die günstigere Strategie um das Auto zu gewinnen ?

Sie A die Nummer der Tür mit dem Auto. Bleibt der Kandidat bei seiner ursprünglichen Wahl, dann beträgt die Gewinnwahrscheinlichkeit offensichtlich $1/4$, da er bei zufälliger Position des Autos zu Beginn mit Wahrscheinlichkeit $1/4$ die richtige Tür gewählt hat. Die Situation beim Wechseln können wir uns durch das folgende Baumdiagramm klarmachen:



Steht das Auto hinter Tür 1, dann gewinnt der Spieler beim Wechseln nie. Steht das Auto dagegen hinter einer anderen Tür, dann öffnet der Moderator eine weitere Tür. Damit bleiben beim Wechseln nur noch zwei Türen zur Auswahl, und der Spieler gewinnt in diesem Fall mit Wahrscheinlichkeit $1/2$. Insgesamt beträgt die Gewinnwahrscheinlichkeit mit Wechseln also

$$p \;=\; \frac{1}{4}\cdot 0 + \frac{3}{4}\cdot\frac{1}{2} \;=\; \frac{3}{8},$$

d.h. Wechseln ist für den Kandidaten vorteilhaft.

Formal könnten wir die Situation durch Zufallsvariablen $A, M : \Omega \to \{1,2,3,4\}$ beschreiben, die die Nummern der Tür mit dem Auto und der vom Moderator geöffneten Tür angeben. Es ist dann naheliegend anzusetzen, dass A gleichverteilt ist, während M gegeben A bedingt gleichverteilt auf $\{2,3,4\} \setminus A$ ist, d.h.

$$P[M = k|A = 1] = 1/3 \quad \text{für } k \neq 1, \qquad P[M = k|A = 2] = \begin{cases} 1/2 & \text{für } k = 3,4, \\ 0 & \text{sonst,} \end{cases} \qquad \text{usw.}$$

Prüfen Sie selbst nach, dass sich in diesem Modell

$$P[A = k|M \neq k] \;=\; 3/8 \qquad \text{für } k = 2,3,4$$

ergibt, d.h. bei Wechseln zu einer Tür $k \neq 1$, die der Moderator nicht geöffnet hat, beträgt die Gewinnwahrscheinlichkeit $3/8$.

Beispiel (Münzwürfe mit partieller Information). Bei 20 fairen Münzwürfen fällt 15-mal „Zahl“. Wie groß ist die Wahrscheinlichkeit, dass die ersten 5 Würfe „Zahl“ ergeben haben ? Sei P die Gleichverteilung auf

$$\Omega \;=\; \{0,1\}^{20} \;=\; \{\omega = (x_1, \ldots x_{20}) \,:\, x_i \in \{0,1\}\},$$

und sei $X_i(\omega) = x_i$ der Ausgang des i-ten Wurfs. Dann gilt:

$$P\left[X_1 = \ldots = X_5 = 1 \;\middle|\; \sum_{i=1}^{20} X_i = 15\right] = \frac{P\left[X_1 = \ldots = X_5 = 1 \text{ und } \sum_{i=6}^{20} X_i = 10\right]}{P\left[\sum_{i=1}^{20} X_i = 15\right]}$$
$$= \frac{\binom{15}{10}}{\binom{20}{15}} = \frac{15 \cdot 14 \cdot \cdots \cdot 11}{20 \cdot 19 \cdot \cdots \cdot 16} \approx \frac{1}{5}.$$

Dagegen ist $P[X_1 = \ldots = X_5 = 1] = 1/32$.

Berechnung von Wahrscheinlichkeiten durch Fallunterscheidung

Wir zeigen nun wie man unbedingte Wahrscheinlichkeiten aus bedingten berechnet. Sei $\Omega = \bigcup H_i$ eine disjunkte Zerlegung von Ω in abzählbar viele Teilmengen H_i , $i \in I$. Die Mengen H_i beschreiben unterschiedliche Fälle (oder auch *Hypothesen* in statistischen Anwendungen).

Satz 2.2 (Formel von der totalen Wahrscheinlichkeit). Für alle $A \in \mathcal{A}$ gilt:

$$P[A] = \sum_{\substack{i \in I \\ P[H_i] \neq 0}} P[A|H_i] \cdot P[H_i] \tag{2.1}$$

Beweis. Es ist $A = A \cap (\bigcup_{i \in I} H_i) = \bigcup_{i \in I} (A \cap H_i)$ eine disjunkte Vereinigung, also folgt aus der σ-Additivität und wegen $P[A \cap H_i] \leq P[H_i]$:

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

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

Durch Bedingen auf die Farbe der ersten Kugel erhalten wir nach Satz 2.2:

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

Ein interessanter Effekt ist, dass bei Wechsel der zugrundeliegenden Wahrscheinlichkeitsverteilung die unbedingte Wahrscheinlichkeit eines Ereignisses A selbst dann abnehmen kann, wenn alle bedingten Wahrscheinlichkeiten in (2.1) zunehmen:

Beispiel (Simpson-Paradoxon). Die folgende (im wesentlichen auf Originaldaten basierende) Tabelle zeigt die Zahl der Bewerber und der aufgenommenen Studierenden an der Universität Berkeley in einem bestimmten Jahr:

Bewerbungen in Berkeley						
Statistik 1:	Männer		angenommen (A)	Frauen		angenommen (A)
Empirische Verteilung:	2083		996	1067		349
	$P[A\|M]$	$\approx$	0,48	$P[A\|F]$	$\approx$	0,33

Genauere Analyse durch Unterteilung in 4 Fachbereiche						
Statistik 2:	Männer	angenommen (A)		Frauen	angenommen (A)	
Bereich 1	825	511	62%	108	89	82%
Bereich 2	560	353	63%	25	17	68%
Bereich 3	325	110	34%	593	219	37%
Bereich 4	373	22	6%	341	24	7%

Sei $P_F[A] = P[A|F]$ die relative Häufigkeit der angenommenen Bewerber unter Frauen, und $P_M[A] = P[A|M]$ die entsprechende Annahmequote unter Männern. Hierbei steht P für die zugrundeliegende empirische Verteilung, und P_F sowie P_M sind dementsprechend die empirischen Verteilungen in den Unterpopulationen der weiblichen und männlichen Bewerber. Die vollständige Aufgliederung nach Fachbereichen ergibt folgende Zerlegung in Hypothesen:

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

Im Beispiel ist $P_F[A|H_i] > P_M[A|H_i]$ für *alle* i, aber *dennoch* $P_F[A] < P_M[A]$. Obwohl die Annahmequoten unter männlichen Bewerbern insgesamt höher sind, schneiden also die Frauen in jedem der Fachbereiche besser ab.

Die Gesamtstatistik im Beispiel vermischt verschiedene Populationen und legt deshalb eventuell eine falsche Schlussfolgerung nahe. Bei statistischen Untersuchungen ist es daher wichtig, die Population zunächst in möglichst homogene Unterpopulationen aufzuspalten.

Das Simpson-Paradox tritt auch an vielen anderen Stellen auf. Beispielsweise kann bei der Steuerprogression der Steueranteil insgesamt steigen obwohl der Steuersatz in jeder Einkommensklasse sinkt, weil Personen in höhere Einkommensklassen aufsteigen.

Bayessche Regel

Eine direkte Konsequenz des Satzes von der totalen Wahrscheinlichkeit ist die Bayessche Regel. Wir betrachten erneut eine disjunkte Zerlegung von Ω in Teilmengen (Hypothesen) H_i.

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

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

Korollar 2.3 (Bayessche Regel). Für $A \in \mathcal{A}$ mit $P[A] \neq 0$ ist

$$P[H_i|A] \;=\; \frac{P[A|H_i] \cdot P[H_i]}{\sum\limits_{\substack{k \in I \\ P[H_k] \neq 0}} P[A|H_k] \cdot P[H_k]} \qquad \text{für alle } i \in I \text{ mit } P[H_i] \neq 0,$$

d.h. es gilt die Proportionalität

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

wobei c eine von i unabhängige Konstante ist.

Beweis. Nach Satz 2.2 und der Definition der bedingten Wahrscheinlichkeit erhalten wir

$$P[H_i|A] \;=\; \frac{P[A \cap H_i]}{P[A]} \;=\; \frac{P[A|H_i] \cdot P[H_i]}{\sum\limits_{\substack{k \in I \\ P[H_k] \neq 0}} P[A|H_k] \cdot P[H_k]}.$$

Die Bayessche Regel besagt, dass die *A-posteriori-Wahrscheinlichkeiten* $P[H_i|A]$ als Funktion von i proportional zum Produkt der *A-priori-Wahrscheinlichkeiten* $P[H_i]$ und der *Likelihood-Funktion* $i \mapsto P[A|H_i]$ sind. In dieser und ähnlichen Formen bildet sie das Fundament der Bayesschen Statistik.

Beispiel (Medizinische Tests). Von 10.000 Personen eines Alters habe einer die Krankheit K. Ein Test sei positiv (+) bei 96% der Kranken und bei 0,1% der Gesunden. Liegen keine weiteren Informationen vor (z.B. über Risikofaktoren), dann ergibt sich für die A-priori-und A-Posteriori-Wahrscheinlichkeiten für die Krankheit K vor und nach einem positiven Test:

$$\begin{array}{rrcrrcr} \text{A priori:} & P[K] & = & 0,0001, & P[K^C] & = & 0,9999. \\ \text{Likelihood:} & P[+|K] & = & 0,96, & P[+|K^C] & = & 0,001. \end{array}$$

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

Daraus folgt insbesondere: $P[K^C|+] \approx \frac{10}{11}$, d.h. ohne zusätzliche Informationen (z.B. durch einen weiteren Test) muss man in diesem Fall davon ausgehen, dass $\frac{10}{11}$ der positiv getesteten Personen in Wirklichkeit gesund sind!

2.2 Mehrstufige Modelle

Wir betrachten nun ein n-stufiges Zufallsexperiment. Der Ausgang des k-ten Teilexperiments ($k = 1, \ldots, n$) werde durch eine Zufallsvariable $X_k : \Omega \to S_k$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ beschrieben, wobei wir wieder voraussetzen, dass der Wertebereich S_k abzählbar ist. Wir nehmen an, dass folgendes gegeben ist:

- Die *Verteilung bzw. Massenfunktion von* X_1:

$$P[X_1 = x_1] \;=\; p_1(x_1) \qquad \text{für alle } x_1 \in S_1, \qquad \text{sowie} \tag{2.2}$$

- die *bedingten Verteilungen/Massenfunktionen von* X_k *gegeben* $X_1, \ldots, X_{k-1}$:

$$P[X_k \;= x_k \mid X_1 = x_1, \ldots, X_{k-1} = x_{k-1}] \;=\; p_k(x_k \mid x_1, \ldots, x_{k-1}) \tag{2.3}$$

für $k = 2, \ldots n$ und alle $x_1 \in S_1, \ldots, x_k \in S_k$ mit $P[X_1 = x_1, \ldots, X_{k-1} = x_{k-1}] \neq 0$.

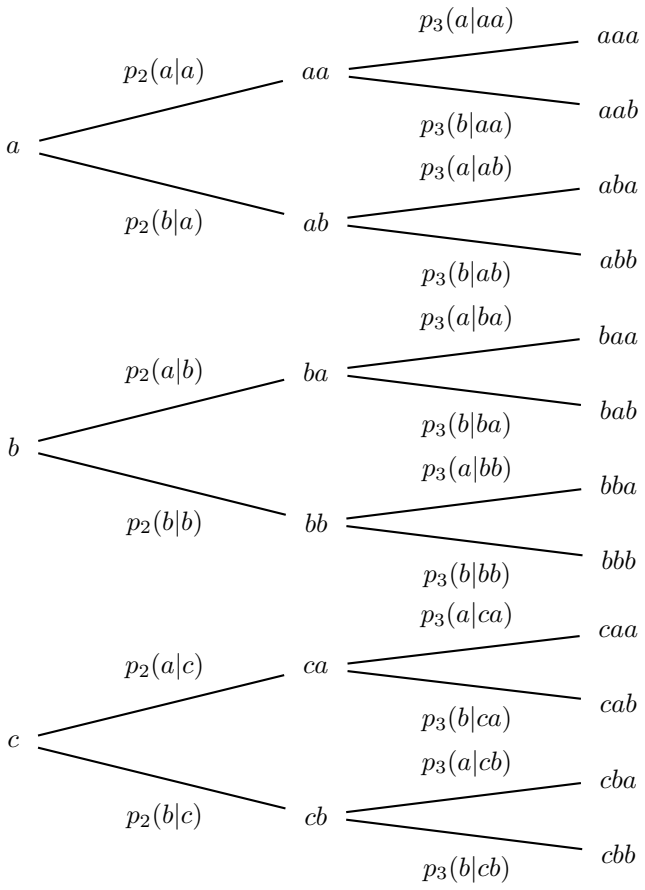


Abbildung 2.1: Dreistufiges Modell mit $S_1 = \{a, b, c\}$ und $S_2 = S_3 = \{a, b\}$.

Zwei wichtige Spezialfälle sind

(i) *Produktmodelle*, in denen die bedingten Massenfunktionen $p_k(\bullet|x_1, \ldots, x_{k-1})$ nicht von den vorherigen Werten $x_1, \ldots, x_{k-1}$ abhängen, sowie

(ii) *Markovketten*, bei denen $p_k(\bullet|x_1, \ldots, x_{k-1})$ nur vom letzten Zustand x_{k-1} abhängt.

Das kanonische Modell

Zufallsvariablen $X_1, \ldots, X_n$, die (2.2) und (2.3) erfüllen, kann man zu gegebenen Massenfunktionen auf unterschiedlichen Wahrscheinlichkeitsräumen realisieren. Im „kanonischen Modell" realisiert man die Zufallsvariablen als Koordinatenabbildungen

$$X_k(\omega) \;=\; \omega_k, \qquad k = 1, \ldots, n,$$

auf dem mit der σ-Algebra $\mathcal{A} = \mathcal{P}(\Omega)$ versehenen Produktraum

$$\Omega \;=\; S_1 \times \ldots \times S_n \;=\; \{(\omega_1, \ldots, \omega_n) \,:\, \omega_i \in S_i\}.$$

Satz 2.4 (Kanonisches Mehrstufenmodell). Seien p_1 und $p_k(\,\bullet \mid x_1, \ldots, x_{k-1})$ für jedes $k = 2, \ldots, n$ und $x_1 \in S_1, \ldots, x_{k-1} \in S_{k-1}$ Massenfunktionen von Wahrscheinlichkeitsverteilungen auf S_k. Dann existiert genau eine Wahrscheinlichkeitsverteilung P auf dem Produktraum $(\Omega, \mathcal{A})$ mit (2.2) und (2.3). Diese ist bestimmt durch die Massenfunktion

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

Beweis. Eindeutigkeit: Wir zeigen durch Induktion, dass für eine Verteilung P mit (2.2) und (2.3) und $k = 1, \ldots n$ gilt:

$$P[X_1 = x_1, \ldots, X_k = x_k] \;=\; p_1(x_1) \cdot p_2(x_2 \mid x_1) \cdots p_k(x_k \mid x_1, \ldots, x_{k-1}). \tag{2.4}$$

Nach (2.2) ist dies für $k = 1$ der Fall. Zudem folgt aus (2.4) für $k - 1$ nach (2.3):

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

also die Behauptung (2.4) für k, falls $P[X_1 = x_1, \ldots, X_{k-1} = x_{k-1}] \neq 0$. Andernfalls verschwinden beide Seiten in (2.4) und die Behauptung ist trivialerweise erfüllt. Für $k = n$ erhalten wir die Massenfunktion von P:

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

Existenz: Die Funktion p ist Massenfunktion einer Wahrscheinlichkeitsverteilung P auf Ω, denn die Gewichte $p(x_1, \ldots, x_n)$ sind nach Voraussetzung nichtnegativ mit

$$\begin{aligned} \sum_{x_1 \in S_1} \cdots \sum_{x_n \in S_n} p(x_1, \ldots, x_n) \;&=\; \sum_{x_1 \in S_1} p_1(x_1) \sum_{x_2 \in S_2} p_2(x_2 \mid x_1) \cdots \underbrace{\sum_{x_n \in S_n} p_n(x_n \mid x_1, \ldots, x_n)}_{=1} \\ &= \; 1. \end{aligned}$$

Hierbei haben wir benutzt, dass die Funktionen $p_k(\bullet | x_1, \ldots, x_{k-1})$ Massenfunktionen von Wahrscheinlichkeitsverteilungen auf S_k sind. Für die Wahrscheinlichkeitsverteilung P auf Ω gilt

$$\begin{aligned} P[X_1 = x_1, \ldots X_k = x_k] \;&=\; \sum_{x_{k+1} \in S_{k+1}} \cdots \sum_{x_n \in S_n} p(x_1, \ldots, x_n) \\ &= p_1(x_1)\, p_2(x_2 \mid x_1) \cdots p_k(x_k \mid x_1, \ldots, x_{k-1}) \end{aligned}$$

für $k = 1, \ldots, n$. Hieraus folgt, dass P die Bedingungen (2.2) und (2.3) erfüllt. ■

Beispiel (Skat). Wie groß ist die Wahrscheinlichkeit, dass beim Skat jeder Spieler genau einen der vier Buben erhält ? Wir beschreiben die Anzahl der Buben der drei Spieler durch die Zufallsvariablen $X_i(\omega) = \omega_i$, $i = 1,2,3$, auf dem Produktraum

$$\Omega \;=\; \{(\omega_1,\omega_2,\omega_3) \;:\; \omega_i \in \{0,1,2,3,4\}\}.$$

Da es insgesamt 32 Karten gibt, von denen jeder Spieler 10 erhält, sind die bedingten Verteilungen der Zufallvariablen X_1, X_2 und X_3 gegeben durch die hypergeometrischen Verteilungen

$$\begin{aligned} p_1(x_1) \;&=\; \binom{4}{x_1}\binom{28}{10-x_1}\Big/\binom{32}{10}, \\ p_2(x_2 \mid x_1) \;&=\; \binom{4-x_1}{x_2}\binom{18+x_1}{10-x_2}\Big/\binom{22}{10} \;\text{ falls } x_1+x_2 \le 4,\;\; 0 \text{ sonst, sowie} \\ p_3(x_3 \mid x_1,x_2) \;&=\; \binom{4-x_1-x_2}{x_3}\binom{8+x_1+x_2}{10-x_3}\Big/\binom{12}{10} \;\text{ falls } 2 \le x_1+x_2+x_3 \le 4,\; 0 \text{ sonst.} \end{aligned}$$

Damit erhalten wir für die gesuchte Wahrscheinlichkeit

$$p(1,1,1) \;=\; p_1(1)\,p_2(1 \mid 1)\,p_3(1 \mid 1,1) \;\approx\; 5{,}56\%.$$

Produktmodelle

Hängt der Ausgang des i-ten Teilexperiments nicht von $x_1,\dots,x_{i-1}$ ab, dann gilt

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

mit einer von $x_1,\dots,x_{i-1}$ unabhängigen Massenfunktion p_i einer Wahrscheinlichkeitsverteilung P_i auf S_i. Sind alle Teilexperimente voneinander unabhängig, dann hat die Wahrscheinlichkeitsverteilung P eines kanonischen n-stufigen Modells die Massenfunktion

$$p(x_1,\dots,x_n) = \prod_{i=1}^{n} p_i(x_i), \qquad x \in S_1 \times \cdot \times S_n. \tag{2.5}$$

Definition 2.5. Seien P_i, $i = 1,\dots n$, Wahrscheinlichkeitsverteilungen auf abzählbaren Mengen S_i mit Massenfunktionen p_i. Die durch die Massenfunktion (2.5) bestimmte Wahrscheinlichkeitsverteilung $P = P_1 \otimes \dots \otimes P_n$ auf $\Omega = S_1 \times \dots \times S_n$ heißt **Produkt** von $P_1,\dots,P_n$.

Beispiel (n-dimensionale Bernoulli-Verteilung). Wir betrachten n unabhängige 0-1-Experimente mit Erfolgswahrscheinlichkeit p, und setzen entsprechend

$$S_i = \{0,1\}, \qquad p_i(1) = p, \qquad p_i(0) = 1-p \qquad \text{für } i = 1,\dots,n.$$

Sei $k = \sum_{i=1}^n x_i$ die Anzahl der Einsen in einem n-Tupel $x \in \Omega = \{0,1\}^n$. Dann hat die Verteilung im Produktmodell die Massenfunktion

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

und wird als **n-dimensionale Bernoulli-Verteilung** bezeichnet.

Beispiel (Produkt von Gleichverteilungen). Sind die Mengen S_i, $i = 1,\dots,n$, endlich, und ist P_i die Gleichverteilung auf S_i, dann ist $P_1 \otimes \dots \otimes P_n$ die Gleichverteilung auf dem Produktraum $S_1 \times \dots \times S_n$.

Die Multiplikativität gilt in Produktmodellen nicht nur für die Massenfunktionen, sondern allgemeiner für die Wahrscheinlichkeiten, dass in den Teilexperimenten bestimmte Ereignisse $A_1,\dots,A_n$ eintreten:

Satz 2.6. Bezüglich des Produkts $P = P_1 \otimes \cdots \otimes P_n$ gilt für beliebige Ereignisse $A_i \subseteq S_i, i = 1, \ldots, n$:

$$\begin{array}{rcl} P[X_1 \in A_1, \ldots, X_n \in A_n] & = & \prod_{i=1}^{n} P[X_i \in A_i] \\ || & & || \\ P[A_1 \times \ldots \times A_n] & & \prod_{i=1}^{n} P_i[A_i] \end{array} \tag{2.6}$$

Beweis. Wegen $(X_1, \ldots, X_n)(\omega) = (\omega_1, \ldots, \omega_n) = \omega$ ist $(X_1, \ldots, X_n)$ die identische Abbildung auf dem Produktraum, und es gilt

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

Insbesondere folgt

$$P[X_i \in A_i] = P[X_1 \in S_1, \ldots, X_{i-1} \in S_{i-1}, X_i \in A_i, X_{i+1} \in S_{i+1}, \ldots, X_n \in S_n] = P_i[A_i],$$

für jedes $i \in \{1, \ldots n\}$, und damit die Behauptung. ■

Bemerkung (Unabhängigkeit). Satz 2.6 besagt, dass die Koordinatenabbildungen $X_i(\omega) = \omega_i$ im Produktmodell *unabhängige Zufallsvariablen* sind, siehe Abschnitt 2.4.

Markovketten

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

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

Oft ist es naheliegend anzunehmen, dass die Weiterentwicklung des Systems nur vom gegenwärtigen Zustand, aber nicht vom vorherigen Verlauf abhängt („kein Gedächtnis“), d.h. es ist

$$p_k(x_k \mid x_0, \ldots, x_{k-1}) = p_k(x_{k-1}, x_k), \tag{2.7}$$

wobei das „Bewegungsgesetz“ $\pi_k : S \times S \to [0, 1]$ folgende Bedingungen erfüllt:

(i) $\pi_k(x, y) \geq 0$ für alle $x, y \in S$,

(ii) $\sum_{y \in S} \pi_k(x, y) = 1$ für alle $x \in S$.

Die Bedingungen (i) und (ii) besagen, dass $\pi_k(x, \bullet)$ für jedes $x \in S$ und $k \in \{1, \ldots n\}$ die Massenfunktion einer Wahrscheinlichkeitsverteilung auf S ist. Diese Wahrscheinlichkeitsverteilung beschreibt die **Übergangswahrscheinlichkeiten** von einem Zustand x zum nächsten Zustand im k-ten Schritt. Die Übergangswahrscheinlichkeiten $\pi_k(x, y)$, $x, y \in S$, kann man in einer Matrix $\pi_k \in \mathbb{R}^{S \times S}$ zusammenfassen. Hat S unendlich viele Elemente, dann ist diese Matrix allerdings unendlich dimensional.

Definition 2.7. Eine Matrix $\pi_k = (\pi_k(x,y))_{x,y\in S} \in \mathbb{R}^{S\times S}$ mit (i) und (ii) heißt **stochastische Matrix** auf S.

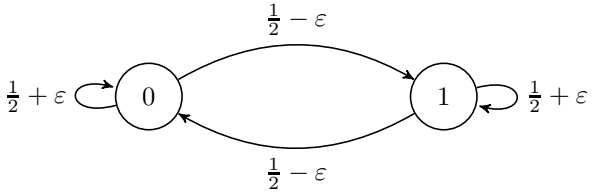
Sei $\nu : S \to [0,1]$ die Massenfunktion der Verteilung von X_0, also der **Startverteilung** der zufälligen Entwicklung. Als Massenfunktion des mehrstufigen Modells ergibt sich dann aus Gleichung (2.7):

$$p(x_0, x_1, \ldots, x_n) \;=\; \nu(x_0)\,\pi_1(x_0,x_1)\,\pi_2(x_1,x_2)\cdots\pi_n(x_{n-1},x_n) \qquad \text{für } x_0,\ldots,x_n \in S,$$

Eine Folge $X_0, X_1, X_2, \ldots, X_n$ von Zufallsvariablen, deren gemeinsame Verteilung durch das beschriebene mehrstufige Modell gegeben ist, nennt man eine **Markovkette** mit Übergangsmatrizen π_k, $k = 1,\ldots,n$. Den Fall, in dem der übergangsmechanismus $\pi_k(x,y) = \pi(x,y)$ unabhängig von k ist, bezeichnet man als ***zeitlich homogen***.

Beispiele. a) Produktmodell: Produktmodelle sind spezielle Markovketten mit Übergangswahrscheinlichkeiten $\pi_k(x,y) = p_k(y)$, die nicht von x abhängen.

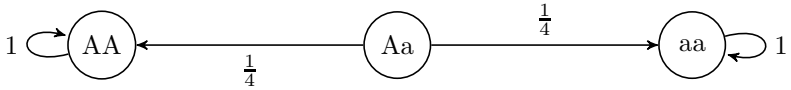
b) Abhängige Münzwürfe: Ein einfaches Modell für abhängige Münzwürfe ist eine Markovkette mit Zustandsraum $S = \{0,1\}$ und den folgenden Übergangswahrscheinlichkeiten:



Hierbei ist $\varepsilon \in \left[-\frac{1}{2}, \frac{1}{2}\right]$ ein Parameter, der die Abhängigkeit des nächsten Münzwurfs vom Ausgang des vorherigen Wurfs bestimmt. Die zeitunabhängige Übergangsmatrix ist

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

c) Selbstbefruchtung von Pflanzen: Die Selbstbefruchtung ist ein klassisches Verfahren zur Züchtung von Pflanzen vom Genotyp AA bzw. aa, wobei A und a zwei mögliche Allele des Pflanzen-Gens sind. Die Übergangswahrscheinlichkeiten zwischen den möglichen Genotypen AA, Aa und aa sind durch



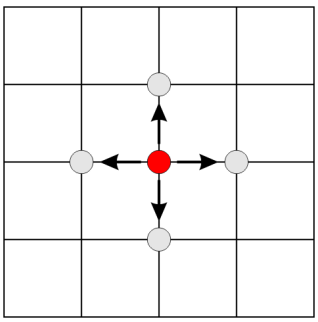
gegeben, und die Übergangsmatrix einer entsprechenden Markovkette ist

$$\pi \;=\; \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 1 \end{pmatrix}.$$

d) Random Walks auf Graphen: Sei $S = V$ die Knotenmenge eines Graphen (V,E). Wir nehmen an, dass jeder Knoten $x \in V$ endlichen Grad $\deg(x)$ hat. Dann ist durch

$$\pi(x,y) \;=\; \begin{cases} \frac{1}{\deg(x)} & \text{falls } \{x,y\} \in E, \\ 0 & \text{sonst,} \end{cases}$$

die zeitunabhängige Übergangsmatrix eines Random Walks auf dem Graphen definiert. Beispielsweise ist der klassische Random Walk (Irrfahrt) auf $S = \mathbb{Z}^d$ die Markovkette, die sich in jedem Schritt zu einem zufällig (gleichverteilt) ausgewählten Nachbarpunkt des gegenwärtigen Zustands weiterbewegt:



Da in d Dimensionen jeder Gitterpunkt $2d$ Nachbarpunkte hat, sind die Übergangswahrscheinlichkeiten durch

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

gegeben. In Dimension $d = 1$ ist die Übergangsmatrix eine unendliche (mit $x \in \mathbb{Z}$ indizierte) Tridiagonalmatrix, die neben der Diagonale die Einträge $1/2$, und auf der Diagonalen die Einträge 0 hat.

Berechnung von Mehr-Schritt-Übergangswahrscheinlichkeiten

Wir berechnen nun die Übergangswahrscheinlichkeiten und Verteilungen einer Markovkette nach mehreren Schritten. Es stellt sich heraus, dass sich diese durch Matrizenmultiplikation der Übergangsmatrizen ergeben. Dazu interpretieren wir die Massenfunktion ν der Startverteilung als Zeilenvektor $(\nu(x))_{x\in S}$ in $\mathbb{R}^S$.

Satz 2.8 (Übergangswahrscheinlichkeiten und Verteilung nach mehreren Schritten).
Für alle $0 \le k < l \le n$ und $x_0, \dots, x_k, y \in S$ mit $P[X_0 = x_0, \dots, X_k = x_k] \neq 0$ gilt

$$\begin{aligned} P[X_l = y \mid X_0 = x_0, \dots, X_k = x_k] &= P[X_l = y \mid X_k = x_k] \\ &= (\pi_{k+1}\, \pi_{k+2} \cdots \pi_l)(x_k, y), \qquad \text{und} \\ P[X_l = y] &= (\nu \pi_1\, \pi_2 \cdots \pi_l)(y). \end{aligned}$$

Hierbei ist

$$(\pi\, \widetilde{\pi})(x,y) \;:=\; \sum_{z\in S} \pi(x,z)\, \widetilde{\pi}(z,y)$$

das Produkt zweier Übergangsmatrizen π und $\widetilde{\pi}$ an der Stelle (x,y), und

$$(\nu\widetilde{\pi})(y) \;=\; \sum_{x\in S} \nu(x)\widetilde{\pi}(x,y)$$

ist das Produkt des Zeilenvektors ν mit einer Übergangsmatrix $\widetilde{\pi}$, ausgewertet an der Stelle y.

Die Matrixprodukte in Satz 2.8 sind auch für abzählbar unendliche Zustandsräume S wohldefiniert, da die Komponenten der Übergangsmatrizen alle nicht-negativ sind.

Bemerkung. a) Markov-Eigenschaft: Der Satz zeigt, dass die Weiterentwicklung einer Markovkette auch für mehrere Schritte jeweils nur vom gegenwärtigen Zustand x_k abhängt, und nicht vom vorherigen Verlauf $x_0, x_1, \dots, x_{k-1}$.

b) n-Schritt-Übergangswahrscheinlichkeiten: Die Übergangswahrscheinlichkeiten für die ersten n Schritte sind nach dem Satz gegeben durch

$$P[X_n = y \mid X_0 = x] \;=\; (\pi_1\, \pi_2 \cdots \pi_n)(x,y).$$

Im *zeitlich homogenen Fall* (d.h. $\pi_i \equiv \pi$ unabhängig von i) ist die n-Schritt-übergangswahrscheinlichkeit von x nach y gleich $\pi^n(x,y)$.

c) Gleichgewichtsverteilungen: Weiterhin ist im *zeitlich homogenen Fall* $\pi_i \equiv \pi$ die Verteilung der Markovkette zur Zeit l gleich $\nu\pi^l$. Gilt $\nu = \nu\pi$, dann stimmt diese für jedes l mit der Startverteilung überein, d.h. die Wahrscheinlichkeitsverteilung ν ist ein *Gleichgewicht* der stochastischen Dynamik, die durch die Übergangsmatrix π beschrieben wird. Gleichgewichte von zeithomogenen Markovketten werden wir in Abschnitt 3.4 weiter untersuchen.

Beweis. Für $x_0, \ldots, x_k, y$ wie im Satz vorausgesetzt gilt

$$\begin{aligned} P[X_l = y \mid X_0 = x_0, \ldots, X_k = x_k] &= \frac{P[X_0 = x_0, \ldots, X_k = x_k, X_l = y]}{P[X_0 = x_0, \ldots, X_k = x_k]} \\ &= \frac{\sum_{x_{k+1},\ldots,x_{l-1}} \nu(x_0)\,\pi_1(x_0,x_1)\cdots\pi_l(x_{l-1},y)}{\nu(x_0)\,\pi_1(x_0,x_1)\cdots\pi_k(x_{k-1},x_k)} \\ &= \sum_{x_{k+1}} \cdots \sum_{x_{l-1}} \pi_{k+1}(x_k,x_{k+1})\,\pi_{k+2}(x_{k+1},x_{k+2})\cdots\pi_l(x_{l-1},y) \\ &= (\pi_{k+1}\,\pi_{k+2}\cdots\pi_l)(x_k,y). \end{aligned}$$

Entsprechend erhalten wir

$$\begin{aligned} P[X_l = y \mid X_k = x_k] &= \frac{P[X_k = x_k, X_l = y]}{P[X_k = x_k]} \\ &= \frac{\sum_{x_1,\ldots,x_{k-1}} \sum_{x_{k+1},\ldots,x_{l-1}} \nu(x_0)\,\pi_1(x_0,x_1)\cdots\pi_l(x_{l-1},y)}{\sum_{x_1,\ldots,x_{k-1}} \nu(x_0)\,\pi_1(x_0,x_1)\cdots\pi_k(x_{k-1},x_k)} \\ &= (\pi_{k+1}\,\pi_{k+2}\cdots\pi_l)(x_k,y). \end{aligned}$$

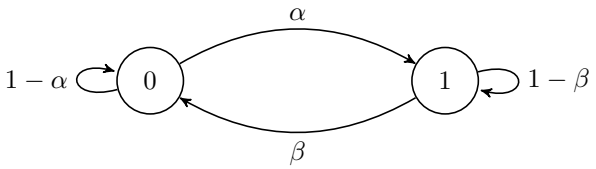
Für die unbedingten Wahrscheinlichkeiten ergibt sich

$$\begin{aligned} P[X_l = y] &= \sum_{\substack{x\in S \\ P[X_0=x]\neq 0}} P[X_0 = x]\,P[X_l = y \mid X_0 = x] \\ &= \sum_{\substack{x\in S \\ \nu(x)\neq 0}} \nu(x)\,(\pi_1\,\pi_2\cdots\pi_l)(x,y) = (\nu\pi_1\,\pi_2\cdots\pi_l)(y). \end{aligned}$$

■

Wir untersuchen abschließend den Spezialfall einer zeithomogenen Markovkette auf einem Zustandsraum mit zwei Elementen. Diesen können wir schon jetzt weitgehend vollständig analysieren:

Beispiel (Explizite Berechnung für Zustandsraum mit zwei Elementen). Wir betrachten eine allgemeine zeithomogene Markovkette mit Zustandsraum $S = \{0,1\}$. Die Übergangswahrscheinlichkeiten



$\pi(x,y)$ sind durch gegeben, wobei wir annehmen, dass $0 < \alpha, \beta \leq 1$ gilt. Die Wahrscheinlichkeitsverteilung μ mit Gewichten $\mu(0) = \frac{\beta}{\alpha+\beta}$ und $\mu(1) = \frac{\alpha}{\alpha+\beta}$ ist ein Gleichgewicht der Übergangsmatrix

$$\pi = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix},$$

denn für den Zeilenvektor $\mu = (\mu(0), \mu(1))$ gilt $\mu\pi = \mu$. Für $n \in \mathbb{N}$ erhalten wir durch Bedingen auf den Wert zur Zeit $n-1$:

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

Daraus folgt mit Induktion

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

Analoge Formeln erhält man für $\pi^n(1,0)$ und $\pi^n(1,1)$ durch Vertauschen von α und β. Für die n-Schritt-Übergangsmatrix ergibt sich also

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

Sind die Übergangswahrscheinlichkeiten α und β nicht beide gleich 1, dann gilt $\pi^n(0,\cdot) \approx \pi^n(1,\cdot) \approx \mu$ für große $n \in \mathbb{N}$. Die Kette „vergisst" also ihren Startwert X_0 exponentiell schnell („Exponentieller Gedächtnisverlust"), und die Verteilung von X_n nähert sich für $n \to \infty$ rasch der Gleichgewichtsverteilung μ an („Konvergenz ins Gleichgewicht") !

2.3 Unabhängigkeit

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

$$\begin{aligned} P[A|B] &= P[A] \qquad \text{falls } P[B] \neq 0, \qquad \text{sowie} \\ P[B|A] &= P[B] \qquad \text{falls } P[A] \neq 0. \end{aligned}$$

Beide Aussagen sind äquivalent zu der Bedingung

$$P[A \cap B] = P[A] \cdot P[B], \tag{2.8}$$

die im Fall $P[A] = 0$ oder $P[B] = 0$ automatisch erfüllt ist. Allgemeiner definieren wir für beliebige (endliche, abzählbare oder überabzählbare) Kollektionen von Ereignissen:

Definition 2.9. Eine Kollektion A_i, $i \in I$, von Ereignissen aus $\mathcal{A}$ heißt **unabhängig** (bzgl. P), falls

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

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

Beispiele. a) Falls $P[A] \in \{0,1\}$ gilt, dann ist A unabhängig von B für alle $B \in \mathcal{A}$. Deterministische Ereignisse sind also von allen anderen Ereignissen unabhängig.

b) Wir betrachten das kanonische Modell für zwei faire Münzwürfe, d.h. P ist die Gleichverteilung auf $\Omega = \{0,1\}^2$. Die drei Ereignisse

$$\begin{aligned} A_1 &= \{(1,0),(1,1)\} \quad \text{„erster Wurf Zahl“,} \\ A_2 &= \{(0,1),(1,1)\} \quad \text{„zweiter Wurf Zahl“,} \\ A_3 &= \{(0,0),(1,1)\} \quad \text{„beide Würfe gleich“,} \end{aligned}$$

sind *paarweise unabhängig*, denn es gilt:

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

Trotzdem ist die Kollektion A_1, A_2, A_3 aller drei Ereignisse *nicht unabhängig*, denn

$$P[A_1 \cap A_2 \cap A_3] \;=\; \frac{1}{4} \;\neq\; \frac{1}{8} \;=\; P[A_1] \cdot P[A_2] \cdot P[A_3].$$

Sind A und B unabhängige Ereignisse, so auch A und B^C, denn es gilt

$$P[A \cap B^C] \;=\; P[A] - P[A \cap B] \;=\; P[A] \cdot (1 - P[B]) \;=\; P[A] \cdot P[B^C].$$

Allgemeiner folgt:

Lemma 2.10 (Stabilität von Unabhängigkeit unter Komplementbildung).
Sind die Ereignisse $A_1, \ldots, A_n \in \mathcal{A}$ unabhängig, und gilt $B_j = A_j$ oder $B_j = A_j^C$ für alle $j = 1, \ldots, n$, dann sind auch die Ereignisse $B_1, \ldots, B_n$ unabhängig.

Beweis. Da wir zum Nachweis der Unabhängigkeit beliebige Unterkollektionen von $\{B_1, \ldots B_n\}$ betrachten müssen, ist zu zeigen, dass

$$P[C_1 \cap \ldots \cap C_n] \;=\; P[C_1] \cdot \cdots \cdot P[C_n]$$

gilt, falls die Ereignisse C_i jeweils gleich A_i, A_i^C oder Ω sind. Sei ohne Beschränkung der Allgemeinheit $C_i = A_i$ für $i \leq k$, $C_i = A_i^C$ für $k < i \leq l$, und $C_i = \Omega$ für $k > l$ mit $0 \leq k \leq l \leq n$. Dann folgt unter Verwendung der Linearität des Erwartungswerts und der Unabhängigkeit von $A_1, \ldots, A_n$:

$$\begin{aligned} P[C_1 \cap \ldots \cap C_n] \;&=\; P\big[A_1 \cap \ldots \cap A_k \cap A_{k+1}^C \cap \ldots \cap A_l^C\big] \\ &=\; E\left[I_{A_1} \cdots I_{A_k} \cdot (1 - I_{A_{k+1}}) \cdots (1 - I_{A_l})\right] \\ &=\; E\Big[I_{A_1} \cdots I_{A_k} \cdot \sum_{J \subseteq \{k+1,\ldots,l\}} (-1)^{|J|} \prod_{j \in J} I_{A_j}\Big] \\ &=\; \sum_{J \subseteq \{k+1,\ldots,l\}} (-1)^{|J|} \, P\Big[A_1 \cap \ldots \cap A_k \cap \bigcap_{j \in J} A_j\Big] \\ &=\; \sum_{J \subseteq \{k+1,\ldots,l\}} (-1)^{|J|} P[A_1] \cdots P[A_k] \cdot \prod_{j \in J} P[A_j] \\ &=\; P[A_1] \cdots P[A_k] \cdot (1 - P[A_{k+1}]) \cdots (1 - P[A_l]) \\ &=\; P[C_1] \cdots P[C_n]. \end{aligned}$$

■

Verteilungen für unabhängige Ereignisse

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

Geometrische Verteilung

Die *„Wartezeit“* auf das erste Eintreten eines der Ereignisse ist durch

$$T(\omega) \;=\; \inf\{n \in \mathbb{N} \;:\; \omega \in A_n\}$$

gegeben, wobei wir hier $\min \emptyset := \infty$ setzen. Mit Lemma 2.10 können wir die Verteilung der Zufallsvariable $T : \Omega \to \mathbb{N} \cup \{\infty\}$ berechnen. Für $n \in \mathbb{N}$ erhalten wir

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

Definition 2.11. Sei $p \in [0,1]$. Die Wahrscheinlichkeitsverteilung μ auf $\mathbb{N} \cup \{\infty\}$ mit Massenfunktion

$$\mu(n) = p \cdot (1-p)^{n-1} \qquad \text{für } n \in \mathbb{N}$$

heißt **geometrische Verteilung zum Parameter** p, und wird kurz mit Geom(p) bezeichnet.

Bemerkung. a) für $n \in \mathbb{N}$ gilt

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

Ist $p \neq 0$, dann folgt insbesondere $P[T = \infty] = 0$, d.h. die geometrische Verteilung ist eine Wahrscheinlichkeitsverteilung auf den natürlichen Zahlen. für $p = 0$ gilt dagegen $P[T = \infty] = 1$.

b) Wegen $T = \sum_{n=0}^{\infty} I_{\{T>n\}}$ ergibt sich als Erwartungswert der geometrischen Verteilung

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

Binomialverteilung

Die *Anzahl* der Ereignisse unter $A_1, \ldots, A_n$, die eintreten, ist durch die Zufallsvariable

$$S_n(\omega) \;=\; |\{1 \le i \le n \;:\; \omega \in A_i\}| \;=\; \sum_{i=1}^{n} I_{A_i}(\omega)$$

gegeben. Mithilfe von Lemma 2.10 können wir auch die Verteilung von S_n berechnen. Für $0 \le k \le n$ gilt

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

d.h. S_n ist *binomialverteilt mit Parametern n und p*.

Exkurs zu gemeinsamen Verteilungen

Um den Zusammenhang zwischen mehreren Zufallsvariablen untersuchen, genügt es nicht, die Verteilungen der einzelnen Zufallsvariablen zu kennen. Stattdessen benötigen wir die *gemeinsame Verteilung* der Zufallvariablen. Diese ist folgendermaßen definiert: Sind $X_1 : \Omega \to S_1, \ldots, X_n : \Omega \to S_n$ diskrete Zufallsvariablen, dann ist auch $(X_1, \ldots, X_n)$ eine diskrete Zufallsvariable mit Werten im Produktraum $S_1 \times \cdots \times S_n$.

Definition 2.12. Die Verteilung $\mu_{X_1,\ldots,X_n}$ des Zufallsvektors $(X_1, \ldots, X_n)$ unter P heißt **gemeinsame Verteilung** der Zufallsvariablen $X_1, \ldots, X_n$.

Die gemeinsame Verteilung ist eine Wahrscheinlichkeitsverteilung auf $S_1 \times \cdots \times S_n$ mit Massenfunktion

$$p_{X_1,\ldots,X_n}(a_1, \ldots, a_n) \;=\; P[X_1 = a_1, \ldots, X_n = a_n] \tag{2.9}$$

Sie enthält Informationen über den Zusammenhang zwischen den Zufallsgrößen X_i. Die Verteilungen der einzelnen Zufallsvariablen X_i nennt man dagegen **Randverteilungen**.

Beispiel (Zwei unabhängige Würfel). Beschreiben die Zufallsvariablen $X, Y : \Omega \to \{1,2,3,4,5,6\}$ auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ die Augenzahlen beim Werfen zweier Würfel, die jede Augenzahl unabhängig voneinander mit Wahrscheinlichkeit $1/6$ annehmen, dann gilt

$$P[X = a, Y = b] \;=\; \frac{1}{36} \qquad \text{für alle } a, b \in \{1,2,3,4,5,6\}.$$

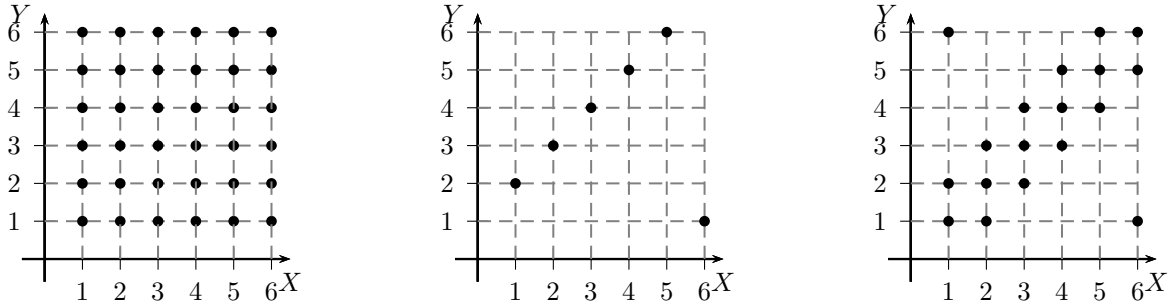
Die gemeinsame Verteilung von X und Y ist also die Gleichverteilung auf $\{1,2,3,4,5,6\}^2$. Sei nun $M = \max(X, Y)$ die größere der beiden Augenzahlen. Dann erhalten wir für die gemeinsame Verteilung von M und X die folgenden Wahrscheinlichkeiten:

$P[M = i, X = j]$	$j = 1$	2	3	4	5	6	$P[M = i]$
$i = 1$	1/36	0	0	0	0	0	1/36
2	1/36	2/36	0	0	0	0	3/36
3	1/36	1/36	3/36	0	0	0	5/36
4	1/36	1/36	1/36	4/36	0	0	7/36
5	1/36	1/36	1/36	1/36	5/36	0	9/36
6	1/36	1/36	1/36	1/36	1/36	6/36	11/36
$P[X = j]$	1/6	1/6	1/6	1/6	1/6	1/6	1

Die Massenfunktionen der Randverteilungen von X und M stehen in der letzten Zeile bzw. Spalte und ergeben sich durch Aufaddieren über die möglichen Werte der jeweils anderen Zufallsvariable.

Das folgende Beispiel zeigt, dass die gemeinsamen Verteilungen auch dann sehr unterschiedlich sein können wenn die Randverteilungen übereinstimmen.

Beispiel (Zwei abhängige Würfel). Seien $X, Y : \Omega \to \{1,2,3,4,5,6\}$ gleichverteilte Zufallsvariablen. Für die Gewichte der gemeinsamen Verteilung von X und Y gibt es dann unter anderem die in Abbildung 2.2 gegebenen Möglichkeiten.



(a) X, Y unabhängig.
Die Gewichte der Punkte sind jeweils gleich $1/36$.
$\mu_{X,Y} = \mu_X \otimes \mu_Y$.

(b) $Y = (X + 1) \mod 6$.
Die Gewichte der Punkte sind jeweils gleich $1/6$.

(c) $Y = (X + Z) \mod 6$,
$Z \sim \mathrm{Unif}\{-1, 0, 1\}$.
Die Gewichte der Punkte sind jeweils gleich $1/18$.

Abbildung 2.2: Gemeinsame Verteilungen mit identischen Randverteilungen.

Die gemeinsame Verteilung der Zufallsvariablen $X_1, \ldots, X_n$ benötigen wir, wenn wir Erwartungswerte von Funktionen berechnen wollen, die von mehreren dieser Zufallsvariablen abhängen. Ist $g : S_1 \times \cdots \times S_n \to [0, \infty)$ eine reellwertige Funktion, dann gilt nach dem Transformationssatz 1.15 nämlich

$$E\left[g(X_1, \ldots, X_n)\right] = \sum_{(a_1, \ldots, a_n)} g(a_1, \ldots, a_n)\, P\left[X_1 = a_1, \ldots, X_n = a_n\right].$$

Beispiel (Erwartungswerte für zwei Würfel). Sind X und Y die Augenzahlen zweier unabhängiger fairer Würfel, dann gilt

$$E\left[X \cdot Y\right] = \frac{1}{36} \sum_{i,j=1}^{6} ij = \left(\frac{1}{6} \sum_{i=1}^{6} i\right)^2 = \left(\frac{7}{2}\right)^2 = \frac{49}{4},$$

$$E\left[\max(X, Y)\right] = \frac{1}{36} \sum_{i,j=1}^{6} \max(i, j) = \frac{1}{36} \sum_{i=1}^{6} i + \frac{2}{36} \sum_{i=1}^{6} \sum_{j=1}^{i-1} i = \frac{161}{36}.$$

Unabhängigkeit von diskreten Zufallsvariablen

Wir erweitern den Begriff der Unabhängigkeit nun von Ereignissen auf Zufallsvariablen. Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, und I eine beliebige Menge.

Definition 2.13. Eine Familie $X_i : \Omega \to S_i$ $(i \in I)$ von Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ mit abzählbaren Wertebereichen S_i heißt **unabhängig**, falls die Ereignisse $\{X_i \in A_i\}$ $(i \in I)$ für alle Teilmengen $A_i \subseteq S_i$ unabhängig sind.

Aus der Definition folgt unmittelbar, dass die Zufallsvariablen X_i $(i \in I)$ genau dann unabhängig sind, wenn jede endliche Teilkollektion unabhängig ist. Daher beschränken wir uns im folgenden auf den Fall $I = \{1, \ldots, n\}$ mit $n \in \mathbb{N}$.

Satz 2.14. Die folgenden Aussagen sind äquivalent:

(i) $X_1, \ldots, X_n$ sind unabhängig.

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

(iii) $p_{X_1,\ldots,X_n}(a_1, \ldots, a_n) = \prod_{i=1}^n p_{X_i}(a_i)$ für alle $a_i \in S_i$, $i = 1, \ldots, n$.

(iv) $\mu_{X_1,\ldots,X_n} = \bigotimes_{i=1}^n \mu_{X_i}$.

Beweis. (i)⇒(ii) folgt durch Wahl von $A_i = \{a_i\}$.

(iii)⇔(iv) gilt nach Definition des Produkts $\bigotimes_{i=1}^n \mu_{X_i}$ der Wahrscheinlichkeitsverteilungen μ_{X_i}.

(iv)⇒(i): Seien $A_i \subseteq S_i$ $(i = 1, \ldots, n)$ und $1 \le i_1 < i_2 < \ldots < i_k \le n$. Um die Produkteigenschaft für die Ereignisse mit Indizes $i_1, \ldots, i_k$ zu zeigen, setzen wir $B_{i_j} := A_{i_j}$ für alle j und $B_i := S_i$ für $i \notin \{i_1, \ldots, i_k\}$. Mit (iv) folgt dann nach Satz 2.6:

$$\begin{aligned} P[X_{i_1} \in A_{i_1}, \ldots, X_{i_k} \in A_{i_k}] &= P[X_1 \in B_1, \ldots, X_n \in B_n] \\ &= P[(X_1, \ldots, X_n) \in B_1 \times \ldots \times B_n] = \mu_{X_1,\ldots,X_n}[B_1 \times \ldots \times B_n] \\ \prod_{i=1}^n \mu_{X_i}[B_i] &= \prod_{i=1}^n P[X_i \in B_i] = \prod_{j=1}^k P[X_{i_j} \in A_{i_j}]. \end{aligned}$$

■

Als Konsequenz aus Satz 2.14 ergibt sich insbesondere:

Korollar 2.15. Sind $X_i : \Omega \to S_i$ $(i = 1, \ldots, n)$ diskrete Zufallsvariablen, und hat die Massenfunktion der gemeinsamen Verteilung eine Darstellung in Produktform

$$p_{X_1,\ldots,X_n}(a_1, \ldots, a_n) = c \cdot \prod_{i=1}^n g_i(a_i) \qquad \forall\, (a_1, \ldots, a_n) \in S_1 \times \ldots \times S_n \tag{2.10}$$

mit Funktionen $g_i : S_i \to [0, \infty)$ und einer Proportionalitätskonstanten $c \in \mathbb{R}$, dann sind $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit Massenfunktionen

$$p_{X_i}(a) = \frac{g_i(a)}{\sum\limits_{b \in S_i} g_i(b)}, \qquad a \in S_i. \tag{2.11}$$

Beweis. Durch Summieren über $a_1, a_2, \ldots, a_n$ in (4.28) folgt $\sum_{b \in S_i} g_i(b) < \infty$ für alle i. Daher ist die Funktion $\widetilde{g}_i(a) := g_i(a) / \sum_{b \in S_i} g_i(b)$ $(a \in S_i)$, die auf der rechten Seite von (2.11) steht, die Massenfunktion einer Wahrscheinlichkeitsverteilung μ_i auf S_i. Nach Voraussetzung gilt für $(a_1, \ldots, a_n) \in S_1 \times \ldots \times S_n$:

$$p_{X_1,\ldots,X_n}(a_1, \ldots, a_n) = \widetilde{c} \cdot \prod_{i=1}^n \widetilde{g}_i(a_i) \tag{2.12}$$

mit einer reellen Konstante $\widetilde{c}$. Da auf beiden Seiten von (2.12) bis auf den Faktor $\widetilde{c}$ die Massenfunktionen von Wahrscheinlichkeitsverteilungen stehen, gilt $\widetilde{c} = 1$. Also ist die gemeinsame Verteilung von $X_1, \ldots, X_n$ das Produkt der Verteilungen μ_i, und somit sind die Zufallsvariablen X_i unabhängig mit Verteilung μ_i, d.h. mit Massenfunktion $\widetilde{g}_i$. ■

Sei nun I eine beliebige Menge, und S_i sowie $\widetilde{S}_i$ $(i \in I)$ abzählbare Mengen. Sind X_i $(i \in I)$ unabhängige diskrete Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit Wertebereichen S_i, dann sind auch die Zufallsvariablen

$$Y_i(\omega) := h_i(X_i(\omega)) = (h_i \circ X_i)(\omega) \qquad (i \in I)$$

für beliebige Funktionen $h_i : S_i \to \widetilde{S}_i$ wieder unabhängig. Dies folgt unmittelbar aus der Definition der Unabhängigkeit, denn für beliebige Teilmengen $A_i \subset \widetilde{S}_i$ gilt $\{Y_i \in A_i\} = \left\{X_i \in h_i^{-1}(A_i)\right\}$, und diese Ereignisse sind unabhängig. Das folgende Korollar liefert eine wichtige Verschärfung dieser Aussage.

Korollar 2.16 (Gruppierungslemma). Sei X_i $(i \in I)$ eine Kollektion unabhängiger Zufallsvariablen, und seien $I_1, I_2, \ldots, I_n$ disjunkte Teilmengen von I. Dann sind auch die Zufallsvariablen

$$X_{I_1} = (X_i)_{i\in I_1},\ X_{I_2} = (X_i)_{i\in I_2},\ \ldots,\ X_{I_n} = (X_i)_{i\in I_n}$$

wieder unabhängig. Zudem sind $h_1(X_{I_1}), \ldots, h_n(X_{I_n})$ für beliebige Funktionen $h_1, \ldots, h_n$ wieder unabhängige Zufallsvariablen.

Wir können also die unabhängigen Zufallsvariablen in disjunkte Blöcke einteilen, und beliebige Funktionen betrachten, die jeweils nur von den Zufallsvariablen in einem Block abhängen. Diese Funktionen sind dann wieder unabhängige Zufallsvariablen. Beispielsweise sind bei sechs unabhängigen Würfelwürfen $X_1, X_2, \ldots, X_6$ die Augensummen $X_1 + X_2$, $X_3 + X_4$ und $X_5 + X_6$ voneinander unabhängig, ebenso die maximalen Augenzahlen $\max(X_1, X_2, X_3)$ und $\max(X_4, X_5, X_6)$ bei den ersten und den letzten drei Würfen.

Beweis. Seien $a_i \in S_i$ $(i \in I)$, und sei $a_{I_k} = (a_i)_{i\in I_k}$ $(k = 0, 1, \ldots, n)$. Dann gilt

$$\begin{aligned} P\left[X_{I_1} = a_{I_1}, \ldots, X_{I_n} = a_{I_n}\right] &= P\left[X_i = a_i \text{ für alle } i \in I_1 \cup \cdots \cup I_n\right] \\ &= P\left[\bigcap_{i\in I_1\cup\cdots\cup I_n} \{X_i = a_i\}\right] = \prod_{i\in I_1\cup\cdots\cup I_n} P\left[X_i = a_i\right] \\ &= \prod_{i\in I_1} P\left[X_i = a_i\right] \cdot \cdots \cdot \prod_{i\in I_n} P\left[X_i = a_i\right] \\ &= P\left[X_{I_1} = a_{I_1}\right] \cdot \cdots \cdot P\left[X_{I_n} = a_{I_n}\right]. \end{aligned}$$

Also ist die Massenfunktion der gemeinsamen Verteilung der Zufallsvariablen $X_{I_1}, \ldots, X_{I_n}$ das Produkt der einzelnen Massenfunktionen, d.h. die Zufallsvariablen sind unabhängig. Die Unabhängigkeit der Zufallsvariablen $h_1(X_{I_1}), \ldots, h_n(X_{I_n})$ folgt dann wie oben. ■

2.4 Summen von unabhängigen Zufallsvariablen

Wir berechnen nun Verteilungen und gemeinsame Verteilungen von Summen unabhängiger Zufallsvariablen.

Faltung von Wahrscheinlichkeitsverteilungen

Seien zunächst X und Y unabhängige diskrete Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ mit Werten im $\mathbb{R}^d$, Verteilungen μ bzw. ν und Massenfunktionen p bzw. q. Wir wollen die Verteilung von $X + Y$ bestimmen. Es gilt

$$P[X + Y = z] = \sum_{x\in X(\Omega)} \underbrace{P[X = x, Y = z - x]}_{=P[X=x]\cdot P[Y=z-x]} = \sum_{x\in X(\Omega)} p(x)q(z - x). \tag{2.13}$$

Die Verteilung von $X + Y$ ist also die Wahrscheinlichkeitsverteilung $\mu \star \nu$ mit Massenfunktion

$$(p \star q)(z) = \sum_{x\in X(\Omega)} p(x)q(z - x). \tag{2.14}$$

Diese Verteilung nennt man die **Faltung** der Wahrscheinlichkeitsverteilungen μ und ν.

Bemerkung (Eigenschaften der Faltung von Wahrscheinlichkeitsverteilungen). Die Faltung $\mu \star \nu$ zweier Wahrscheinlichkeitsverteilungen μ und ν auf $\mathbb{R}^d$ ist wieder eine Wahrscheinlichkeitsverteilung auf $\mathbb{R}^d$. Da die Addition von Zufallsvariablen kommutativ und assoziativ ist, hat die Faltung von Wahrscheinlichkeitsverteilungen dieselben Eigenschaften:

$$\begin{aligned} \mu \star \nu &= \nu \star \mu \qquad (\text{da } X + Y = Y + X), \\ (\mu \star \nu) \star \eta &= \mu \star (\nu \star \eta) \qquad (\text{da } (X + Y) + Z = X + (Y + Z)\,). \end{aligned}$$

Beispiele. (i) Sind X und Y unabhängig und $\mathrm{Bin}(n,p)$ bzw. $\mathrm{Bin}(m,p)$-verteilt, dann ist $X + Y$ eine $\mathrm{Bin}(n+m,p)$-verteilte Zufallsvariable. Zum Beweis bemerkt man, dass die gemeinsame Verteilung von X und Y mit der gemeinsamen Verteilung von $Z_1 + \ldots + Z_n$ und $Z_{n+1} + \ldots + Z_{n+m}$ übereinstimmt, wobei die Zufallsvariablen Z_i $(1 \le i \le n+m)$ unabhängig und Bernoulli(p)-verteilt sind. Also folgt:

$$\mu_{X+Y} = \mu_{Z_1+\ldots+Z_n+Z_{n+1}+\ldots+Z_{n+m}} = \mathrm{Bin}(n+m,p)\,.$$

Als Konsequenz erhalten wir (ohne zu rechnen):

$$\mathrm{Bin}(n,p) \star \mathrm{Bin}(m,p) = \mathrm{Bin}(n+m,p)\,,$$

d.h. die Binomialverteilungen bilden eine *Faltungshalbgruppe*. Explizit ergibt sich:

$$\sum_{k=0}^{l} \binom{n}{k} p^k (1-p)^{n-k} \binom{m}{l-k} p^{l-k} (1-p)^{m-(l-k)} = \binom{n+m}{l} p^l (1-p)^{n+m-l}\,, \quad \text{d.h.}$$

$$\sum_{k=0}^{l} \binom{n}{k}\binom{m}{l-k} = \binom{n+m}{l}. \tag{2.15}$$

Diese kombinatorische Formel ist auch als *Vandermonde-Identität* bekannt.

(ii) Sind X und Y unabhängig und Poisson-verteilt mit Parametern λ bzw. $\widetilde{\lambda}$, dann ist $X + Y$ Poisson-verteilt mit Parameter $\lambda + \widetilde{\lambda}$, denn nach der binomischen Formel gilt für $n \ge 0$:

$$\begin{aligned} (\mu_X \star \mu_Y)(n) &= \sum_{k=0}^{n} \mu_X(k) \cdot \mu_Y(n-k) \\ &= \sum_{k=0}^{n} \frac{\lambda^k}{k!}\, e^{-\lambda} \cdot \frac{\widetilde{\lambda}^{n-k}}{(n-k)!}\, e^{-\widetilde{\lambda}} \\ &= e^{-\lambda+\widetilde{\lambda}} \cdot \sum_{k=0}^{n} \frac{\lambda^k}{k!} \frac{\widetilde{\lambda}^{n-k}}{(n-k)!} \\ &= e^{-\lambda+\widetilde{\lambda}} \cdot \frac{(\lambda+\widetilde{\lambda})^n}{n!}\,. \end{aligned}$$

Also bilden auch die Poissonverteilungen eine Faltungshalbgruppe:

$$\mathrm{Poisson}(\lambda) \star \mathrm{Poisson}(\widetilde{\lambda}) = \mathrm{Poisson}(\lambda + \widetilde{\lambda}).$$

Irrfahrten auf $\mathbb{Z}$

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

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

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

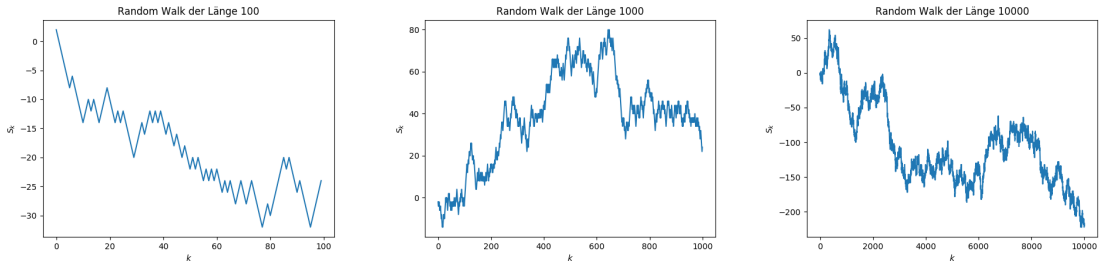
$$\begin{aligned} S_0 &= a, \\ S_{n+1} &= S_n + X_{n+1}, \end{aligned}$$

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

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

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

Beispiel (Symmetrische Irrfahrt, $p = 1/2$). Die folgenden Graphiken zeigen Simulationen der ersten 50, 500 bzw. 5000 Schritte eines Random Walks für $p = 1/2$.



(a) $n = 50$ (b) $n = 500$ (c) $n = 5000$

Wir wollen nun die Verteilung von verschiedenen, durch die Irrfahrt gegebenen, Zufallsvariablen berechnen. Die Verteilung von S_n selbst ist eine verzerrte Bimomialverteilung.

Lemma 2.17 (Verteilung von S_n). *Für $k \in \mathbb{Z}$ gilt*

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

Beweis. Es gilt

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

■

Sei $\lambda \in \mathbb{Z}$. Weiter unten werden wir (im Fall $p = 1/2$) die Verteilung der Zufallsvariable

$$T_\lambda(\omega) \;:=\; \min\{n \in \mathbb{N} \;:\; S_n(\omega) = \lambda\}$$

bestimmen, wobei wir wieder $\min \emptyset := \infty$ setzen. Für $\lambda \neq a$ ist T_λ die erste *Trefferzeit von* λ, für $\lambda = a$ ist es hingegen die erste *Rückkehrzeit nach* a. Beschreibt die Irrfahrt beispielsweise die Kapitalentwicklung in einem Glücksspiel, dann kann man T_0 als Ruinzeitpunkt interpretieren. Da das Ereignis

$$\{T_\lambda \leq n\} \;=\; \bigcup_{i=1}^{n} \{S_i = \lambda\}$$

von den Positionen der Irrfahrt zu *mehreren* Zeiten abhängt, benötigen wir die *gemeinsame* Verteilung der entsprechenden Zufallsvariablen. Sei dazu

$$S_{0:n}(\omega) \;:=\; (S_0(\omega), S_1(\omega), \ldots, S_n(\omega))$$

der *Bewegungsverlauf bis zur Zeit* n . Dann ist $S_{0:n}$ eine Zufallsvariable, die Werte im Raum

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

aller möglichen Verläufe (Pfade) der Irrfahrt annimmt. Sei μ_a die Verteilung von $S_{0:n}$ unter P.

Lemma 2.18. *Für* $(s_0, s_1, \ldots, s_n) \in \widehat{\Omega}_a^{(n)}$ *gilt*

$$\mu_a[\{(s_0, \ldots, s_n)\}] \;=\; p^{\frac{n+k}{2}} (1-p)^{\frac{n-k}{2}}, \qquad \text{wobei } \; k = s_n - s_0. \tag{2.16}$$

Insbesondere ist μ_a *im Fall* $p = 1/2$ *die Gleichverteilung auf dem Pfadraum* $\widehat{\Omega}_a^{(n)} \subseteq \mathbb{Z}^{n+1}$.

Beweis. Für $s_0, \ldots, s_n \in \mathbb{Z}$ gilt

$$\begin{aligned} \mu_a[\{(s_0, \ldots, s_n)\}] &= P[S_0 = s_0, \ldots, S_n = s_n] \\ &= P[S_0 = s_0, X_1 = s_1 - s_0, \ldots, X_n = s_n - s_{n-1}]. \end{aligned}$$

Diese Wahrscheinlichkeit ist gleich 0, falls $s_0 \neq a$ oder $|s_i - s_{i-1}| \neq 1$ für ein $i \in \{1, \ldots, n\}$ gilt. Andernfalls, d.h. für $(s_0, \ldots, s_n) \in \widehat{\Omega}_a^{(n)}$, gilt (2.16), da für $s_n - s_0 = k$ genau $\frac{n+k}{2}$ der Inkremente $s_1 - s_0, \ldots, s_n - s_{n-1}$ gleich $+1$ und die übrigen gleich -1 sind. ■

Symmetrische Irrfahrt und Reflektionsprinzip

Ab jetzt betrachten wir nur noch die symmetrische Irrfahrt mit $p = \frac{1}{2}$. Lemma 2.18 ermöglicht es uns, Wahrscheinlichkeiten für die symmetrische Irrfahrt durch Abzählen zu berechnen. Dazu zeigen wir eine nützliche Invarianzeigenschaft bezüglich der Reflektion der Pfade beim ersten Erreichen eines Levels λ. Den Beweis des folgenden Satzes macht man sich am besten zunächst anhand von Abbildung 2.4 klar.

Satz 2.19 (Reflektionsprinzip). Seien $\lambda, b \in \mathbb{Z}$. Es gelte entweder ($a < \lambda$ und $b \leq \lambda$), oder ($a > \lambda$ und $b \geq \lambda$). Dann folgt

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

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

Beweis. Es gilt

$$\begin{aligned} P[T_\lambda \leq n, S_n = b] &= \mu_a[\overbrace{\{(s_0, \ldots, s_n) \;:\; s_n = b, s_i = \lambda \text{ für ein } i \in \{1, \ldots, n\}\}}^{=:A}], \\ P[S_n = b^\star] &= \mu_a[\underbrace{\{(s_0, \ldots, s_n) \;:\; s_n = b^\star\}}_{=:B}]. \end{aligned}$$

Die in Abbildung 2.4 dargestellte Transformation (Reflektion des Pfades nach Treffen von λ) definiert eine Bijektion von A nach B. Also gilt $|A| = |B|$. Da μ_a die Gleichverteilung auf $\widehat{\Omega}_a^{(n)}$ ist, folgt

$$\mu_a[A] \;=\; \frac{|A|}{\left|\widehat{\Omega}_a^{(n)}\right|} \;=\; \frac{|B|}{\left|\widehat{\Omega}_a^{(n)}\right|} \;=\; \mu_a[B],$$

und damit die Behauptung. ■

Mithilfe des Reflektionsprinzips können wir nun die Verteilung der ersten Trefferzeiten explizit aus den uns schon bekannten Verteilungen der Zufallsvariablen S_n berechnen.

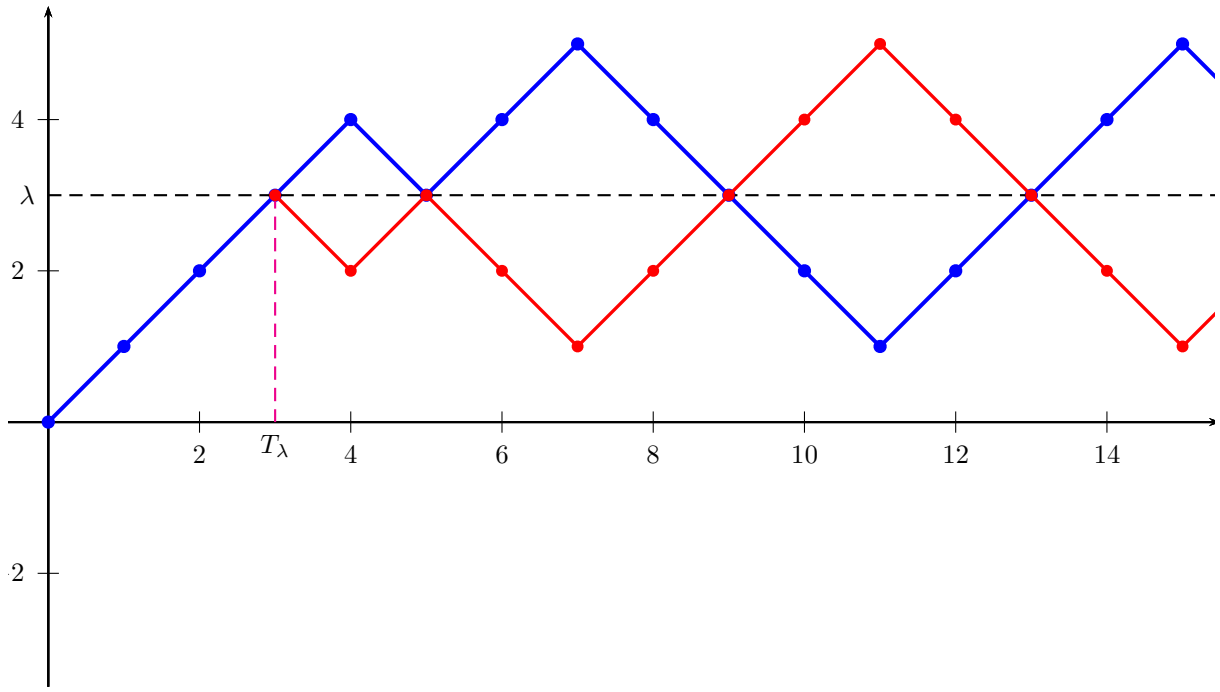


Abbildung 2.4: Reflektionsprinzip

Korollar 2.20 (Verteilung der Trefferzeiten). Für $\lambda \in \mathbb{Z}$ und $n \in \mathbb{N}$ gilt:

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

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

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

(i) Ist $S_n \geq \lambda$, dann gilt stets $T_\lambda \leq n$. Daher folgt nach Satz 2.19:

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

(ii) Aus (i) folgt

$$\begin{aligned} P[T_\lambda = n] &= P[T_\lambda \le n] - P[T_\lambda \le n-1] \\ &= \underbrace{P[S_n \ge \lambda] - P[S_{n-1} \ge \lambda]}_{=:\mathbf{I}} + \underbrace{P[S_n \ge \lambda+1] - P[S_{n-1} \ge \lambda+1]}_{=:\mathbf{II}} \end{aligned}$$

Wegen

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

erhalten wir für den ersten Term:

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

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

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

■

Aus der Verteilung der Trefferzeiten T_λ ergibt sich auch unmittelbar die Verteilung des Maximums

$$M_n := \max(S_0, S_1, \ldots, S_n)$$

der Irrfahrt bis zur Zeit n.

Korollar 2.21 (Verteilung des Maximums). Für $\lambda > a$ gilt

$$P[M_n \ge \lambda] = P[T_\lambda \le n] = P[S_n \ge \lambda] + P[S_n > \lambda].$$

3 Gesetze der großen Zahlen

In diesem Kapitel beweisen wir zwei ganz unterschiedliche Arten von Konvergenzaussagen für Folgen von Zufallsvariablen bzw. deren Verteilungen: zum einen Gesetze der großen Zahlen für relative Häufigkeiten von unabhängigen Ereignissen, und allgemeiner für Mittelwerte von schwach korrelierten Zufallsvariablen, zum anderen die Konvergenz ins Gleichgewicht der Verteilungen irreduzibler, aperiodischer Markovketten mit endlichem Zustandsraum. Beide Aussagen lassen sich auch zu einem Gesetz der großen Zahlen für Markovketten kombinieren.

3.1 Gesetz der großen Zahlen für unabhängige Ereignisse

Das empirische Gesetz der großen Zahlen (GGZ) besagt, dass sich die relative Häufigkeit für das Eintreten von gleich wahrscheinlichen unabhängigen Ereignissen $A_1, \dots, A_n$ für $n \to \infty$ der Erfolgswahrscheinlichkeit p annähert. Wir können diese Aussage nun mathematisch präzisieren, und aus den Kolmogorovschen Axiomen herleiten. Je nach Präzisierung des Konvergenzbegriffs unterscheidet man zwischen dem schwachen und dem starken Gesetz der großen Zahlen.

Bernstein-Ungleichung und schwaches Gesetz der großen Zahlen

Sei $A_1, A_2, \dots$ eine Folge unabhängiger Ereignisse auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit fester Wahrscheinlichkeit $P[A_i] = p \in [0,1]$, und sei

$$S_n(\omega) \;=\; |\{1 \le i \le n \;:\; \omega \in A_i\}| \;=\; \sum_{i=1}^{n} I_{A_i}(\omega)$$

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

Satz 3.1 (Bernstein-Ungleichung, Schwaches GGZ für unabhängige Ereignisse).
für alle $\varepsilon > 0$ und $n \in \mathbb{N}$ gilt

$$P\left[\frac{S_n}{n} \ge p + \varepsilon\right] \;\le\; e^{-2\varepsilon^2 n}, \qquad \text{und} \qquad P\left[\frac{S_n}{n} \le p - \varepsilon\right] \;\le\; e^{-2\varepsilon^2 n}.$$

Insbesondere ist

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

d.h. die Wahrscheinlichkeit für eine Abweichung der relativen Häufigkeit S_n/n von der Wahrscheinlichkeit p um mehr als ε fällt exponentiell schnell in n ab.

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

b) Die Aussage kann man zum empirischen *Schätzen der Wahrscheinlichkeit* p verwenden: für große n gilt

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

Simuliert man die Stichproben künstlich auf dem Computer, dann ergibt sich ein *Monte-Carlo-Verfahren* zur näherungsweisen Berechnung von p. Der Satz liefert eine recht präzise Fehlerabschätzung für den Schätz- bzw. Approximationsfehler.

c) Bemerkenswert ist, dass die Abschätzung aus der Bernstein-Ungleichung nicht nur asymptotisch für $n \to \infty$, sondern für jedes feste n gilt. Solche präzisen *nicht-asymptotischen Abschätzungen* sind für Anwendungen sehr wichtig, und oft nicht einfach herzuleiten.

Beweis. Der Beweis von Satz 3.1 besteht aus zwei Teilen: Wir leiten zunächst exponentielle Abschätzungen für die Wahrscheinlichkeiten her, welche von einem Parameter $\lambda \geq 0$ abhängen. Anschließend optimieren wir die erhaltene Abschätzung durch Wahl von λ.

Wir setzen $q := 1 - p$. Wegen $S_n \sim \mathrm{Bin}(n, p)$ gilt für $\lambda \geq 0$:

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

Wir werden unten zeigen, dass für alle $\lambda \geq 0$ die Abschätzung

$$p\, e^{\lambda q} + q\, e^{-\lambda p} \;\leq\; e^{\lambda^2/8} \tag{3.1}$$

gilt. Damit erhalten wir dann

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

Der Exponent auf der rechten Seite ist minimal für $\lambda = 4\varepsilon$. Mit dieser Wahl von λ folgt schließlich

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

Die Abschätzung für $P[S_n \leq n\,(p-\varepsilon)]$ zeigt man analog, und erhält so die Aussage des Satzes.

Nachzutragen bleibt nur noch der Beweis der Abschätzung (3.1). Sei dazu

$$f(\lambda) := \log\left(p\, e^{\lambda q} + q\, e^{-\lambda p}\right) = \log\left(e^{-\lambda p}\,(p\, e^{\lambda} + q)\right) = -\lambda\, p + \log\left(p\, e^{\lambda} + q\right).$$

Zu zeigen ist $f(\lambda) \leq \lambda^2/8$ für alle $\lambda \geq 0$. Es gilt $f(0) = 0$,

$$\begin{aligned} f'(\lambda) &= -p + \frac{p\, e^{\lambda}}{p\, e^{\lambda} + q} = -p + \frac{p}{p + q\, e^{-\lambda}}, \qquad f'(0) = 0, \\ f''(\lambda) &= \frac{p\, q\, e^{-\lambda}}{(p + q\, e^{-\lambda})^2} \;\leq\; \frac{1}{4}. \end{aligned}$$

Hierbei haben wir im letzten Schritt die elementare Ungleichung

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

benutzt. Damit folgt für $\lambda \geq 0$ wie behauptet

$$f(\lambda) \;=\; \int_0^{\lambda} f'(x)\,dx \;=\; \int_0^{\lambda} \int_0^{x} f''(y)\,dy\,dx \;\leq\; \int_0^{\lambda} \frac{x}{4}\,dx \;\leq\; \frac{\lambda^2}{8}. \qquad \blacksquare$$

Zur Illustration des Satzes simulieren wir den Verlauf von S_k und S_k/k für $k \leq n$ und $p = 0.7$ mehrfach (30 mal, siehe Abbildung 3.1 und 3.2), und plotten die Massenfunktionen von S_n, Abbildung 3.3.

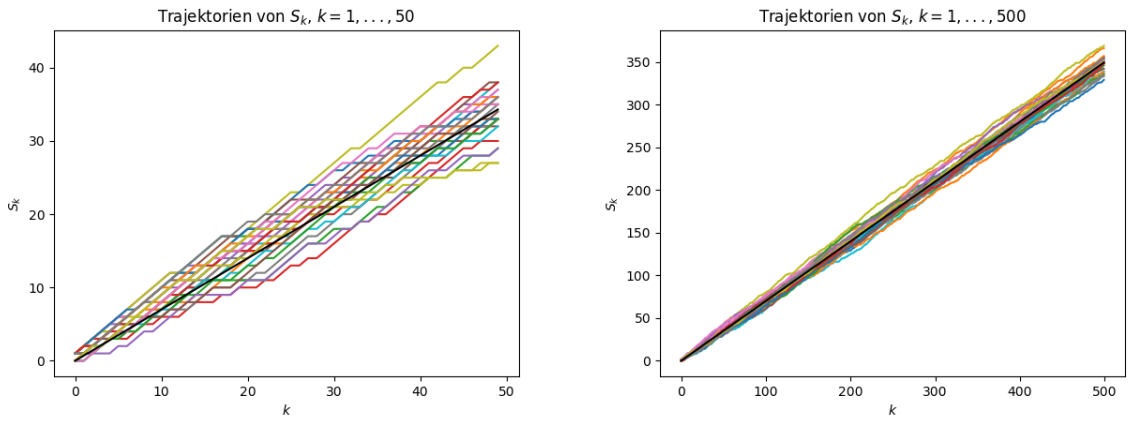


Abbildung 3.1: Verlauf von S_k für $k \leq 50$ bzw. $k \leq 500$

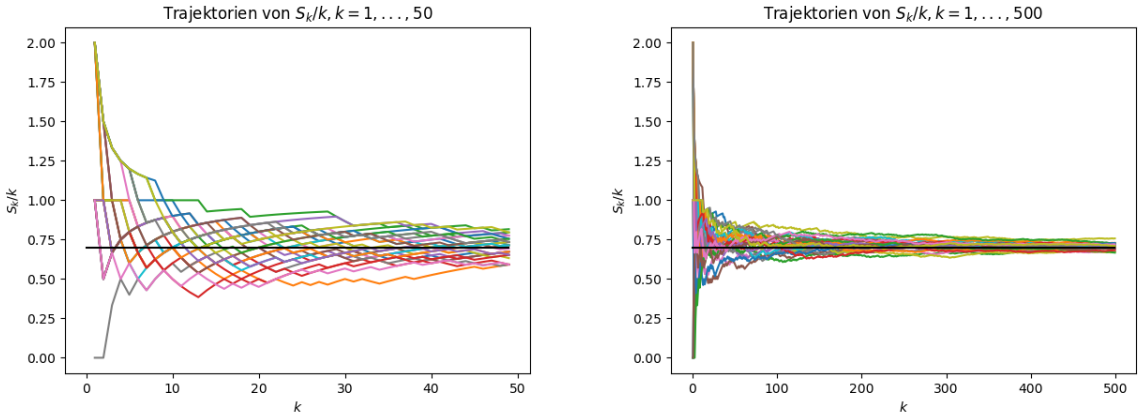


Abbildung 3.2: Verlauf von S_k/k für $k \leq 50$ bzw. $k \leq 500$.

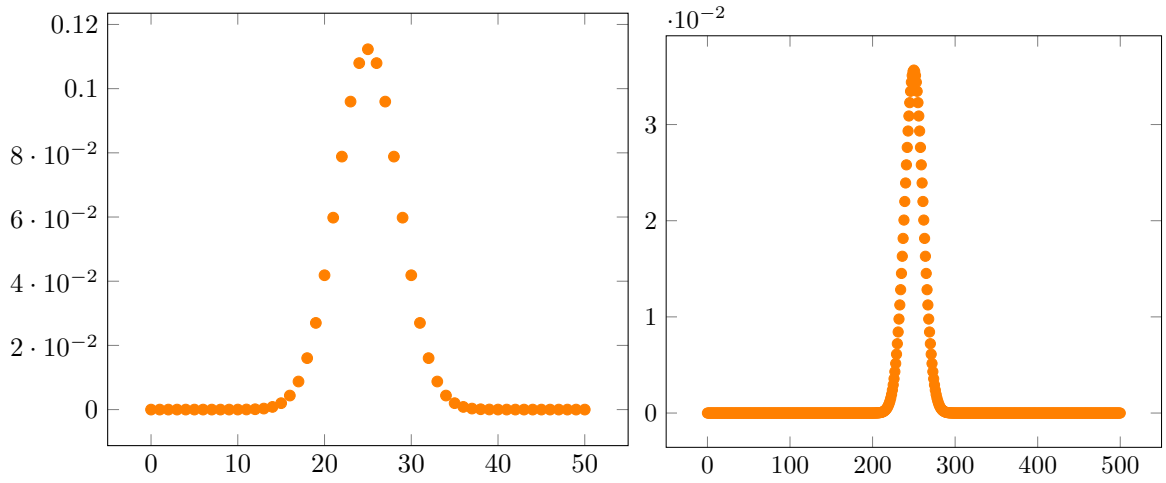


Abbildung 3.3: Massenfunktion von S_{50} bzw. S_{500}.

Beispiel (Konfidenzintervalle bei Wahlumfragen). Sei p der Stimmenanteil einer Partei in der Gesamtheit aller Wähler. Um p zu schätzen befragen wir n Wähler. Der Einfachheit halber nehmen wir an, dass diese unabhängig voneinander und rein zufällig aus der Gesamtheit aller Wähler ausgewählt werden. Sei A_i das Ereignis, dass der i-te Wähler in unserer Zufallsstichprobe die Partei wählt. Dann sind die Ereignisse $A_1, \dots, A_n$ unabhängig mit Wahrscheinlichkeit p, und $\hat{p}_n := S_n/n$ ist der Stimmenanteil der Partei in unserer Stichprobe. Nach der Bernstein-Ungleichung gilt also

$$P[|\hat{p}_n - p| \geq \varepsilon] \;\leq\; 2e^{-2\varepsilon^2 n}.$$

Ist zum Beispiel $\varepsilon = 2\%$ und $n = 5000$, dann ist die Wahrscheinlichkeit kleiner als $0,04$, d.h. für den gesuchten Stimmenanteil p gilt

$$P\left[p \in (\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon)\right] \;\geq\; 0,96\,.$$

In diesem Fall nennt man das Intervall $(\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon)$ ein *Konfidenzintervall (Vertrauensintervall) zum Niveau* 96% für den gesuchten Wert p. In der Praxis ist es natürlich nicht möglich, die Stichprobe so zu wählen, dass jeder Wähler mit der gleichen Wahrscheinlichkeit befragt wird, und die Auswertung einer Befragung ist daher wesentlich komplizierter.

Starkes Gesetz der großen Zahlen für unabhängige Ereignisse

Wir zeigen nun, dass aus der Bernstein-Ungleichung auch ein *starkes Gesetz der großen Zahlen* für die relativen Häufigkeiten folgt. Dieses besagt, dass die Zufallsfolge S_n/n mit Wahrscheinlichkeit 1 für $n \to \infty$ gegen p konvergiert. Wir bemerken zunächst, dass $\{\lim S_n/n = p\}$ ein Ereignis in der σ-Algebra $\mathcal{A}$ ist, denn es gilt

$$\lim_{n\to\infty} \frac{S_n(\omega)}{n} = p \quad\Leftrightarrow\quad \forall\, k \in \mathbb{N}\ \exists\, n_0 \in \mathbb{N}\ \forall\, n \geq n_0 \,:\, \left|\frac{S_n(\omega)}{n} - p\right| \leq \frac{1}{k},$$

und damit

$$\left\{\lim_{n\to\infty} \frac{S_n}{n} = p\right\} \;=\; \bigcap_{k=1}^{\infty} \bigcup_{n_0=1}^{\infty} \bigcap_{n=n_0}^{\infty} \left\{\left|\frac{S_n}{n} - p\right| \leq \frac{1}{k}\right\} \;\in\; \mathcal{A}. \tag{3.2}$$

Korollar 3.2 (Starkes GGZ für unabhängige Ereignisse). Es gilt

$$P\left[\lim_{n\to\infty} \frac{S_n}{n} = p\right] \;=\; 1.$$

Ein schwaches Gesetz der großen Zahlen für unabhängige Ereignisse wurde bereits 1689 von Jakob Bernoulli formuliert und bewiesen. Der erste Beweis eines starken Gesetzes der großen Zahlen wurde dagegen erst zu Beginn des 20. Jahrhunderts von Borel, Hausdorff und Cantelli gegeben.

Beweis. Wir zeigen mithilfe der Bernstein-Ungleichung, dass das Gegenereignis $\{S_n/n \not\to p\}$ Wahrscheinlichkeit Null hat. Nach (3.2) gilt

$$\left\{\lim \frac{S_n}{n} \neq p\right\} \;=\; \bigcup_{k=1}^{\infty} A_k \qquad \text{mit} \qquad A_k \;=\; \bigcap_{n_0=1}^{\infty} \bigcup_{n=n_0}^{\infty} \left\{\left|\frac{S_n}{n} - p\right| > \frac{1}{k}\right\}.$$

Es genügt also $P[A_k] = 0$ für jedes $k \in \mathbb{N}$ zu zeigen. Sei dazu $k \in \mathbb{N}$ fest gewählt. Aus der Bernstein-Ungleichung folgt für $n_0 \in \mathbb{N}$:

$$P[A_k] \;\leq\; P\left[\bigcup_{n=n_0}^{\infty} \left\{\left|\frac{S_n}{n} - p\right| > \frac{1}{k}\right\}\right] \;\leq\; \sum_{n=n_0}^{\infty} 2e^{-2n/k^2}.$$

Wegen $\sum_{n=1}^{\infty} e^{-2n/k^2} < \infty$ konvergieren die Partialsummen auf der rechten Seite für $n_0 \to \infty$ gegen Null. Also folgt $P[A_k] = 0$, und damit die Behauptung. ■

Im Beweis haben wir die folgende Aussage benutzt, die aus den Kolmogorovschen Axiomen folgt.

Lemma 3.3 (σ-Subadditivität). *Für beliebige Ereignisse $A_1, A_2, \ldots \in \mathcal{A}$ gilt*

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

Beweis. Die Mengen $B_n := A_n \setminus (A_{n-1} \cup \cdots \cup A_1)$ sind disjunkt mit $\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$. Wegen $B_n \subseteq A_n$ erhalten wir $P\left[\bigcup_{n=1}^{\infty} A_n\right] = P\left[\bigcup_{n=1}^{\infty} B_n\right] = \sum_{n=1}^{\infty} P[B_n] \leq \sum_{n=1}^{\infty} P[A_n]$. ■

Beispiel (Irrfahrt auf $\mathbb{Z}$). Wir betrachten einen Random Walk

$$Z_n = X_1 + X_2 + X_3 + \ldots + X_n \quad (n \in \mathbb{N})$$

mit unabhängigen identisch verteilten Inkrementen $X_i, i \in \mathbb{N}$, mit

$$P[X_i = 1] = p \quad \text{und} \quad P[X_i = -1] = 1 - p, \quad p \in (0,1) \text{ fest.}$$

Die Ereignisse $A_i := \{X_i = 1\}$ sind unabhängig mit $P[A_i] = p$ und es gilt:

$$X_i = I_{A_i} - I_{A_i^C} = 2I_{A_i} - 1,$$

also

$$Z_n = 2S_n - n, \quad \text{wobei} \quad S_n = \sum_{i=1}^{n} I_{A_i}.$$

Nach Korollar 3.2 folgt

$$\lim_{n\to\infty} \frac{Z_n}{n} = 2 \lim_{n\to\infty} \frac{S_n}{n} - 1 = 2p - 1 \qquad P\text{-fast sicher (d.h. mit Wahrscheinlichkeit 1).}$$

Für $p \neq \frac{1}{2}$ wächst (bzw. fällt) Z_n mit Wahrscheinlichkeit 1 asymptotisch linear (siehe Abbildung 3.4), d.h. für $n \to \infty$ gilt

$$Z_n \sim (2p-1) \cdot n \quad P\text{-fast sicher.}$$

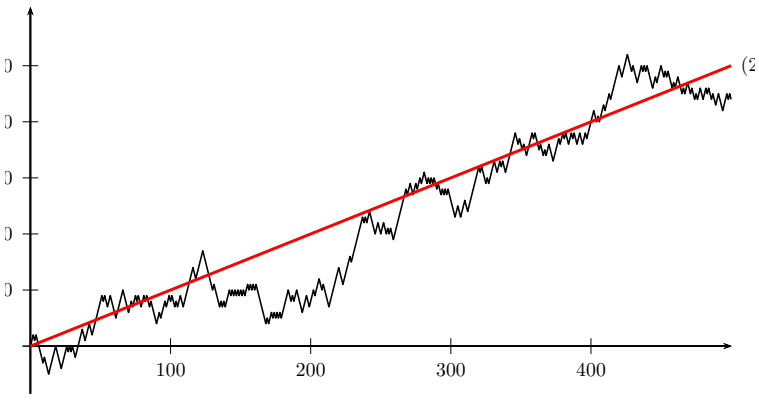


Abbildung 3.4: Random Walk mit Drift: $p = 0.55, n = 500$

Für $p = \frac{1}{2}$ dagegen wächst der Random Walk sublinear, d.h. $\frac{Z_n}{n} \to 0$ P-fast sicher. In diesem Fall liegt für hinreichend große n der Graph einer typischen Trajektorie $Z_n(\omega)$ in einem beliebig kleinen Sektor um die x-Achse (siehe Abbildung 3.5).

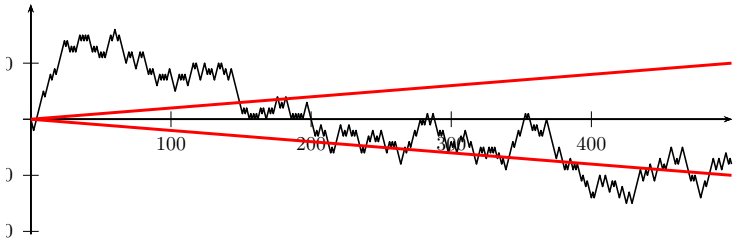


Abbildung 3.5: Random Walk ohne Drift: $p = 0.5, n = 500$

3.2 Varianz und Kovarianz

Im nächsten Abschnitt werden wir ein Gesetz der großen Zahlen für schwach korrelierte Zufallsvariablen beweisen. Als Vorbereitung führen wir in diesem Abschnitt die Begriffe der Varianz und Standardabweichung, sowie Kovarianz und Korrelation reellwertiger Zufallsvariablen ein, und beweisen zwei wichtige Ungleichungen.

Varianz und Standardabweichung

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $X : \Omega \to S \subseteq \mathbb{R}$ eine reellwertige Zufallsvariable auf $(\Omega, \mathcal{A}, P)$ mit abzählbarem Wertebereich S. Wir setzen voraus, dass $E[|X|]$ endlich ist.

Definition 3.4. Die **Varianz** von X ist definiert als mittlere quadratische Abweichung vom Erwartungswert, d.h.

$$\mathrm{Var}[X] \;=\; E\left[(X - E[X])^2\right] \;\in\; [0, \infty].$$

Die Größe $\sigma[X] \;=\; \sqrt{\mathrm{Var}[X]}$ heißt **Standardabweichung** von X.

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

Bemerkung (Eigenschaften der Varianz). a) Die Varianz einer Zufallsvariable hängt nur von ihrer Verteilung ab. Es gilt

$$\mathrm{Var}[X] \;=\; \sum_{a \in S} (a - m)^2 \, p_X(a),$$

wobei $m := E[X] = \sum_{a \in S} a \, p_X(a)$ der Erwartungswert von X ist.

b) Aus der Linearität des Erwartungswerts folgt

$$\mathrm{Var}[X] \;=\; E\left[X^2 - 2X \cdot E[X] + E[X]^2\right] \;=\; E\left[X^2\right] - E[X]^2.$$

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

c) Entsprechend folgt aus der Linearität des Erwartungswerts

$$\mathrm{Var}[aX + b] \;=\; \mathrm{Var}[aX] \;=\; a^2 \,\mathrm{Var}[X] \qquad \text{für alle } a, b \in \mathbb{R}.$$

d) Die Varianz von X ist genau dann gleich 0, wenn X *deterministisch* ist, d.h. falls

$$P[X = E[X]] \;=\; 1.$$

Beispiele. a) Varianz von Bernoulli-Verteilungen: Sei $X = 1$ mit Wahrscheinlichkeit p, und $X = 0$ mit Wahrscheinlichkeit $1 - p$. Dann gilt $E\left[X^2\right] = E[X] = p$, und damit

$$\mathrm{Var}[X] \;=\; p - p^2 \;=\; p\,(1-p).$$

b) Varianz von geometrischen Verteilungen: Sei T geometrisch verteilt mit Parameter $p \in (0,1]$. Dann gilt $P[T = k] = (1-p)^{k-1}p$ für alle $k \in \mathbb{N}$. Durch zweimaliges Differenzieren der Identität $\sum_{k=0}^{\infty}(1-p)^k = 1/p$ erhalten wir

$$E[T] \;=\; \sum_{k=1}^{\infty} k\,(1-p)^{k-1}\,p \;=\; -p\,\frac{d}{dp}\,\frac{1}{p} \;=\; \frac{1}{p}, \qquad \text{sowie}$$

$$E[(T+1)\,T] \;=\; \sum_{k=1}^{\infty}(k+1)\,k\,(1-p)^{k-1}\,p \;=\; \sum_{k=2}^{\infty} k\,(k-1)\,(1-p)^{k-2}\,p \;=\; p\frac{d^2}{dp^2}\frac{1}{p} \;=\; \frac{2}{p^2}.$$

Damit ergibt sich $E[T^2] = \frac{2}{p^2} - \frac{1}{p}$, und somit

$$\mathrm{Var}[T] \;=\; E\left[T^2\right] - E[T]^2 \;=\; \frac{1}{p^2} - \frac{1}{p} \;=\; \frac{1-p}{p^2}.$$

Im folgenden bezeichnen wir mit $\mathcal{L}^p(\Omega, \mathcal{A}, P)$ für $p \in [1, \infty)$ den Raum aller (diskreten) Zufallsvariablen $X : \Omega \to \mathbb{R}$ mit $E\left[|X|^p\right] < \infty$. Dieser Raum ist ein Vektorraum. Ist der Wahrscheinlichkeitsraum fest vorgegeben, dann schreiben wir auch kurz $\mathcal{L}^p$ statt $\mathcal{L}^p(\Omega, \mathcal{A}, P)$. Die Zufallsvariablen aus $\mathcal{L}^1(\Omega, \mathcal{A}, P)$ haben einen endlichen Erwartungswert. Gilt $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$, dann ist auch die Varianz von X endlich.

Die folgende wichtige Ungleichung spielt unter anderem im Beweis des Gesetzes der großen Zahlen im nächsten Abschnitt eine zentrale Rolle.

Satz 3.5 (Čebyšev-Ungleichung). Für $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ und $c > 0$ gilt:

$$P[|X - E[X]| \geq c] \;\leq\; \frac{1}{c^2}\,\mathrm{Var}[X].$$

Beweis. Es gilt

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

denn der Term auf der rechten Seite ist nichtnegativ und ≥ 1 auf $\{|X - E[X]| \geq c\}$. Durch Bilden des Erwartungswerts folgt

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

und damit die Behauptung. ∎

Kovarianz und Korrelation

Für Zufallsvariablen $X, Y \in \mathcal{L}^2$ können wir die Kovarianz und die Korrelation definieren.

Definition 3.6. Seien X und Y Zufallsvariablen in $\mathcal{L}^2(\Omega, \mathcal{A}, P)$.

(i) Die **Kovarianz** von X und Y ist definiert als

$$\mathrm{Cov}[X,Y] \;=\; E[(X - E[X])\,(Y - E[Y])] \;=\; E[XY] - E[X]\,E[Y].$$

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

$$\varrho[X,Y] \;=\; \frac{\mathrm{Cov}[X,Y]}{\sigma[X]\,\sigma[Y]}$$

Korrelationskoeffizient von X und Y.

(iii) Die Zufallsvariablen X und Y heißen **unkorreliert**, falls $\mathrm{Cov}[X,Y] = 0$, d.h. falls

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

Gilt $\mathrm{Cov}[X,Y] > 0$ bzw. < 0, dann heißen X und Y **positiv** bzw. **negativ korreliert**.

Satz 3.7 (Cauchy-Schwarz-Ungleichung für Kovarianz).

(i) Die Kovarianz ist eine symmetrische und bilineare Abbildung von $\mathcal{L}^2 \times \mathcal{L}^2$ nach $\mathbb{R}$ mit

$$\mathrm{Cov}[X,X] \;=\; \mathrm{Var}[X] \;\geq\; 0 \qquad \text{für alle } X \in \mathcal{L}^2.$$

(ii) Für $X, Y \in \mathcal{L}^2$ gilt die *Cauchy-Schwarz-Ungleichung*

$$|\mathrm{Cov}[X,Y]| \;\leq\; \sqrt{\mathrm{Var}[X]} \cdot \sqrt{\mathrm{Var}[Y]} \;=\; \sigma[X] \cdot \sigma[Y]. \tag{3.3}$$

Insbesondere gilt für den Korrelationskoeffizienten im Fall $\sigma[X] \cdot \sigma[Y] \neq 0$ stets

$$|\varrho[X,Y]| \quad\leq\quad 1. \tag{3.4}$$

(iii) Gleichheit gilt in den Ungleichungen (3.3) bzw. (3.4) genau dann, wenn Konstanten $a, b \in \mathbb{R}$ existieren, sodass

$$Y \;=\; aX + b \qquad \text{mit Wahrscheinlichkeit 1.}$$

In diesem Fall ist $\varrho[X,Y] = 1$ falls $a > 0$, und $\varrho[X,Y] = -1$ falls $a < 0$.

Beweis. Nach Definition gilt $\mathrm{Cov}[X,Y] = \mathrm{Cov}[Y,X]$ und $\mathrm{Cov}[X,X] = \mathrm{Var}[X]$. Außerdem folgt aus der Linearität des Erwartungswerts für $X, Y, Z \in \mathcal{L}^2$ und $a \in \mathbb{R}$:

$$\mathrm{Cov}[X, aY + Z] = E\left[(X - E[X])(aY + Z - E[aY + Z])\right] = a\,\mathrm{Cov}[X,Y] + \mathrm{Cov}[X,Z].$$

Somit ist die Kovarianz linear in der zweiten Komponente und damit wegen der Symmetrie auch bilinear. Cov ist also eine nicht-negative definite symmetrische Bilinearform auf dem Vektorraum $\mathcal{L}^2$. Damit gilt insbesondere die Cauchy-Schwarz-Ungleichung, siehe die Vorlesung Lineare Algebra. Den letzten Teil der Aussage und auch die Cauchy-Schwarz-Ungleichung werden wir gleich nebenbei im Rahmen eines Exkurses zu linearen Prognosen beweisen. ■

Die Bilinearität und Symmetrie der Kovarianz können wir benutzen, um die Varianz von Summen von Zufallsvariablen zu berechnen. Zum Beispiel erhalten wir für $X, Y \in \mathcal{L}^2$:

$$\begin{aligned} \mathrm{Var}[X+Y] &= \mathrm{Cov}[X+Y, X+Y] = \mathrm{Cov}[X,X] + 2\,\mathrm{Cov}[X,Y] + \mathrm{Cov}[Y,Y] \\ &= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}[X,Y]. \end{aligned}$$

Der Kovarianzterm ist gleich 0 falls X und Y unkorreliert sind. Dies ist insbesondere für unabhängige Zufallsvariablen der Fall, denn für diese gilt

$$\mathrm{Cov}[X,Y] = E[X \cdot Y] - E[X] \cdot E[Y] = 0.$$

Allgemeiner gilt sogar:

Satz 3.8 (Zusammenhang von Unabhängigkeit und Unkorreliertheit). Seien $X : \Omega \to S$ und $Y : \Omega \to T$ diskrete Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$. Dann sind äquivalent:

(i) X und Y sind unabhängig.

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

Bemerkung. Nach Satz 2.14 ist Bedingung (i) äquivalent zu

$$P[X = a, Y = b] = P[X = a]\, P[Y = b] \qquad \text{für alle } a \in S \text{ und } b \in T.$$

Entsprechend ist Bedingung (ii) genau dann erfüllt, wenn

$$E[f(X)\, g(Y)] = E[f(X)]\, E[g(Y)] \qquad \text{für alle } f, g : S \to \mathbb{R} \text{ mit } f(X), g(Y) \in \mathcal{L}^2 \text{ gilt.}$$

Beweis. (i)⇒(ii): Sind X und Y unabhängig, und $f(X), g(Y) \in \mathcal{L}^2$, dann folgt

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

(ii)⇒(i): Durch Wahl von $f = I_{\{a\}}$ und $g = I_{\{b\}}$ folgt aus (ii) für $a \in S$ und $b \in T$:

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

■

Das folgende einfache Beispiel zeigt, dass allein aus der Unkorreliertheit zweier Zufallsvariablen X und Y nicht deren Unabhängigkeit folgt.

Beispiel (Unkorreliertheit ohne Unabhängigkeit). Sei $X = +1$, 0, bzw. -1, jeweils mit Wahrscheinlichkeit $1/3$, und sei $Y = X^2$. Dann sind X und Y nicht unabhängig, aber unkorreliert, denn

$$\begin{aligned} P[X = 0, Y = 0] &= 1/3 \neq 1/9 = P[X = 0]\, P[Y = 0], \\ E[XY] &= 0 = E[X]\, E[Y]. \end{aligned}$$

Lineare Prognosen und Regressionsgeraden

Angenommen, wir wollen den Ausgang eines Zufallsexperiments vorhersagen, dass durch eine reellwertige Zufallsvariable $Y : \Omega \to \mathbb{R}$ beschrieben wird. Welches ist der *beste Prognosewert* b für $Y(\omega)$, wenn uns keine weiteren Informationen zur Verfügung stehen?

Die Antwort hängt offensichtlich davon ab, wie wir den Prognosefehler messen. Häufig verwendet man den mittleren quadratischen Fehler (*M*ean *S*quare *E*rror)

$$\text{MSE} \quad = \quad E[(Y-b)^2]\,.$$

Satz 3.9 (Erwartungswert als bester Prognosewert im quadratischen Mittel). Ist Y eine Zufallsvariable in $L^2(\Omega, \mathcal{A}, P)$, dann gilt für alle $b \in \mathbb{R}$:

$$E[(Y-b)^2] \quad = \quad \text{Var}[Y] + (b - E[Y])^2 \quad \geq \quad E[(Y-E[Y])^2].$$

Der mittlere quadratische Fehler des Prognosewertes b ist also die Summe der Varianz von Y und des Quadrats des systematischen bzw. mittleren Prognosefehlers (engl. *Bias*) $b - E[Y]$:

$$\text{MSE} \quad = \quad \text{Varianz} \; + \; \text{Bias}^2.$$

Insbesondere ist der mittlere quadratische Fehler genau für $b = E[Y]$ minimal.

Beweis. Für $b \in \mathbb{R}$ gilt wegen der Linearität des Erwartungswertes:

$$E[(Y-b)^2] \quad = \quad \text{Var}[Y-b] \; + \; E[Y-b]^2 \; = \; \text{Var}[Y] + (E[Y]-b)^2. \qquad \blacksquare$$

Seien nun $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ quadratintegrierbare Zufallsvariablen mit $\sigma[X] \neq 0$. Angenommen, wir kennen bereits den Wert $X(\omega)$ in einem Zufallsexperiment und suchen die beste *lineare* Vorhersage

$$\hat{Y}(\omega) \quad = \quad aX(\omega) + b, \qquad (a, b \in \mathbb{R}) \tag{3.5}$$

für $Y(\omega)$ im quadratischen Mittel. Zu minimieren ist jetzt der mittlere quadratischen Fehler

$$\text{MSE} \quad := \quad E[(\hat{Y} - Y)^2]$$

unter allen Zufallsvariablen $\hat{Y}$, die affine Funktionen von X sind. In diesem Fall erhalten wir

$$\text{MSE} \quad = \quad \text{Var}[Y - \hat{Y}] + E[Y-\hat{Y}]^2 \; = \; \text{Var}[Y - aX] + (E[Y] - aE[X] - b)^2.$$

Den zweiten Term können wir für gegebenes a minimieren, indem wir

$$b \quad = \quad E[Y] - aE[X]$$

wählen. Für den ersten Term ergibt sich

$$\begin{aligned} \text{Var}[Y-aX] \quad &= \quad \text{Cov}[Y-aX, Y-aX] \; = \; \text{Var}[Y] - 2a\,\text{Cov}[X,Y] + a^2\,\text{Var}[X] \\ &= \quad \left(a \cdot \sigma[X] - \frac{\text{Cov}[X,Y]}{\sigma[X]}\right)^2 + \text{Var}[Y] - \frac{\text{Cov}[X,Y]^2}{\text{Var}[X]}. \end{aligned} \tag{3.6}$$

Dieser Ausdruck wird minimiert, wenn wir $a = \text{Cov}[X,Y]/\sigma[X]^2$ wählen. Die bzgl. des mittleren quadratischen Fehlers optimale Prognose für Y gestützt auf X ist dann

$$\hat{Y}_{\text{opt}} \quad = \quad aX + b \quad = \quad E[Y] + a(X - E[X]).$$

Damit haben wir gezeigt:

Satz 3.10 (Lineare Prognose/Regression von Y gestützt auf X). Der mittlere quadratische Fehler $E[(\hat{Y} - Y)^2]$ ist minimal unter allen Zufallsvariablen der Form $\hat{Y} = aX + b$ mit $a, b \in \mathbb{R}$ für

$$\hat{Y}(\omega) \quad = \quad E[Y] \; + \; \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]} \cdot (X(\omega) - E[X]).$$

Das Problem der linearen Prognose steht in engem Zusammenhang mit der Cauchy-Schwarz-Ungleichung für die Kovarianz. In der Tat ergibt sich diese Ungleichung unmittelbar aus Gleichung (3.6):

Beweis (Cauchy-Schwarz-Ungleichung, Satz 3.7 (ii) und (iii)). Im Fall $\sigma[X] = 0$ gilt $X = E[X]$ mit Wahrscheinlichkeit 1, und die Ungleichung (3.3) ist trivialerweise erfüllt. Wir nehmen nun an, dass $\sigma[X] \neq 0$ gilt. Wählt man dann wie oben $a = \mathrm{Cov}[X,Y]/\sigma[X]^2$, so folgt aus (3.6) die Cauchy-Schwarz-Ungleichung

$$\mathrm{Var}[Y] - \frac{\mathrm{Cov}[X,Y]^2}{\mathrm{Var}[X]} \quad \geq \quad 0.$$

Die Ungleichung (3.4) folgt unmittelbar. Zudem erhalten wir nach (3.6) genau dann Gleichheit in (3.3) bzw. (3.4), wenn $\mathrm{Var}[Y - aX] = 0$ gilt, also wenn $Y - aX$ mit Wahrscheinlichkeit 1 konstant ist. In diesem Fall folgt $\mathrm{Cov}[X,Y] = \mathrm{Cov}[X, aX] = a\,\mathrm{Var}[X]$, also hat $\varrho[X,Y]$ dasselbe Vorzeichen wie a. ■

Beispiel (Regressionsgerade, Methode der kleinsten Quadrate). Wenn die gemeinsame Verteilung von X und Y eine empirische Verteilung von Daten $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots n$, ist, d.h. wenn

$$(X,Y) \; = \; (x_i, y_i) \qquad \text{mit Wahrscheinlichkeit } 1/n$$

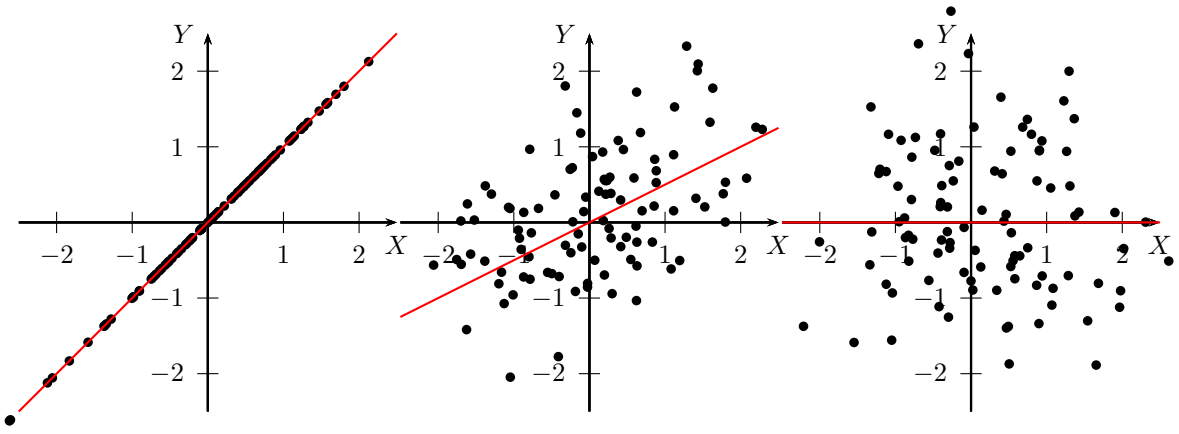
für $1 \leq i \leq n$ gilt, dann sind die Erwartungswerte und die Kovarianz gegeben durch

$$\begin{aligned} E[X] \quad &= \quad \frac{1}{n}\sum_{i=1}^{n} x_i \;=:\; \overline{x}_n, \qquad E[Y] \;=\; \overline{y}_n, \\ \mathrm{Cov}[X,Y] \quad &= \quad \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n) \quad = \quad \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x}_n \overline{y}_n. \end{aligned}$$

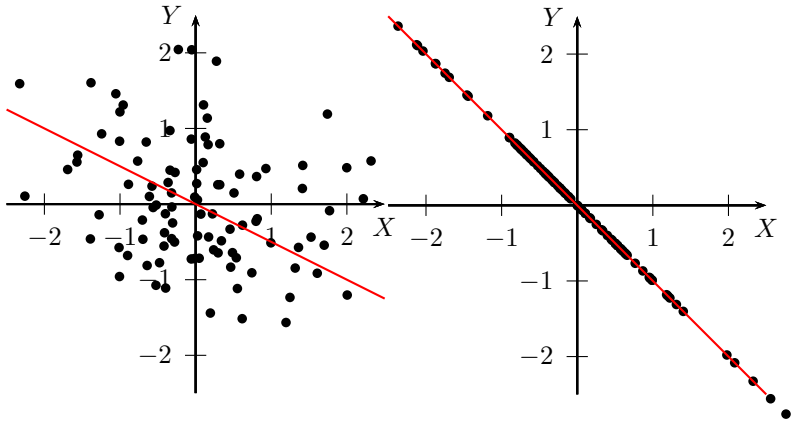
Der entsprechende *empirische Korrelationskoeffizient* der Daten (x_i, y_i), $1 \leq i \leq n$, ist

$$\varrho[X,Y] \quad = \quad \frac{\mathrm{Cov}[X,Y]}{\sigma[X]\sigma[Y]} \quad = \quad \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\left(\sum\limits_{i=1}^{n}(x_i - \overline{x}_n)^2\right)^{1/2}\left(\sum\limits_{i=1}^{n}(y_i - \overline{y}_n)^2\right)^{1/2}}$$

Diesen verwendet man als Schätzer für die Korrelation von Zufallsgrößen mit unbekannten Verteilungen. Die Grafiken in Abbildung 3.6 zeigen Datensätze mit verschiedenen Korrelationskoeffizienten ϱ.



(a) $\varrho = 1$ (b) $\varrho = \frac{1}{2}$ (c) $\varrho = 0$



(d) $\varrho = -\frac{1}{2}$ (e) $\varrho = -1$

Abbildung 3.6: Korrelationskoeffizienten und Regressionsgeraden für verschiedene Datensätze

Als beste lineare Prognose von Y gestützt auf X im quadratischen Mittel erhalten wir die *Regressionsgerade* $y = ax + b$, die die Quadratsumme

$$\sum_{i=1}^{n} (ax_i + b - y_i)^2 \quad = \quad n \cdot \mathrm{MSE}$$

der Abweichungen minimiert. Hierbei gilt nach Satz 3.10:

$$a = \frac{\mathrm{Cov}[X,Y]}{\sigma[X]^2} = \frac{\sum (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum (x_i - \overline{x}_n)^2} \qquad \text{und} \qquad b = E[Y] - a \cdot E[X] = \overline{y}_n - a \cdot \overline{x}_n.$$

Die Regressionsgeraden sind in Abbildung 3.6 eingezeichnet.

3.3 Gesetz der großen Zahlen für schwach korrelierte Zufallsvariablen

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

$$S_n(\omega) \quad = \quad X_1(\omega) + \cdots + X_n(\omega).$$

Wir betrachten die empirischen Mittelwerte

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

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

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

bzw.

$$\frac{S_n}{n} - \frac{E[S_n]}{n} \quad \stackrel{n\to\infty}{\longrightarrow} \quad 0. \tag{3.7}$$

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

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

Andererseits erwartet man ein Wegmitteln des Zufalls bei *unabhängigen* Wiederholungen desselben Zufallsexperiments. Wir werden nun zeigen, dass schon ein rasches Abklingen der Kovarianzen der Zufallsvariablen X_i genügt, um ein Gesetz der großen Zahlen zu erhalten. Dazu berechnen wir die Varianzen der Mittelwerte S_n/n, und schätzen Anschließend die Wahrscheinlichkeiten, dass die zentrierten Mittelwerte in (3.7) einen Wert größer als ε annehmen, durch die Varianzen ab.

Varianz von Summen

Die Varianz einer Summe von reellwertigen Zufallsvariablen können wir mithilfe der Kovarianzen berechnen:

Lemma 3.11. *Für Zufallsvariablen $X_1, \ldots, X_n \in \mathcal{L}^2$ gilt:*

$$\operatorname{Var}[X_1 + \cdots + X_n] \;=\; \sum_{i=1}^{n} \operatorname{Var}[X_i] \;+\; 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \operatorname{Cov}[X_i, X_j].$$

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

$$\operatorname{Var}[X_1 + \cdots + X_n] \;=\; \sum_{i=1}^{n} \operatorname{Var}[X_i].$$

Beweis. Aufgrund der Bilinearität und Symmetrie der Kovarianz gilt

$$\begin{aligned} \operatorname{Var}[X_1 + \cdots + X_n] \;&=\; \operatorname{Cov}\Big[\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\Big] \;=\; \sum_{i,j=1}^{n} \operatorname{Cov}[X_i, X_j] \\ &=\; \sum_{i=1}^{n} \operatorname{Var}[X_i] \;+\; 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \operatorname{Cov}[X_i, X_j]. \end{aligned}$$

■

Beispiel (Varianz der Binomialverteilung). Eine mit Parametern n und p binomialverteilte Zufallsvariable ist gegeben durch $S_n = \sum_{i=1}^{n} X_i$ mit unabhängigen, Bernoulli(p)-verteilten Zufallsvariablen X_i, d.h.

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

Da unabhängige Zufallsvariablen auch unkorreliert sind, erhalten wir mit Lemma 3.11 für die Varianz der Binomialverteilung:

$$\operatorname{Var}[S_n] \;=\; \sum_{i=1}^{n} \operatorname{Var}[X_i] \;=\; n\,p\,(1-p).$$

Insbesondere ist die Standardabweichung einer $\operatorname{Bin}(n,p)$-verteilten Zufallsvariable von der Ordnung $O(\sqrt{n})$.

Gesetz der großen Zahlen

Für den Beweis des Gesetzes der großen Zahlen nehmen wir an, dass $X_1, X_2, \ldots$ diskrete Zufallsvariablen aus $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ sind, die die folgende Voraussetzung erfüllen:

Annahme (Schneller Abfall der Korrelationen): Es existiert eine Folge $c_n \in \mathbb{R}$ ($n \in \mathbb{Z}_+$) mit

$$\sum_{n=0}^{\infty} c_n \;<\; \infty \qquad \text{und} \qquad \operatorname{Cov}[X_i, X_j] \;\le\; c_{|i-j|} \quad \text{für alle } i, j \in \mathbb{N}. \tag{3.8}$$

Die Annahme ist z.B. immer erfüllt, wenn die beiden folgenden Bedingungen erfüllt sind:

(i) Die Zufallsvariablen sind unkorreliert: $\operatorname{Cov}[X_i, X_j] \;=\; 0$ für alle $i \neq j$.

(ii) Die Varianzen sind beschränkt: $v \;:=\; \sup_{i \in \mathbb{N}} \operatorname{Var}[X_i] \;<\; \infty$.

In diesem Fall können wir in (3.8) $c_0 = v$ und $c_n = 0$ für $n \neq 0$ wählen. Insbesondere setzen wir keine Unabhängigkeit voraus, sondern nur Bedingungen an die Kovarianzen.

Satz 3.12 (Gesetz der großen Zahlen für schwach korrelierte Zufallsvariablen). Ist die Annahme erfüllt, dann gilt für alle $\varepsilon > 0$ und $n \in \mathbb{N}$:

$$P\left[\left|\frac{S_n}{n} - \frac{E[S_n]}{n}\right| \ge \varepsilon\right] \;\le\; \frac{C}{\varepsilon^2\, n} \qquad \text{mit} \qquad C \;:=\; c_0 + 2\sum_{n=1}^{\infty} c_n \;<\; \infty.$$

Ist insbesondere $E[X_i] = m$ für alle $i \in \mathbb{N}$, dann ***konvergieren*** die Mittelwerte ***stochastisch*** gegen den Erwartungswert m, d.h.

$$\lim_{n\to\infty} P\left[\left|\frac{S_n}{n} - m\right| \ge \varepsilon\right] = 0 \qquad \text{für jedes } \varepsilon > 0.$$

Der Beweis des Gesetzes der großen Zahlen ergibt sich unmittelbar aus Lemma 3.11 und Satz 3.5:

Beweis. Nach der Annahme und Lemma 3.11 gilt

$$\operatorname{Var}\left[\frac{S_n}{n}\right] \;=\; \frac{1}{n^2} \operatorname{Var}\left[\sum_{i=1}^{n} X_i\right] \;=\; \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}[X_i, X_j] \;\le\; \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} c_{|i-j|} \;\le\; \frac{C}{n}.$$

Die Varianz der Mittelwerte fällt also mit Ordnung $O(1/n)$ ab. Mithilfe der Čebyšev-Ungleichung erhalten wir

$$P\left[\left|\frac{S_n}{n} - \frac{E[S_n]}{n}\right| \ge \varepsilon\right] \;\le\; \frac{1}{\varepsilon^2} \operatorname{Var}\left[\frac{S_n}{n}\right] \;\le\; \frac{C}{n\,\varepsilon^2}.$$

für alle $\varepsilon > 0$ und $n \in \mathbb{N}$. ■

Beispiel. Sind $X_1, X_2, \ldots$ unkorrelierte (also beispielsweise unabhängige) und identisch verteilte Zufallsvariablen aus $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ mit $E[X_i] = m$ und $\mathrm{Var}[X_i] = v$ für alle i, dann ist die Annahme mit $c_0 = v$ und $c_n = 0$ für $n \neq 0$ erfüllt, und wir erhalten die Abschätzung

$$P\left[\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right] \;\leq\; \frac{C}{\varepsilon^2 n}$$

für den Abstand des Mittelwerts der Zufallsvariablen vom Erwartungswert.

Unter den Voraussetzungen aus Satz 3.12 gilt auch ein starkes Gesetz der großen Zahlen:

Satz 3.13 (Starkes Gesetz der großen Zahlen für schwach korrelierte Zufallsvariablen). Ist die Annahme oben erfüllt, und gilt $E[X_i] = m$ für alle $i \in \mathbb{N}$, dann ***konvergieren*** die Mittelwerte ***fast sicher*** gegen den Erwartungswert m, d.h.

$$P\left[\lim_{n\to\infty} \frac{S_n}{n} = m\right] = 1\,.$$

Der Beweis dieser Aussage wird in der Vorlesung Einführung in die Wahrscheinlichkeitstheorie gegeben.

Schätzen von Kenngrößen

Sei X eine reellwertige Zufallsvariable mit $E[X^2] < \infty$. In vielen Anwendungen kennen wir die Verteilung μ von X nicht, oder wir können Erwartungswerte und Wahrscheinlichkeiten nicht explizit berechnen. In diesen Fällen können wir solche Kenngrößen aus unabhängigen Stichproben von X schätzen. Dies wird sowohl in der Statistik bei der Parameterschätzung, als auch in der stochastischen Simulation bei der Monte-Carlo Berechnung von Erwartungswerten verwendet. Im ersten Fall sind die Stichproben Beobachtungswerte, im zweiten Fall werden sie auf dem Computer simuliert.

Wir interpretieren die Stichproben als Realisierungen unabhängiger Zufallsvariablen $X_1, X_2, \ldots$, die auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definiert sind.

(i) Schätzen des Erwartungswertes: Um den Erwartungswert $m = E[X]$ zu schätzen, verwenden wir die *empirischen Mittelwerte*

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

Das empirische Mittel ist ein *erwartungstreuer Schätzer* für m, d.h. $\overline{X}_n$ ist eine Funktion von den Beobachtungswerten $X_1, \ldots, X_n$ mit

$$E[\overline{X}_n] \;=\; m\,.$$

Nach dem Gesetz der großen Zahlen ist $(\overline{X}_n)_{n\in\mathbb{N}}$ zudem eine *konsistente Folge von Schätzern* für m, d.h. es gilt

$$\overline{X}_n \quad\longrightarrow\quad m \qquad P\text{-stochastisch bzw. } P\text{-fast sicher.}$$

Zudem können wir basierend auf den Stichproben *Konfidenzintervalle* für m angeben. Sei dazu $\varepsilon > 0$. Dann gilt wie oben gezeigt

$$P\left[m \notin (\overline{X}_n - \varepsilon, \overline{X}_n + \varepsilon)\right] \;=\; P\left[|\overline{X}_n - m| \geq \varepsilon\right] \;\leq\; \frac{\mathrm{Var}[X]}{\varepsilon^2 n}\,.$$

Sind ε und n beispielsweise so gewählt, dass die rechte Seite kleiner oder gleich $0{,}05$ ist, dann folgt, dass das zufällige Intervall $I = (\overline{X}_n - \varepsilon, \overline{X}_n + \varepsilon)$ ein 95%-*Konfidenzintervall für* m ist, d.h. die Wahrscheinlichkeit, dass der tatsächliche Wert von m in diesem zufälligen Intervall liegt, beträgt mindestens $0{,}95$.

(ii) Schätzen der Varianz: Um die Varianz $v = \mathrm{Var}[X]$ zu schätzen, verwendet man meistens die *renormierte Stichprobenvarianz*

$$\widetilde{V}_n \quad := \quad \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Der Vorfaktor $\frac{1}{n-1}$ (statt $\frac{1}{n}$) gewährleistet, dass $\widetilde{V}_n$ ein *erwartungstreuer* Schätzer für v ist, denn aus

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\overline{X}_n - m)^2 \qquad (3.9)\\ \text{Stichprobenvarianz} &= \text{MSE} - \text{Stichprobenbias}^2 \end{aligned}$$

folgt

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[X_i] - \mathrm{Var}[\overline{X}_n] = \left(\frac{1}{n} - \frac{1}{n^2}\right)\sum_{i=1}^{n}\mathrm{Var}[X_i] = \frac{n-1}{n}v,$$

also $E[\widetilde{V}_n] = v$. Um zu zeigen, dass $(\widetilde{V}_n)_{n\in\mathbb{N}}$ eine *konsistente* Folge von Schätzern für v ist, können wir erneut das Gesetz der großen Zahlen anwenden. Da die Zufallsvariablen $X_i - \overline{X}_n, 1 \le i \le n$, selbst nicht unabhängig sind, verwenden wir dazu die Zerlegung (3.9). Aus dem starken Gesetz der großen Zahlen folgt dann

$$\frac{n-1}{n}\widetilde{V}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\overline{X}_n - m)^2 \longrightarrow v \qquad P\text{-fast sicher},$$

also auch $\widetilde{V}_n \to v$ P-fast sicher.

(iii) Schätzen von allgemeinen Erwartungswerten: Allgemeiner können wir für jede Funktion f mit $E[|f(X)|] < \infty$ den Erwartungswert $\theta = E[f(X)]$ erwartungstreu durch die *empirischen Mittelwerte*

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

schätzen. Da die Zufallsvariablen $f(X_i)$ wieder unabhängig und identisch verteilt sind mit Erwartungswert θ, gilt nach dem Gesetz der großen Zahlen:

$$\widehat{\theta}_n \longrightarrow \theta \qquad P\text{-stochastisch und } P\text{-fast sicher}. \qquad (3.10)$$

(iv) Schätzen der Verteilung: Die gesamte Verteilung

$$\mu[B] = P[X \in B]$$

können wir durch die *empirische Verteilung*

$$\widehat{\mu}_n[B] = \frac{1}{n}\sum_{i=1}^{n} I_B(X_i) = \frac{|\{i = 1,\dots,n : X_i \in B\}|}{n}$$

der Zufallsstichprobe $X_1,\dots,X_n$ schätzen. Diese ist eine „*zufällige Wahrscheinlichkeitsverteilung*“. Nach (iii) ist $\widehat{\mu}_n[B]$ ein erwartungstreuer Schätzer für $\mu[B]$, und es gilt

$$\widehat{\mu}_n[B] \overset{n\to\infty}{\longrightarrow} \mu[B] \qquad P\text{-stochastisch und } P\text{-fast sicher}. \qquad (3.11)$$

Konfidenzintervalle für $\mu[B]$ kann man entweder wie oben mithilfe der Čebyšev-Ungleichung oder, wie im Beispiel in Abschnitt 3.1, über die Bernstein-Ungleichung herleiten.

3.4 Konvergenzsätze für Markov-Ketten

Sei S eine abzählbare Menge, ν eine Wahrscheinlichkeitsverteilung auf S, und $\pi = (\pi(x,y))_{x,y\in S}$ eine stochastische Matrix. Hier und im folgenden bezeichnen wir diskrete Wahrscheinlichkeitsverteilungen und die entsprechenden Massenfunktionen mit demselben Buchstaben, d.h. $\nu(x) := \nu[\{x\}]$. Wir interpretieren $\nu = (\nu(x))_{x\in S}$ auch als Zeilenvektor in $\mathbb{R}^S$.

In Abschnitt 2.2 haben wir das kanonische Modell für eine (zeithomogene) Markovkette mit Startverteilung ν und Übergangsmatrix π eingeführt. Allgemeiner definieren wir:

Definition 3.14. Eine Folge $X_0, X_1, \ldots : \Omega \to S$ von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ heißt **zeitlich homogene Markov-Kette** mit Startverteilung ν und Übergangsmatrix π, falls die folgenden Bedingungen erfüllt sind:

(i) für alle $x_0 \in S$ gilt $P[X_0 = x_0] = \nu(x_0)$.

(ii) für alle $n \in \mathbb{N}$ und $x_0, \ldots, x_{n+1} \in S$ mit $P[X_0 = x_0, \ldots, X_n = x_n] \neq 0$ gilt

$$P[X_{n+1} = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n] \;=\; \pi(x_n, x_{n+1}).$$

Die Bedingungen (i) und (ii) sind äquivalent dazu, dass

$$P[X_0 = x_0, \ldots, X_n = x_n] \;=\; \nu(x_0)\,\pi(x_0, x_1) \cdots \pi(x_{n-1}, x_n)$$

für alle $n \in \mathbb{Z}_+$ und $x_0, x_1, \ldots, x_n \in S$ gilt. Eine Folge $(X_k)_{k\in\mathbb{Z}_+}$ von Zufallsvariablen mit Werten in S ist also genau dann eine zeithomogene Markovkette mit Startverteilung ν und Übergangsmatrix π, wenn die gemeinsame Verteilung von $X_0, X_1, \ldots, X_n$ für jedes n mit der Verteilung im entsprechenden kanonischen Modell übereinstimmt.

Gleichgewichte und Detailed Balance

Satz 2.8 zeigt, dass die Verteilung einer zeithomogenen Markovkette zur Zeit n durch das Produkt $\nu\pi^n$ des Zeilenvektors ν der Massenfunktion der Startverteilung mit dem n fachen Matrixprodukt der Übergangsmatrix π gegeben ist. Gilt $\nu\,\pi = \nu$, dann folgt $X_n \sim \nu$ für alle $n \in \mathbb{Z}_+$, d.h. die Markovkette mit Startverteilung ν ist „stationär“.

Definition 3.15. i) Eine Wahrscheinlichkeitsverteilung μ auf S heißt **Gleichgewichtsverteilung** (oder **invariante Verteilung**) der Übergangsmatrix π, falls $\mu\,\pi = \mu$ gilt, d.h. falls

$$\sum_{x\in S} \mu(x)\,\pi(x,y) = \mu(y) \qquad \text{für alle } y \in S.$$

ii) μ erfüllt die **Detailed Balance-Bedingung** bzgl. der Übergangsmatrix π, falls gilt:

$$\mu(x)\,\pi(x,y) = \mu(y)\,\pi(y,x) \qquad \text{für alle } x, y \in S \tag{3.12}$$

Satz 3.16. Erfüllt μ die Detailed Balance-Bedingung (3.12), dann ist μ eine Gleichgewichtsverteilung von π.

Beweis. Aus der Detailed Balance-Bedingung folgt

$$\sum_{x \in S} \mu(x)\,\pi(x,y) \;=\; \sum_{x \in S} \mu(y)\,\pi(y,x) \;=\; \mu(y) \qquad \text{für alle } y \in S.$$

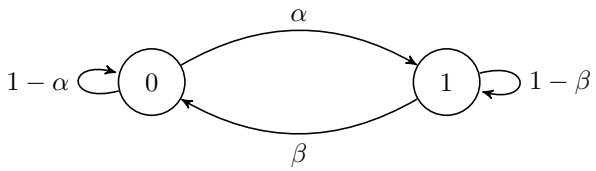
Bemerkung. Bei Startverteilung μ gilt:

$$\mu(x)\,\pi(x,y) = P[X_0 = x, X_1 = y].$$

Wir können diese Größe als „Fluss der Wahrscheinlichkeitsmasse von x nach y“ interpretieren. Die Detailed Balance- und die Gleichgewichtsbedingung haben dann die folgenden anschaulichen Interpretationen:

Detailed Balance:	$\mu(x)\,\pi(x,y)$	=	$\mu(y)\,\pi(y,x)$
	„Fluss von x nach y“	=	„Fluss von y nach x“
Gleichgewicht:	$\sum_{x \in S} \mu(x)\,\pi(x,y)$	=	$\sum_{x \in S} \mu(y)\,\pi(y,x)$
	„Gesamter Fluss nach y“		„Gesamter Fluss von y“.

Beispiele. a) Markov-Kette auf $S = \{0,1\}$:



Seien $\alpha, \beta \in [0,1]$ und $\pi = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$. Dann ist die Gleichgewichtsbedingung $\mu\, p = \mu$ äquivalent zu den folgenden Gleichungen:

$$\mu(0) = \mu(0)\,(1-\alpha) + \mu(1)\,\beta,$$
$$\mu(1) = \mu(0)\,\alpha + \mu(1)\,(1-\beta).$$

Da μ eine Wahrscheinlichkeitsverteilung ist, sind beide Gleichungen äquivalent zu

$$\beta\,(1 - \mu(0)) = \alpha\,\mu(0).$$

Die letzte Gleichung ist äquivalent zur Detailed Balance-Bedingung (3.12). Auf einem Zustandsraum mit zwei Elementen erfüllt also jede Gleichgewichtsverteilung die Detailed Balance-Bedingung. Falls $\alpha + \beta > 0$ gilt, ist $\mu = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)$ das eindeutige Gleichgewicht. Falls $\alpha = \beta = 0$ gilt, ist jede Wahrscheinlichkeitsverteilung μ eine Gleichgewichtsverteilung.

b) Zyklischer Random Walk: Sei $S = \mathbb{Z}/n\mathbb{Z}$ ein diskreter Kreis, und

$$\pi(k + n\mathbb{Z}, k+1+n\mathbb{Z}) = p, \qquad \pi(k+n\mathbb{Z}, k-1+n\mathbb{Z}) = 1-p.$$

Dann ist die Gleichverteilung $\mu(x) = \frac{1}{n}$ für jedes $p \in [0,1]$ ein Gleichgewicht von π. Die Detailed Balance-Bedingung ist dagegen nur für $p = \frac{1}{2}$, d.h. im symmetrischen Fall, erfüllt.

c) Random Walks auf Graphen:
Sei (V, E) ein endlicher Graph, und $S = V$ die Menge der Knoten. Wir nehmen an, dass von jedem Knoten mindestens eine Kante ausgeht. Der klassische Random Walk auf dem Graphen hat die Übergangswahrscheinlichkeiten

$$\pi(x,y) \;=\; \begin{cases} \frac{1}{\deg(x)} & \text{falls } \{x,y\} \in E, \\ 0 & \text{sonst.} \end{cases}$$

Die Detailed Balance-Bedingung lautet in diesem Fall:

$$\frac{\mu(x)}{\deg(x)} \;=\; \frac{\mu(y)}{\deg(y)} \qquad \text{für alle } \{x,y\} \in E.$$

Sie ist erfüllt, falls

$$\mu(x) = \deg(x)/Z$$

gilt, wobei Z eine positive Konstante ist. Damit μ eine Wahrscheinlichkeitsverteilung ist, muss

$$Z \;=\; \sum_{x \in B} \deg(x) \;=\; 2\,|E|$$

gelten. Somit ergibt sich als Gleichgewichtsverteilung

$$\mu(x) \;=\; \frac{\deg(x)}{2|E|}.$$

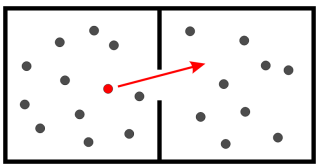
Alternativ können wir einen modifizierten Random Walk definieren, der die Gleichverteilung auf V als Gleichgewicht hat. Sei dazu $\triangle := \max_{x \in V} \deg(x)$ der maximale Grad, und

$$\pi(x,y) = \begin{cases} \frac{1}{\triangle} & \text{falls } \{x,y\} \in E, \\ 1 - \frac{\deg(x)}{\triangle} & \text{falls } x = y, \\ 0 & \text{sonst.} \end{cases}$$

Dann gilt $\pi(x,y) = \pi(y,x)$, und somit ist die Gleichverteilung auf V ein Gleichgewicht.

Ist der Graph regulär, also $\deg(x)$ konstant, dann stimmen die beiden Arten von Random Walks überein.

d) Urnenmodell von P. und T. Ehrenfest: Das Ehrenfestsche Urnenmodell ist ein einfaches Modell, dass den Austausch von Gasmolekülen zwischen zwei Behältern beschreibt, ohne die räumliche Struktur zu berücksichtigen. Im Modell ist eine feste Anzahl n von Kugeln (Molekülen) auf zwei Urnen (Behälter) verteilt. Typischerweise ist n sehr groß, z.B. $n = 10^{23}$. Zu jedem Zeitpunkt $t \in \mathbb{N}$ wechselt eine zufällig ausgewählte Kugel die Urne.



Wir können diesen Vorgang auf zwei ganz verschiedene Arten durch Markovketten beschreiben.

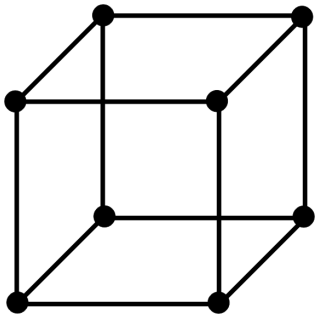
Mikroskopische Beschreibung: Ein detailliertes Modell ergibt sich, wenn wir für jede einzelne Kugel notieren, ob sich diese in der ersten Urne befindet. Der Zustandsraum ist dann

$$S = \{0,1\}^n = \{(\sigma_1, \ldots, \sigma_n) \;:\; \sigma_i \in \{0,1\}\ \forall\, i\},$$

wobei $\sigma_i = 1$ dafür steht, dass sich die i-te Kugel in der ersten Urne befindet. Man beachte, dass dieser Konfigurationsraum enorm viele Elemente enthält (z.B. $2^{10^{23}}$). Die Übergangswahrscheinlichkeiten sind durch

$$\pi(\sigma, \widetilde{\sigma}) \;=\; \begin{cases} \frac{1}{n} & \text{falls } \sum_{i=1}^n |\sigma_i - \widetilde{\sigma}_i| = 1, \\ 0 & \text{sonst,} \end{cases}$$

gegeben. Die resultierende Markov-Kette ist ein Random Walk auf dem (in der Regel sehr hochdimensionalen) diskreten HyperWürfel $\{0,1\}^n$, d.h. sie springt in jedem Schritt von einer Ecke des HyperWürfels zu einer zufällig ausgewählten benachbarten Ecke. Die Gleichverteilung auf dem Hyperwürfel ist das eindeutige Gleichgewicht.



Makroskopische Beschreibung: Wir betrachten nur die Anzahl der Kugeln in der ersten Urne. Der Zustandsraum ist dann

$$S = \{0, 1, 2, \ldots, n\},$$

und die Übergangswahrscheinlichkeiten sind durch

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

gegeben, da in jedem Schritt mit Wahrscheinlichkeit x/n eine Kugel aus der ersten Urne gezogen wird, wenn sich x Kugeln dort befinden. Da sich im mikroskopischen Gleichgewicht jede Kugel mit Wahrscheinlichkeit $\frac{1}{2}$ in jeder der beiden Urnen befindet, können wir erwarten, dass die Binomialverteilung $\mu(x) = \binom{n}{x} 2^{-n}$ mit Parameter $p = \frac{1}{2}$ ein Gleichgewicht der makroskopischen Dynamik ist. Tatsächlich erfüllt die Binomialverteilung die Detailed Balance-Bedingung

$$\mu(x-1)\,\pi(x-1,x) \;=\; \mu(x)\,\pi(x,x-1) \qquad \text{für } x = 1,\ldots,n,$$

denn es gilt

$$2^{-n}\,\frac{n!}{(x-1)!(n-(x-1))!}\,\frac{n-(x-1)}{n} \;=\; 2^{-n}\,\frac{n!}{x!(n-x)!}\,\frac{x}{n}.$$

Konvergenz ins Gleichgewicht

Wir wollen nun zeigen, dass sich unter geeigneten Voraussetzungen die Verteilung einer Markovkette zur Zeit n für $n \to \infty$ einer Gleichgewichtsverteilung annähert, die nicht von der Startverteilung abhängt. Um dies mathematisch zu präzisieren, benötigen wir einen Abstandsbegriff für Wahrscheinlichkeitsverteilungen. Sei

$$\mathrm{WV}(S) := \{\nu = (\nu(x))_{x\in S} : \nu(x) \geq 0 \;\forall\, x, \quad \sum_{x\in S} \nu(x) = 1\}$$

die Menge aller (Massenfunktionen von) Wahrscheinlichkeitsverteilungen auf der abzählbaren Menge S. Ist S endlich mit m Elementen, dann ist $\mathrm{WV}(S)$ ein Simplex im $\mathbb{R}^m$. Wir führen nun einen Abstandsbegriff auf $\mathrm{WV}(S)$ ein:

Definition 3.17. Die **(totale) Variationsdistanz** zweier Wahrscheinlichkeitsverteilungen μ, ν auf S ist:

$$d_{TV}(\mu,\nu) \;:=\; \frac{1}{2}\|\mu-\nu\|_1 \;:=\; \frac{1}{2}\sum_{x\in S} |\mu(x)-\nu(x)|.$$

Man prüft leicht nach, dass d_{TV} tatsächlich eine Metrik auf $\mathrm{WV}(S)$ ist.

Bemerkung. a) Für alle $\mu, \nu \in \mathrm{WV}(S)$ gilt:

$$d_{TV}(\mu,\nu) \;\le\; \frac{1}{2}\sum_{x\in S}(\mu(x)+\nu(x)) \;=\; 1.$$

b) Seien $\mu, \nu \in \mathrm{WV}(S)$ und $B := \{x \in S : \mu(x) \ge \nu(x)\}$. Dann gilt

$$d_{TV}(\mu,\nu) \;=\; \sum_{x\in B}(\mu(x)-\nu(x)) \;=\; \max_{A\subseteq S}|\mu(A)-\nu(A)|.$$

Diese Aussage zeigt, dass d_{TV} eine sehr natürliche Abstandsfunktion auf Wahrscheinlichkeitsverteilungen ist. Der Beweis der Aussage ist eine Übungsaufgabe.

Wir betrachten nun eine stochastische Matrix $(\pi(x,y))_{x,y\in S}$ mit Gleichgewichtsverteilung μ. Die Verteilung einer Markov-Kette mit Startverteilung ν und Übergangsmatrix π zur Zeit n ist $\nu\,\pi^n$. Um Konvergenz ins Gleichgewicht zu zeigen, verwenden wir die folgende Annahme:

Minorisierungsbedingung: Es gibt ein $\delta \in (0,1]$ und ein $r \in \mathbb{N}$, so dass

$$\pi^r(x,y) \ge \delta\cdot\mu(y) \qquad \text{für alle } x,y\in S \text{ gilt.} \tag{3.13}$$

Satz 3.18 (Konvergenzsatz von W. Doeblin). Gilt die Minorisierungsbedingung (3.13), dann konvergiert $\nu\,\pi^n$ für jede Startverteilung ν exponentiell schnell gegen μ. Genauer gilt für alle $n \in \mathbb{Z}_+$ und $\nu \in \mathrm{WV}(S)$:

$$d_{TV}(\nu\,\pi^n,\mu) \;\le\; (1-\delta)^{\lfloor n/r\rfloor}.$$

Bemerkung. Insbesondere ist μ unter der Voraussetzung des Satzes das *eindeutige* Gleichgewicht von π, denn für eine beliebige Wahrscheinlichkeitsverteilung ν mit $\nu\,\pi = \nu$ gilt

$$d_{TV}(\nu,\mu) \;=\; d_{TV}(\nu\,\pi^n,\mu) \;\to\; 0 \qquad \text{für } n\to\infty,$$

und damit $\nu = \mu$.

Beweis. 1. Durch die Zerlegung

$$\pi^r(x,y) \;=\; \delta\,\mu(y) + (1-\delta)\,q(x,y)$$

der r-Schritt-Übergangswahrscheinlichkeiten wird eine *stochastische* Matrix q definiert, denn:

(i) Aus der Minorisierungsbedingung (3.13) folgt $q(x,y) \ge 0$ für alle $x,y \in S$.

(ii) Aus $\sum_{y\in S}\pi^r(x,y) = 1$, $\sum_{y\in S}\mu(y) = 1$ folgt $\sum_{y\in S} q(x,y) = 1$ für alle $x \in S$.

Wir setzen im folgenden $\lambda := 1-\delta$. Dann gilt für alle $\nu \in \mathrm{WV}(S)$:

$$\nu\,\pi^r = (1-\lambda)\,\mu + \lambda\,\nu\,q. \tag{3.14}$$

2. Wir zeigen mit vollständiger Induktion:

$$\nu\,\pi^{kr} = (1-\lambda^k)\,\mu + \lambda^k\,\nu\,q^k \qquad \text{für alle } k\ge 0,\quad \nu\in\mathrm{WV}(S). \tag{3.15}$$

für $k = 0$ ist die Aussage offensichtlich wahr. Gilt (3.15) für ein $k \geq 0$, dann erhalten wir durch Anwenden von Gleichung (3.14) auf $\widetilde{\nu}\,\pi^r$ mit $\widetilde{\nu} = \nu\, q^k$:

$$\begin{aligned}
\nu\,\pi^{(k+1)r} &= \nu\,\pi^{kr}\,\pi^r \\
&= ((1-\lambda^k)\,\mu + \lambda^k \underbrace{\nu\, q^k}_{=\widetilde{\nu}})\,\pi^r \\
&= (1-\lambda^k)\underbrace{\mu\,\pi^r}_{=\mu} + (1-\lambda)\,\lambda^k\mu + \lambda^{k+1}\,\nu\, q^k\, q \\
&= (1-\lambda^{k+1})\,\mu + \lambda^{k+1}\,\nu\, q^{k+1}.
\end{aligned}$$

3. Sei $n \in \mathbb{Z}_+$. Dann gilt $n = k\,r + i$ mit $k \in \mathbb{Z}_+$ und $0 \leq i < r$. Damit folgt für $\nu \in \mathrm{WV}(S)$:

$$\nu\,\pi^n = \nu\,\pi^{kr}\,\pi^i \;=\; (1-\lambda^k)\underbrace{\mu\,\pi^i}_{=\mu} + \lambda^k\,\nu\, q^k\,\pi^i, \qquad \text{also}$$

$$\nu\,\pi^n - \mu = \lambda^k\,(\nu\, q^k\,\pi^i - \mu), \qquad \text{und damit}$$

$$d_{TV}(\nu\,\pi^n, \mu) = \frac{1}{2}\,\|\nu\,\pi^n - \mu\|_1 \;=\; \lambda^k\, d_{TV}(\nu\, q^k\,\pi^i, \mu) \;\leq\; \lambda^k. \qquad \blacksquare$$

Auf abzählbar unendlichen Zustandsräumen ist die Minorisierungsbedingung eine relativ restriktive Annahme. Es gibt Erweiterungen des obigen Satzes, die unter deutlich schwächeren Voraussetzungen ähnliche Konvergenzaussagen liefern. Ist der Zustandsraum dagegen endlich, dann können wir den obigen Konvergenzsatz verwenden, um die Konvergenz ins Gleichgewicht unter minimalen Voraussetzungen zu beweisen. Dazu zeigen wir, dass die Minorisierungsbedingung immer erfüllt ist, wenn der Zustandsraum endlich, und die Übergangsmatrix *irreduzibel* ist und einen *aperiodischen Zustand* besitzt:

Definition 3.19. i) Eine stochastische Matrix π heißt **irreduzibel**, falls es für alle $x, y \in S$ ein $n \in \mathbb{N}$ gibt, so dass $\pi^n(x,y) > 0$ gilt.

ii) Ein Zustand $x \in S$ heißt **aperiodisch bzgl.** π, falls ein $n_0 \in \mathbb{N}$ existiert, so dass $\pi^n(x,x) > 0$ für alle $n \geq n_0$ gilt.

Bemerkung. a) Allgemeiner definiert man die **Periode** eines Zustands $x \in S$ als

$$\mathrm{Periode}(x) \;:=\; \mathrm{ggT}\,\{n \in \mathbb{N} \mid \pi^n(x,x) > 0\}.$$

Man kann dann zeigen, dass x genau dann aperiodisch ist, wenn $\mathrm{Periode}(x) = 1$ gilt. Ein Beispiel für eine Übergangsmatrix mit Periode 2 ist die Matrix $\pi = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ auf einem zweielementigen Zustandsraum. Die entsprechende Markovkette wechselt in jedem Schritt mit Wahrscheinlichkeit 1 den Zustand.

b) Ist π irreduzibel, dann folgt aus der Existenz eines aperiodischen Zustands bereits, dass alle Zustände aperiodisch sind.

Beispiel (Irreduzibilität von Random Walks auf Graphen). Die Übergangsmatrix eines Random Walks auf einem endlichen Graphen ist genau dann irreduzibel, wenn der Graph zusammenhängend ist.

Korollar 3.20 (Konvergenzsatz für endliche Markov-Ketten). Ist der Zustandsraum S endlich, die Übergangsmatrix π irreduzibel, und existiert ein aperiodischer Zustand $a \in S$, dann gilt:

$$\lim_{n\to\infty} d_{TV}(\nu\,\pi^n, \mu) = 0 \qquad \text{für alle } \nu \in \mathrm{WV}(S).$$

Beweis. Wir zeigen, dass zu jedem $x, y \in S$ eine natürliche Zahl $k(x,y)$ existiert, so dass

$$\pi^n(x,y) > 0 \qquad \text{für alle } n \geq k(x,y) \tag{3.16}$$

gilt. Da der Zustandsraum endlich ist, folgt hieraus, dass die Minorisierungsbedingung (3.13) mit

$$r = \max_{x,y\in S} k(x,y) < \infty \qquad \text{und} \qquad \delta = \min_{x,y\in S} \pi^r(x,y) > 0$$

erfüllt ist.

Zum Beweis der obigen Behauptung seien $x, y \in S$ fest gewählt. Wegen der Irreduzibilität von π existieren dann $i, j \in \mathbb{N}$ mit $\pi^i(x,a) > 0$ und $\pi^j(a,y) > 0$. Da a aperiodisch ist, existiert zudem ein $n_0 \in \mathbb{N}$ mit $\pi^n(a,a) > 0$ für alle $n \geq n_0$. Damit folgt

$$\pi^{i+n+j}(x,y) \;\geq\; \pi^i(x,a)\,\pi^n(a,a)\,\pi^j(a,y) > 0 \qquad \text{für alle } n \geq n_0,$$

und somit $\pi^n(x,y) > 0$ für alle $n \geq i + n_0 + j$. Also ist die Behauptung für x, y mit $k(x,y) = i + n_0 + j$ erfüllt. ∎

Beispiel (Träger Random Walk auf endlichem Graphen). Ein Random Walk auf einem endlichen Graphen ist im Allgemeinen nicht aperiodisch; zum Beispiel hat der Random Walk auf $\mathbb{Z}/(n\mathbb{Z})$ Periode 2 falls n gerade ist. Um Aperiodizität zu gewährleisten genügt aber eine kleine Modifikation der Übergangsmatrix: Setzen wir

$$\pi(x,y) \;=\; \begin{cases} \varepsilon & \text{für } y = x, \\ \frac{1-\varepsilon}{\deg(x)} & \text{für } \{x,y\} \in E \text{ mit } x \neq y, \\ 0 & \text{sonst,} \end{cases}$$

mit einer festen Konstanten $\varepsilon > 0$, dann sind alle Zustände aperiodisch, und π hat weiterhin das Gleichgewicht $\mu(x) = \deg(x)/(2|E|)$. Die Markovkette mit Übergangsmatrix π ist ein „träger" Random Walk, der in jedem Schritt mit Wahrscheinlichkeit ε beim selben Zustand bleibt. Ist der Graph zusammenhängend, dann ist π irreduzibel. Es folgt, dass die Verteilung des trägen Random Walks zur Zeit n für eine beliebige Startverteilung gegen μ konvergiert.

Gesetz der großen Zahlen für stationäre Markovketten

Das Gesetz der großen Zahlen kann auch auf Mittelwerte von stationären Markovketten angewendet werden. Sei $(Y_n)_{n\in\mathbb{Z}_+}$ eine auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definierte Markovkette mit abzählbarem Zustandsraum S und Übergangsmatrix $\pi = (\pi(x,y))_{x\in S}$. Wir nehmen an, dass die Markovkette im Gleichgewicht startet, d.h. die Verteilung μ von Y_0 ist ein Gleichgewicht von π. Dann gilt

$$Y_n \sim \mu \qquad \text{für alle } n \geq 0. \tag{3.17}$$

Wir betrachten nun die *Anzahl der Besuche*

$$S_n \;=\; \sum_{i=0}^{n-1} I_A(Y_i)$$

in einer Teilmenge A des Zustandsraums S während der ersten n Schritte der Markovkette. Erfüllt die Übergangsmatrix eine Minorisierungsbedingung, dann können wir zeigen, dass die Kovarianzen der Zufallsvariablen $X_i = I_A(Y_{i-1})$ rasch abklingen, und daher das Gesetz der großen Zahlen anwenden:

Korollar 3.21 (Gesetz der großen Zahlen für stationäre Markovketten). Ist die Minorisierungsbedingung (3.13) erfüllt, dann existiert eine Konstante $C \in (0, \infty)$, so dass

$$P\left[\left|\frac{S_n}{n} - \mu[A]\right| \geq \varepsilon\right] \;\leq\; \frac{C}{\varepsilon^2\, n}$$

für alle $\varepsilon > 0$, $n \in \mathbb{N}$ und $A \subseteq S$ gilt.

Die Zufallsvariable S_n/n beschreibt die relative Häufigkeit von Besuchen in der Menge A während der ersten n Schritte der Markovkette. Das Korollar zeigt, dass sich diese relative Häufigkeit für $n \to \infty$ der Wahrscheinlichkeit $\mu[A]$ der Menge A bezüglich der Gleichgewichtsverteilung μ annähert. Dies kann zum näherungsweisen Berechnen der relativen Häufigkeiten für große n, oder aber umgekehrt zum Schätzen der Gleichgewichts-Wahrscheinlichkeiten durch relative Häufigkeiten verwendet werden.

Beweis (Beweis des Korollars). Seien $A \subseteq S$ und $i, n \in \mathbb{Z}_+$. Um die Annahme in Satz 3.12 zu verifizieren, schätzen wir die Kovarianzen der Zufallsvariablen $I_A(Y_i)$ und $I_A(Y_{i+n})$ ab. Nach (3.17) haben Y_i und Y_{i+n} beide die Verteilung μ. Zudem folgt aus der Markov-Eigenschaft, dass

$$P[Y_i = a \text{ und } Y_{i+n} = b] \;=\; \mu(a)\pi^n(a,b) \qquad \text{für alle } a, b \in S$$

gilt. Damit erhalten wir

$$\begin{aligned}
\operatorname{Cov}\left[I_A(Y_i), I_A(Y_{i+n})\right] \;&=\; E\left[I_A(Y_i)\, I_A(Y_{i+n})\right] \;-\; E\left[I_A(Y_i)\right] E\left[I_A(Y_{i+n})\right] \\
&= \sum_{a\in A}\sum_{b\in A} P[Y_i = a, Y_{i+n} = b] \;-\; \sum_{a\in A} P[Y_i = a] \sum_{b\in A} P[Y_{i+n} = b] \\
&= \sum_{a\in A}\sum_{b\in A} \mu(a)\pi^n(a,b) \;-\; \sum_{a\in A}\mu(a) \sum_{b\in A}\mu(b) \\
&= \sum_{a\in A}\mu(a) \sum_{b\in A}(\pi^n(a,b) - \mu(b)) \\
&\leq\; 2\sum_{a\in A}\mu(a)\, d_{TV}(\pi^n(a,\cdot),\mu) \\
&\leq\; 2\sum_{a\in A}\mu(a)\,(1-\delta)^{\lfloor n/r\rfloor} \;\leq\; 2\,(1-\delta)^{\lfloor n/r\rfloor}.
\end{aligned}$$

Hierbei ist $\pi^n(a,\cdot)$ die Verteilung der Markovkette mit Start in a nach n Schritten. Die Abschätzung in der vorletzten Zeile gilt nach Definition der Variationsdistanz, und die zentrale Abschätzung in der letzten Zeile folgt nach Satz 3.18 aus der Minorisierungsbedingung (3.13).

Aus der Abschätzung sehen wir, dass die Zufallsvariablen $X_i := I_A(Y_{i-1})$ die Annahme in (3.8) mit $c_n = 2(1-\delta)^{\lfloor n/r\rfloor}$ erfüllen. Wegen $\sum c_n < \infty$ können wir das Gesetz der großen Zahlen aus Satz 3.12 anwenden. Die Behauptung folgt dann wegen $S_n = \sum_{i=1}^n X_i$ und

$$E[X_i] \;=\; P[Y_{i-1} \in A] \;=\; \mu[A] \qquad \text{für alle } i \in \mathbb{N}. \qquad \blacksquare$$

4 Reelle Zufallsvariablen

4.1 Allgemeine Wahrscheinlichkeitsräume

Bisher haben wir uns noch nicht mit der Frage befasst, ob überhaupt ein Wahrscheinlichkeitsraum existiert, auf dem unendlich viele unabhängige Ereignisse bzw. Zufallsvariablen realisiert werden können. Auch die Realisierung einer auf einem endlichen reellen Intervall gleichverteilten Zufallsvariable auf einem geeigneten Wahrscheinlichkeitsraum haben wir noch nicht gezeigt. Die Existenz solcher Räume wurde stillschweigend vorausgesetzt.

Tatsächlich ist es oft nicht notwendig, den zugrunde liegenden Wahrscheinlichkeitsraum explizit zu kennen - die Kenntnis der gemeinsamen Verteilungen aller relevanten Zufallsvariablen genügt, um Wahrscheinlichkeiten und Erwartungswerte zu berechnen. Dennoch ist es an dieser Stelle hilfreich, die grundlegenden Existenzfragen zu klären, und unsere Modelle auf ein sicheres Fundament zu stellen. Die dabei entwickelten Begriffsbildungen werden sich beim Umgang mit stetigen und allgemeinen Zufallsvariablen als unverzichtbar erweisen.

Beispiele von Wahrscheinlichkeitsräumen

Wir beginnen mit einer Auflistung von verschiedenen Wahrscheinlichkeitsverteilungen. Während wir die ersten beiden Maße direkt auf der Potenzmenge $\mathcal{P}(\Omega)$ realisieren können, erfordert das Aufstellen eines geeigneten Wahrscheinlichkeitsraums in den nachfolgenden Beispielen zusätzliche Überlegungen.

Dirac-Maße.

Sei Ω beliebig und $a \in \Omega$ ein festes Element. Das **Dirac-Maß** in a ist die durch

$$\delta_a[A] \;:=\; I_A(a) \;=\; \begin{cases} 1 & \text{falls } a \in A, \\ 0 & \text{sonst,} \end{cases}$$

definierte Wahrscheinlichkeitsverteilung $P = \delta_a$ auf der σ-Algebra $\mathcal{A} = \mathcal{P}(\Omega)$. Dirac-Maße sind „deterministische Verteilungen" – es gilt

$$\delta_a[\{a\}] \;=\; 1.$$

Konvexkombinationen von Dirac-Maßen.

Ist C eine abzählbare Teilmenge von Ω, und $p : C \to [0,1]$ eine Gewichtsfunktion mit $\sum_{\omega \in C} p(\omega) = 1$, dann ist durch

$$P[A] \;=\; \sum_{a \in A \cap C} p(a) \;=\; \sum_{a \in C} p(a)\, \delta_a[A] \quad \forall A \subseteq \Omega.$$

eine eindeutige Wahrscheinlichkeitsverteilung P auf der Potenzmenge $\mathcal{A} = \mathcal{P}(\Omega)$ gegeben. Die Verteilung P ist „rein atomar", d.h. die Masse sitzt auf abzählbaren vielen „Atomen" (den Elementen von C). Jede diskrete Wahrscheinlichkeitsverteilung ist von dieser Form.

Unendliches Produktmodell (z.B. Münzwurffolge)

Mehrstufige diskrete Modelle mit endlich vielen Stufen können wir auf der Potenzmenge des Produkts $\Omega = \{(\omega_1, \ldots, \omega_n) : \omega_i \in \Omega_i\} = \Omega_1 \times \ldots \times \Omega_n$ der Grundräume Ω_i realisieren. Es stellt sich die Frage, ob wir auch unendlich viele Zufallsvariablen auf einem ähnlichen Produktraum realisieren können. Im einfachsten Fall möchten wir eine Folge unabhängiger fairer Münzwürfe (0-1-Experimente) auf dem Grundraum

$$\Omega \;=\; \{\omega = (\omega_1, \omega_2, \ldots) : \omega_i \in \{0,1\}\} \;=\; \{0,1\}^{\mathbb{N}}$$

modellieren. Ω ist überabzählbar, denn die Abbildung $X : [0,1) \to \Omega$, die einer reellen Zahl die Ziffernfolge ihrer Binärdarstellung zuordnet, ist injektiv. Diese Abbildung ist explizit gegeben durch $X(u) = (X_1(u),\, X_2(u),\, X_3(u),\, \ldots)$, wobei

$$X_n(u) \;=\; I_{D_n}(u) \qquad \text{mit} \qquad D_n = \bigcup_{i=1}^{2^{n-1}} [(2i-1)\cdot 2^{-n}, 2i \cdot 2^{-n}) \tag{4.1}$$

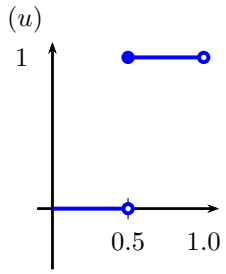


Abbildung 4.1: $X_1(u)$.

Wir suchen eine Wahrscheinlichkeitsverteilung P auf Ω, sodass

$$P[\{\omega \in \Omega : \omega_1 = a_1, \omega_2 = a_2, \ldots, \omega_n = a_n\}] = 2^{-n} \tag{4.2}$$

für alle $n \in \mathbb{N}$ und $a_1, \ldots, a_n \in \{0,1\}$ gilt. Gibt es eine σ-Algebra $\mathcal{A}$, die alle diese Ereignisse enthält, und eine eindeutige Wahrscheinlichkeitsverteilung P auf $\mathcal{A}$ mit (4.2) ?

Die Antwort ist positiv, wobei aber

(i) $\mathcal{A} \neq \mathcal{P}(\Omega)$ und

(ii) $P[\{\omega\}] = 0$ für alle $\omega \in \Omega$

gelten muss. Das entsprechende Produktmodell unterscheidet sich in dieser Hinsicht grundlegend von diskreten Modellen.

Kontinuierliche Gleichverteilung

Für die Gleichverteilung auf einem endlichen reellen Intervall $\Omega = (a,b)$ oder $\Omega = [a,b]$, $-\infty < a < b < \infty$, sollte gelten:

$$P[(c,d)] = P[[c,d]] = \frac{d-c}{b-a} \quad \forall\, a \le c < d \le b. \tag{4.3}$$

Gibt es eine σ-Algebra $\mathcal{B}$, die alle Teilintervalle von $[a,b]$ enthält, und eine Wahrscheinlichkeitsverteilung P auf $\mathcal{B}$ mit (4.3)?

Wieder ist die Antwort positiv, aber erneut gilt notwendigerweise $\mathcal{B} \neq \mathcal{P}(\Omega)$ und $P[\{\omega\}] = 0$ für alle $\omega \in \Omega$.

Tatsächlich sind die Probleme in den letzten beiden Abschnitten weitgehend äquivalent: die durch die Binärdarstellung (4.1) definierte Abbildung X ist eine Bijektion von $[0,1)$ nach $\{0,1\}^{\mathbb{N}} \setminus A$, wobei $A = \{\omega \in \Omega : \omega_n = 1 \text{ für alle hinreichend großen } n\}$ eine abzählbare Teilmenge ist. Eine Gleichverteilung auf $[0,1)$ wird durch X auf eine Münzwurffolge auf $\{0,1\}^{\mathbb{N}}$ abgebildet, und umgekehrt.

Um Wahrscheinlichkeitsverteilungen wie in den letzten beiden Beispielen zu konstruieren, benötigen wir zunächst geeignete σ-Algebren, die die relevanten Ereignisse bzw. Intervalle enthalten. Dazu verwenden wir die folgende Konstruktion:

Konstruktion von σ-Algebren

Sei Ω eine beliebige Menge, und $\mathcal{J} \subseteq \mathcal{P}(\Omega)$ eine Kollektion von Ereignissen, die auf jeden Fall in der zu konstruierenden σ-Algebra enthalten sein sollen, z.B. die Mengen in (4.2) bei unendlichen Produktmodellen, oder die reellen Intervalle im Fall kontinuierlicher Gleichverteilungen.

Definition 4.1 (Von einem Mengensystem erzeugte σ-Algebra). Die Kollektion

$$\sigma(\mathcal{J}) := \bigcap_{\substack{\mathcal{F} \supseteq \mathcal{J} \\ \mathcal{F} \text{ } \sigma\text{-Algebra auf } \Omega}} \mathcal{F}$$

von Teilmengen von Ω heißt *die von $\mathcal{J}$-erzeugte σ-Algebra*.

Bemerkung. Wie man leicht nachprüft (Übung), ist $\sigma(\mathcal{J})$ tatsächlich eine σ-Algebra, und damit die kleinste σ-Algebra, die $\mathcal{J}$ enthält.

Beispiel (Borelsche σ-Algebra auf $\mathbb{R}$). Sei $\Omega = \mathbb{R}$ und $\mathcal{J} = \{(s,t) : -\infty \leq s \leq t \leq \infty\}$ die Kollektion aller offenen Intervalle. Die von $\mathcal{J}$ erzeugte σ-Algebra

$$\mathcal{B}(\mathbb{R}) := \sigma(\mathcal{J})$$

heißt *Borelsche σ-Algebra* auf $\mathbb{R}$. Man prüft leicht nach, dass $\mathcal{B}(\mathbb{R})$ auch alle abgeschlossenen und halboffenen Intervalle enthält. Beispielsweise kann ein abgeschlossenes Intervall als Komplement der Vereinigung zweier offener Intervalle dargestellt werden. Die Borelsche σ-Algebra wird auch erzeugt von der Kollektion aller abgeschlossenen bzw. aller kompakten Intervalle. Ebenso gilt

$$\mathcal{B}(\mathbb{R}) \;=\; \sigma(\{(-\infty, c] : c \in \mathbb{R}\}).$$

Bemerkung (Nicht-Borelsche Mengen). Nicht jede Teilmenge von $\mathbb{R}$ ist in der Borelschen σ-Algebra $\mathcal{B}(\mathbb{R})$ enthalten. Es ist allerdings gar nicht so einfach, nicht-Borelsche Mengen anzugeben. Tatsächlich enthält $\mathcal{B}(\mathbb{R})$ so gut wie alle Teilmengen von $\mathbb{R}$, die in Anwendungsproblemen auftreten; z.B. alle offenen und abgeschlossenen Teilmengen von $\mathbb{R}$, sowie alle Mengen, die durch Bildung von abzählbar vielen Vereinigungen, Durchschnitten und Komplementbildungen daraus entstehen.

Beispiel (Produkt σ-Algebra auf $\{0,1\}^{\mathbb{N}}$). Auf dem Folgenraum

$$\Omega = \{0,1\}^{\mathbb{N}} = \{(\omega_1, \omega_2, \ldots) : \omega_i \in \{0,1\}\}$$

betrachten wir Teilmengen A von Ω von der Form

$$A = \{\omega \in \Omega : \omega_1 = a_1, \omega_2 = a_2, \ldots, \omega_n = a_n\}, \quad n \in \mathbb{N},\ a_1, \ldots, a_n \in \{0,1\}.$$

Im Beispiel unendlicher Produktmodelle von oben verwenden wir die von der Kollektion C aller dieser Mengen erzeugte σ-Algebra $\mathcal{A} = \sigma(C)$ auf $\{0,1\}^{\mathbb{N}}$. $\mathcal{A}$ heißt *Produkt-σ-Algebra* auf Ω.

Existenz und Eindeutigkeit von Wahrscheinlichkeitsverteilungen

Ein Mengensystem $\mathcal{J} \subseteq \mathcal{A}$ heißt *durchschnittsstabil*, falls für alle $A, B \in \mathcal{J}$ auch der Durchschnitt $A \cap B$ in $\mathcal{J}$ enthalten ist. Der folgende wichtige Satz zeigt, dass Wahrscheinlichkeitsverteilungen bereits durch die Wahrscheinlichkeiten aller Ereignisse aus einem durchschnittsstabilen Erzeugendensystem der σ-Algebra eindeutig fesgelegt sind.

Satz 4.2 (Eindeutigkeitssatz). Stimmen zwei Wahrscheinlichkeitsverteilungen P und $\widetilde{P}$ auf $(\Omega, \mathcal{A})$ überein auf einem **durchschnittsstabilen Mengensystem** $\mathcal{J} \subseteq \mathcal{A}$, so auch auf $\sigma(\mathcal{J})$.

Der Satz wird in der Vorlesung Einführung in die Wahrscheinlichkeitstheorie bewiesen.

Beispiel. (i) Eine Wahrscheinlichkeitsverteilung P auf $\mathcal{B}(\mathbb{R})$ ist eindeutig festgelegt durch die Wahrscheinlichkeiten $P[(-\infty, c]]$, $c \in \mathbb{R}$.

(ii) Die Wahrscheinlichkeitsverteilung P im Modell der unendlich vielen Münzwürfe ist eindeutig festgelegt durch die Wahrscheinlichkeiten der Ausgänge der ersten n Würfe für alle $n \in \mathbb{N}$.

Auch die folgende Aussage setzen wir hier ohne Beweis voraus. Sie folgt aus dem Eindeutigkeitssatz und dem Fortsetzungssatz von Carathéodory, siehe die Vorlesungen Analysis III und Einführung in die Wahrscheinlichkeitstheorie.

Satz 4.3 (Existenz und Eindeutigkeit der kontinuierlichen Gleichverteilung). Es existiert genau eine Wahrscheinlichkeitsverteilung $\mathrm{Unif}_{(0,1)}$ auf $\mathcal{B}((0,1))$ mit

$$\mathrm{Unif}(0,1)[(a,b)] = b - a \qquad \text{für alle } 0 < a \le b < 1. \tag{4.4}$$

Bemerkung (Lebesgue-Maß im $\mathbb{R}^d$). Auf ähnliche Weise folgt die Existenz und Eindeutigkeit des durch

$$\lambda[(a_1, b_1) \times \ldots \times (a_d, b_d)] = \prod_{i=1}^{d} (b_i - a_i) \qquad \text{für alle } a_i, b_i \in \mathbb{R} \text{ mit } a_i \le b_i$$

eindeutig festgelegten Lebesguemaßes λ auf der von den offenen Rechtecken erzeugten Borelschen σ-Algebra $\mathcal{B}(\mathbb{R}^d)$, siehe Analysis III.

4.2 Zufallsvariablen und ihre Verteilung

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Wir wollen nun Zufallsvariablen $X : \Omega \to S$ mit Werten in einem allgemeinen messbaren Raum $(S, \mathcal{B})$ betrachten. Beispielsweise ist $S = \mathbb{R}$ und $\mathcal{B}$ ist die Borelsche σ-Algebra. Oft interessieren uns die Wahrscheinlichkeiten von Ereignissen der Form

$$\{X \in B\} \;=\; \{\omega \in \Omega : X(\omega) \in B\} \;=\; X^{-1}(B),$$

„Der Wert der Zufallsgröße X liegt in B“

wobei $B \subseteq S$ eine Menge aus der σ-Algebra $\mathcal{B}$ auf dem Bildraum ist, also z.B. ein Intervall oder eine allgemeinere Borelmenge, falls $S = \mathbb{R}$ gilt.

Definition 4.4 (Zufallsvariable). Eine Abbildung $X : \Omega \to S$ heißt *messbar bzgl.* $\mathcal{A}/\mathcal{B}$, falls

$$X^{-1}(B) \in \mathcal{A} \qquad \text{für alle } B \in \mathcal{B}. \tag{4.5}$$

Eine *Zufallsvariable* ist eine auf einem Wahrscheinlichkeitsraum definierte messbare Abbildung.

Um Zufallsexperimente zu analysieren, müssen wir wissen, mit welchen Wahrscheinlichkeiten die relevanten Zufallsvariablen Werte in bestimmten Bereichen annehmen. Dies wird durch die Verteilung beschrieben. Seien $\mathcal{A}$ und $\mathcal{B}$ σ-Algebren auf den Mengen Ω und S.

Satz 4.5 (Bild einer Wahrscheinlichkeitsverteilung unter einer Zufallsvariable). Ist P eine Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$, und $X : \Omega \to S$ messbar bzgl. $\mathcal{A}/\mathcal{B}$, dann ist durch

$$\mu_X[B] := P[X \in B] = P[X^{-1}(B)] \qquad (B \in \mathcal{B})$$

eine Wahrscheinlichkeitsverteilung auf $(S, \mathcal{B})$ definiert.

Beweis. Es gilt $\mu_X[S] = P[X^{-1}(S)] = P[\Omega] = 1$. Sind $B_n \in \mathcal{B}$ $(n \in \mathbb{N})$ paarweise disjunkte Mengen, dann sind auch die Urbilder $X^{-1}(B_n)$ $(n \in \mathbb{N})$ paarweise disjunkt. Also gilt wegen der σ-Additivität von P:

$$\mu_X\left[\bigcup_n B_n\right] = P\left[X^{-1}\left(\bigcup_n B_n\right)\right] = P\left[\bigcup_n X^{-1}(B_n)\right] = \sum_n P[X^{-1}(B_n)] = \sum_n \mu_X[B_n].$$

Somit ist μ_X eine Wahrscheinlichkeitsverteilung. ■

Definition 4.6 (Verteilung einer Zufallsvariable). Die Wahrscheinlichkeitsverteilung μ_X auf $(S, \mathcal{B})$ heißt *Verteilung (law) von X unter P*.

Für μ_X werden häufig auch die folgenden Notationen verwendet:

$$\mu_X = P \circ X^{-1} = \mathcal{L}_X = P_X = X(P)$$

Stimmt die σ-Algebra $\mathcal{B}$ auf dem Bildraum nicht mit der Potenzmenge $\mathcal{P}(S)$ überein, dann ist es meist schwierig, die Bedingung (4.5) für *alle* Mengen $B \in \mathcal{B}$ explizit zu zeigen. Aufgrund des folgenden Lemmas reicht es aber aus, die Bedingung (4.5) für alle Mengen aus einem Erzeugendensystem $\mathcal{J}$ der σ-Algebra $\mathcal{B}$ zu überprüfen.

Lemma 4.7. *Sei $\mathcal{J} \subseteq \mathcal{P}(S)$ ein Mengensystem mit $\mathcal{B} = \sigma(\mathcal{J})$. Gilt $X^{-1}(B) \in \mathcal{A}$ für alle Mengen $B \in \mathcal{J}$, dann ist X eine Zufallsvariable. Ist das Mengensystem $\mathcal{J}$ außerdem* durchschnittsstabil *(d.h. für $A, B \in \mathcal{J}$ ist auch $A \cap B$ in $\mathcal{J}$ enthalten), dann ist die Verteilung von X bereits durch die Wahrscheinlichkeiten*

$$\mu_X[B] = P[X \in B] = P[X^{-1}(B)] \qquad (B \in \mathcal{J})$$

eindeutig festgelegt.

Beweis. Das Mengensystem $\{B \in \mathcal{B} : X^{-1}(B) \in \mathcal{A}\}$ ist eine σ-Algebra, wie man leicht nachprüft. Diese σ-Algebra enthält $\mathcal{J}$ nach Voraussetzung, also enthält sie auch die von $\mathcal{J}$ erzeugte σ-Algebra $\mathcal{B}$. Somit ist die Bedingung (4.5) erfüllt, d.h., X ist eine Zufallsvariable. Der zweite Teil der Aussage folgt nun aus dem Eindeutigkeitssatz 4.2. ■

Wir werden gleich sehen, wie wir mit Hilfe des Lemmas reellwertige Zufallsvariablen und deren Verteilung auf einfache Weise beschreiben können. Zuvor vervollständigen wir die bereits oben skizzierte Konstruktion des unendlichen Münzwurfmodells aus der Gleichverteilung auf dem Intervall $(0, 1)$.

Korollar 4.8 (Existenz und Eindeutigkeit des unendlichen Münzwurfmodells). Es existiert genau eine Wahrscheinlichkeitsverteilung P auf dem unendlichen Produktraum $\Omega = \{0, 1\}^{\mathbb{N}}$ mit Produkt σ-Algebra, sodass

$$P[\{\omega \in \Omega : \omega_1 = a_1, \omega_2 = a_2, \ldots, \omega_n = a_n\}] = 2^{-n} \qquad \text{für alle } n \in \mathbb{N} \text{ und } a_1, \ldots, a_n \in \{0, 1\}. \tag{4.6}$$

Beweis. Wir betrachten den Wahrscheinlichkeitsraum $((0, 1), \mathcal{B}((0, 1)), \mathrm{Unif}(0, 1))$. Aus Lemma 4.7 folgt, dass die durch (4.1) definierte Abbildung $X(u) = (X_1(u), X_2(u), X_3(u), \ldots)$, die einer reellen Zahl $u \in (0, 1)$ die Ziffernfolge ihrer Binärdarstellung zuordnet, eine Zufallsvariable mit Werten im Produktraum $\Omega = \{0, 1\}^{\mathbb{N}}$ ist. Die Verteilung P dieser Zufallsvariable erfüllt die Bedingung (4.6), denn das Intervall aller reellen Zahlen $u \in (0, 1)$ deren Binärdarstellung mit einer bestimmten Ziffernfolge $0.a_1a_2 \ldots a_n$ beginnt, hat die Länge 2^{-n}. Damit haben wir die Existenz gezeigt. Die Eindeutigkeit folgt aus dem Eindeutigkeitssatz 4.2. ■

Reellwertige Zufallsvariablen; Verteilungsfunktion

Aus Lemma 4.7 folgt sofort, dass Definition 4.4 für diskrete Zufallsvariablen konsistent mit unserer Definition aus dem ersten Kapitel ist. Für reellwertige Zufallsvariablen ergibt sich ebenfalls eine einfache Charakterisierung.

Korollar 4.9 (Diskrete und reellwertige Zufallsvariablen). Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum.

(i) Ist S eine abzählbare Menge mit σ-Algebra $\mathcal{B} = \mathcal{P}(S)$, dann ist eine Abbildung $X : \Omega \to S$ genau dann eine Zufallsvariable, wenn

$$\{X = a\} = \{\omega \in \Omega : X(\omega) = a\} \in \mathcal{A} \quad \forall\, a \in S.$$

Die Verteilung von X ist eindeutig durch die Wahrscheinlichkeiten dieser Ereignisse bestimmt.

(ii) Eine Abbildung $X : \Omega \to \mathbb{R}$ ist genau dann eine Zufallsvariable bzgl. der Borelschen σ-Algebra, wenn

$$\{X \le c\} \;=\; \{\omega \in \Omega : X(\omega) \le c\} \in \mathcal{A} \quad \forall\, c \in \mathbb{R}.$$

Die Verteilung von X ist eindeutig durch die Wahrscheinlichkeiten dieser Ereignisse bestimmt.

Beweis. (i) Es gilt $\{X = a\} = X^{-1}(\{a\})$. Die einelementigen Mengen erzeugen die σ-Algebra $\mathcal{P}(S)$, da jede Teilmenge einer abzählbaren Menge S eine abzählbare Vereinigung von einelementigen Mengen ist.

(ii) Es gilt $\{X \le c\} = X^{-1}((-\infty, c])$. Die Intervalle $(-\infty, c]$ $(c \in \mathbb{R})$ erzeugen $\mathcal{B}(\mathbb{R})$, also folgt die Aussage aus Lemma 4.7. ■

Die Verteilung μ_X einer Zufallsvariable X mit abzählbarem Wertebereich S ist eindeutig durch die Massenfunktion

$$p_X(a) = P[X = a] = \mu_X[\{a\}] \qquad (a \in S)$$

festgelegt. Die Verteilung μ_X einer reellwertigen Zufallsvariablen $X : \Omega \to \mathbb{R}$ ist eine Wahrscheinlichkeitsverteilung auf $\mathcal{B}(\mathbb{R})$. Sie ist eindeutig festgelegt durch die Wahrscheinlichkeiten

$$\mu_X[(-\infty, c]] = P[X \le c] \qquad (c \in \mathbb{R})$$

festgelegt, da die Intervalle $(-\infty, c], c \in \mathbb{R}$, ein durchschnittsstabiles Erzeugendensystem der Borelschen σ-Algebra bilden.

Definition 4.10 (Verteilungsfunktion). Die Funktion $F_X : \mathbb{R} \to [0, 1]$,

$$F_X(c) := P[X \le c] = \mu_X[(-\infty, c]]$$

heißt *Verteilungsfunktion (distribution function)* der Zufallsvariable $X : \Omega \to \mathbb{R}$ bzw. der Wahrscheinlichkeitsverteilung μ_X auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Der folgende Satz nennt einige grundlegende Eigenschaften der Verteilungsfunktion einer auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definierten Zufallsvariable $X : \Omega \to \mathbb{R}$. Wir werden in Abschnitt 4.5 sehen, dass umgekehrt jede Funktion mit den Eigenschaften (i)-(iii) aus Satz 4.11 die Verteilungsfunktion einer reellen Zufallsvariable ist.

Satz 4.11 (Eigenschaften der Verteilungsfunktion).
Für die Verteilungsfunktion $F_X : \mathbb{R} \to [0, 1]$ einer reellwertigen Zufallsvariable X 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 ist *rechtsstetig*, d.h. $F_X(c) = \lim_{y \searrow c} F_X(y)$ für alle $c \in \mathbb{R}$,

(iv) $F_X(c) = \lim_{y \nearrow c} F_X(y) + \mu_X[\{c\}]$.
Insbesondere ist F_X genau dann stetig bei c, wenn $\mu_X[\{c\}] = 0$ gilt.

Beweis. Die Aussagen folgen unmittelbar aus der monotonen Stetigkeit und Normiertheit der zugrundeliegenden Wahrscheinlichkeitsverteilung P. Der Beweis der Eigenschaften (i)-(iii) wird dem Leser als Übung überlassen. Zum Beweis von (iv) bemerken wir, dass für $y < c$ gilt:

$$F_X(c) - F_X(y) = P[X \le c] - P[X \le y] = P[y < X \le c].$$

Für eine monoton wachsende Folge $y_n \nearrow c$ erhalten wir daher aufgrund der monotonen Stetigkeit von P:

$$\begin{aligned} F_X(c) - \lim_{n \to \infty} F_X(y_n) &= \lim_{n \to \infty} P[y_n < X \le c] = P\left[\bigcap_n \{y_n < X \le c\}\right] \\ &= P[X = c] = \mu_X[\{c\}]. \end{aligned}$$

Da dies für alle Folgen $y_n \nearrow c$ gilt, folgt die Behauptung. ∎

Diskrete und absolutstetige Verteilungen

Nach Satz 4.11 (iv) sind die Unstetigkeitsstellen der Verteilungsfunktion F_X einer reellwertigen Zufallsvariable X gerade die *Atome* der Verteilung, d.h. die $c \in \mathbb{R}$ mit $\mu_X[\{c\}] > 0$. Nimmt X nur endlich viele Werte in einem Intervall I an, dann ist F auf I stückweise konstant, und springt nur bei diesen Werten. Allgemeiner definieren wir:

Definition 4.12 (Diskrete und absolutstetige Verteilung; Dichtefunktion). (i) Die Verteilung μ_X einer reellwertigen Zufallsvariable X heißt *diskret*, falls $\mu_X[A] = 1$ für eine abzählbare Menge A gilt, d.h., falls die Verteilungsfunktion gegeben ist durch

$$F_X(c) \;=\; \mu_X[(-\infty, c]] \;=\; \sum_{\substack{a \in A \\ a \le c}} \mu_X[\{a\}] \qquad \text{für alle } c \in \mathbb{R}.$$

(ii) Die Verteilung μ_X heißt *absolutstetig* (oder auch kurz *stetig*), falls eine integrierbare Funktion $f_X : \mathbb{R} \to [0, \infty)$ existiert mit

$$F_X(c) \;=\; \mu_X[(-\infty, c]] \;=\; \int_{-\infty}^{c} f_X(x)\, dx \qquad \text{für alle } c \in \mathbb{R}. \tag{4.7}$$

(iii) Eine integrierbare Funktion $f_X : \mathbb{R} \to [0, \infty)$ mit (4.7) heißt *Dichtefunktion* der Zufallsvariable X bzw. der Verteilung μ_X.

Das Integral in (4.7) ist dabei im Allgemeinen als Lebesgueintegral zu interpretieren, siehe Analysis III. Ist die Funktion f_X stetig, dann stimmt dieses mit dem Riemannintegral überein. Da μ_X eine Wahrscheinlichkeitsverteilung ist, folgt, dass f_X eine *Wahrscheinlichkeitsdichte* ist, d.h. $f_X \ge 0$ und

$$\int_{\mathbb{R}} f_X(x)\, dx \;=\; 1.$$

Bemerkung. (i) Nach dem Hauptsatz der Differential- und Integralrechnung gilt

$$F_X'(x) \;=\; f_X(x) \tag{4.8}$$

für alle $x \in \mathbb{R}$, falls f stetig ist. Im Allgemeinen gilt (4.8) für λ-fast alle x, wobei λ das Lebesguemaß auf $\mathbb{R}$ ist.

(ii) Aus (4.7) folgt aufgrund der Eigenschaften des Lebesgueintegrals:

$$P[X \in B] = \mu_X[B] = \int_B f_X(x)\, dx, \tag{4.9}$$

für alle Mengen $B \in \mathcal{B}(\mathbb{R})$. Zum Beweis zeigt man, dass beide Seiten von (4.9) Wahrscheinlichkeitsverteilungen definieren, und wendet den Eindeutigkeitssatz an.

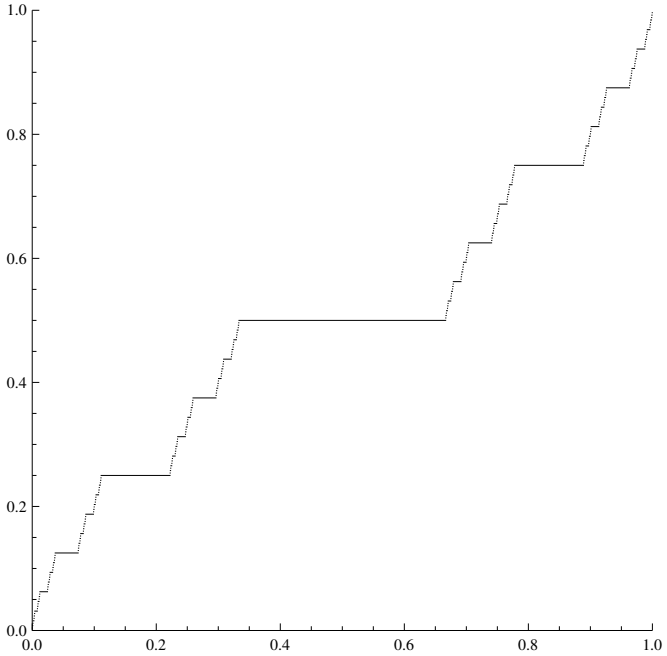
Beispiele (Diskrete und absolutstetige Verteilungen). (i) Ist X deterministisch mit $P[X = a] = 1$ für ein $a \in \mathbb{R}$, dann ist die Verteilung von X das Dirac-Maß δ_a. Das Dirac-Maß ist diskret mit Verteilungsfunktion $F_X(c) = I_{[a,\infty)}(c)$.

(ii) Die Gleichverteilung $\mathrm{Unif}_{(0,1)}$ ist eine absloutstetige Verteilung mit Verteilungsfunktion $F(c) = 0$ für $c \le 0$, $F(c) = c$ für $c \in [0,1]$, und $F(c) = 1$ für $c \ge 1$, und Dichtefunktion $f(x) = I_{(0,1)}(x)$.

(iii) Die Wahrscheinlichkeitsverteilung $\frac{1}{2}\delta_a + \frac{1}{2}\mathrm{Unif}_{(0,1)}$ ist weder absolutstetig noch diskret.

Die Verteilung aus dem letzten Beispiel ist zwar weder absolutstetig noch diskret, sie kann aber sofort zerlegt werden in einen absolutstetigen und einen diskreten Anteil. Für die im folgenden Beispiel betrachtete Gleichverteilung auf der Cantor-Menge existiert keine solche Zerlegung:

Beispiel (Devil's staircase). Wir betrachten die wie folgt definierte Verteilungsfunktion F einer Wahrscheinlichkeitsverteilung auf dem Intervall $(0,1)$: $F(c) = 0$ für $c \le 0$, $F(c) = 1$ für $c \ge 1$, $F(c) = 1/2$ für $c \in [1/3, 2/3]$, $F(c) = 1/4$ für $c \in [1/9, 2/9]$, $F(c) = 3/4$ für $c \in [7/9, 8/9]$, $F(c) = 1/8$ für $c \in [1/27, 2/27]$, usw. Man überzeugt sich leicht, dass auf diese Weise eine eindeutige *stetige* monotone wachsende Funktion $F : \mathbb{R} \to [0,1]$ definiert ist. Nach Satz 4.26 unten ist F also die Verteilungsfunktion einer Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}$.



Da F stetig ist, hat das Maß μ *keinen diskreten Anteil*, d.h. $\mu[\{a\}] = 0$ für alle $a \in \mathbb{R}$. Da F auf den Intervallen $[1/3, 2/3]$, $[1/9, 2/9]$, $[7/9, 8/9]$ usw. jeweils konstant ist, sind alle diese Intervalle μ-Nullmengen. Die Verteilung μ sitzt also auf dem Komplement

$$C = \left\{ \sum_{i=1}^{\infty} a_i 3^{-i} : a_i \in \{0,2\} (i \in \mathbb{N}) \right\}$$

der Vereinigung der Intervalle. Die *Cantor-Menge* C ist ein Fraktal, das aus den reellen Zahlen zwischen 0 und 1 besteht, die sich im Dreier-System ohne Verwendung der Ziffer 1 darstellen lassen. Sie ist eine Lebesgue-Nullmenge, aber der Träger des Maßes μ. Da F der Grenzwert der Verteilungsfunktionen von Gleichverteilungen auf den Mengen $C_1 = (0,1)$, $C_2 = (0,1/3) \cup (2/3,1)$, $C_3 = (0,1/9) \cup (2/9,1/3) \cup (2/3,7/9) \cup (8/9,1)$,... ist, können wir μ als *Gleichverteilung auf der Cantor-Menge* $C = \bigcap C_n$ interpretieren.

Weiterhin ist die Verteilungsfunktion F auf den oben betrachteten Intervallen konstant, und daher Lebesgue-fast überall differenzierbar mit Ableitung $F'(x) = 0$. Die „Teufelstreppe" F wächst also von 0 auf 1, obwohl sie stetig ist und ihre Ableitung fast überall gleich Null ist ! Hieraus folgt, dass die Verteilung μ *nicht absolutstetig* sein kann, denn in diesem Fall wäre F gleich dem Integral der Lebesgue-fast überall definierten Funktion F', also konstant.

4.3 Spezielle Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$

Im Folgenden betrachten wir einige wichtige Beispiele von eindimensionalen Verteilungen und ihren Verteilungsfunktionen. Wir geben zunächst die Verteilungsfunktion für einige elementare diskrete Verteilungen an, und betrachten dann kontinuierliche Analoga dieser Verteilungen.

Diskrete Verteilungen

Beispiele. (i) Gleichverteilung auf einer endlichen Menge $\{a_1, \ldots, a_n\} \subset \mathbb{R}$:

Ist $S = \{a_1, \ldots, a_n\} \subset \mathbb{R}$ eine n-elementige Teilmenge von $\mathbb{R}$, dann ist die Gleichverteilung μ auf S durch

$$\mu = \frac{1}{n} \sum_{i=1}^{n} \delta_{a_i} \tag{4.10}$$

gegeben. Der Wert der Verteilungsfunktion F von μ springt an jeder der Stellen $a_1, \ldots, a_n$ um $1/n$ nach oben. Die *empirische Verteilung* von $a_1, \ldots, a_n$ ist ebenfalls durch (4.10) definiert, wobei hier aber nicht vorausgesetzt wird, dass $a_1, \ldots, a_n$ verschieden sind. Die Sprunghöhen der empirischen Verteilungsfunktion sind dementsprechend Vielfache von $1/n$.

(ii) Geometrische Verteilung mit Parameter $p \in [0, 1]$: Hier ist

$$\mu[\{k\}] = (1-p)^{k-1} \cdot p \quad \text{für } k \in \mathbb{N}.$$

Für eine geometrisch verteilte Zufallsvariable T gilt:

$$F(c) = P[T \le c] = 1 - \underbrace{P[T > c]}_{=P[T > \lfloor c \rfloor]} = 1 - (1-p)^{\lfloor c \rfloor} \quad \text{für } c \ge 0,$$

wobei $\lfloor c \rfloor := \max\{n \in \mathbb{Z} : n \le c\}$ der ganzzahlige Anteil von c ist.

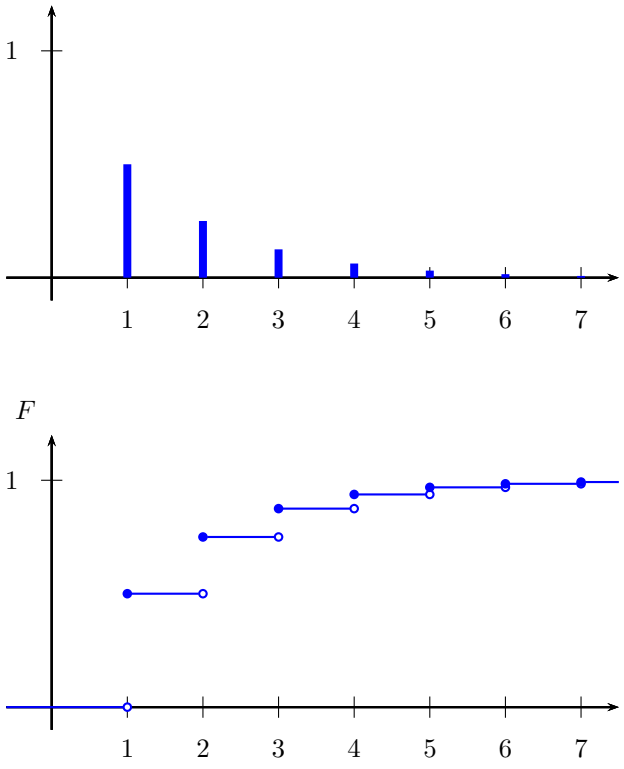


Abbildung 4.2: Massen- und Verteilungsfunktion einer Geom(1/2)-verteilten Zufallsvariable.

(iii) Binomialverteilung mit Parametern n und p: Die Massenfunktion ist

$$\mu[\{k\}] = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{für} \quad k = 0, 1, \ldots, n.$$

Somit ist die Verteilungsfunktion gegeben durch $F(c) = \sum_{k=0}^{\lfloor c \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$.

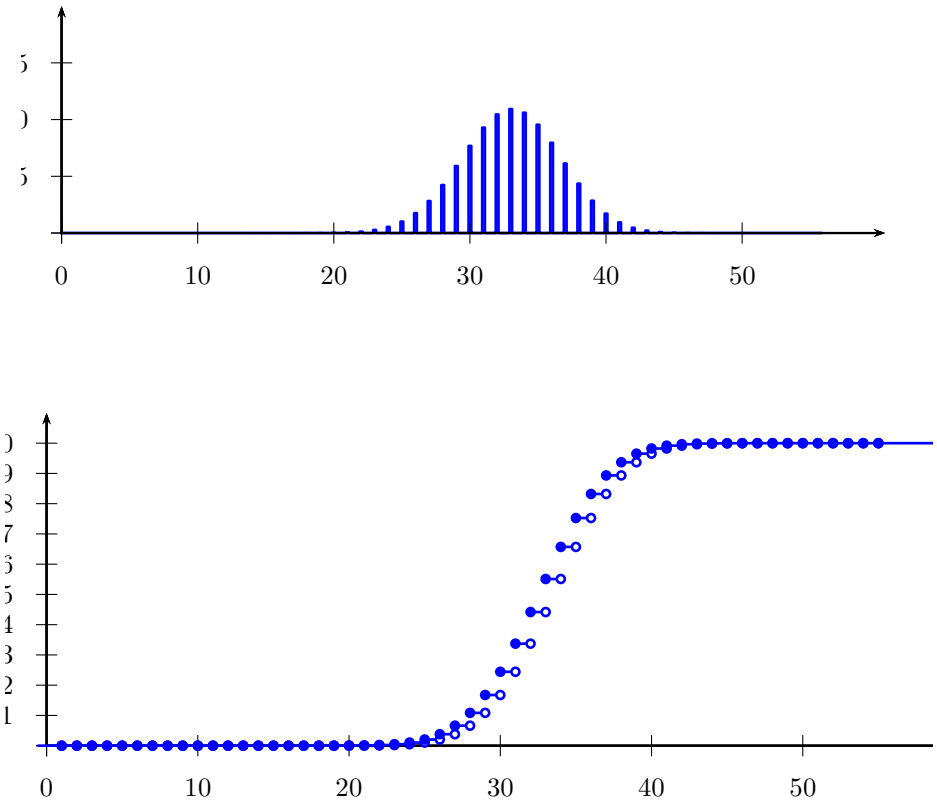


Abbildung 4.3: Massen- und Verteilungsfunktion von Bin(55, 0.6)

Kontinuierliche Gleichverteilung

Seien $a, b \in \mathbb{R}$ mit $a < b$. Eine Zufallsvariable $X : \Omega \to \mathbb{R}$ ist gleichverteilt auf dem Intervall (a, b), falls

$$P[X \le c] \;=\; \mathrm{Unif}_{(a,b)}[(a,c)] \;=\; \frac{c-a}{b-a} \qquad \text{für alle } c \in (a,b)$$

gilt. Eine auf $(0, 1)$ gleichverteilte Zufallsvariable ist zum Beispiel die Identität

$$U(\omega) = \omega$$

auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P) = ((0,1), \mathcal{B}((0,1)), \mathrm{Unif}_{(0,1)})$. Ist U gleichverteilt auf $(0, 1)$, dann ist die Zufallsvariable

$$X(\omega) = a + (b-a)U(\omega)$$

gleichverteilt auf (a, b).

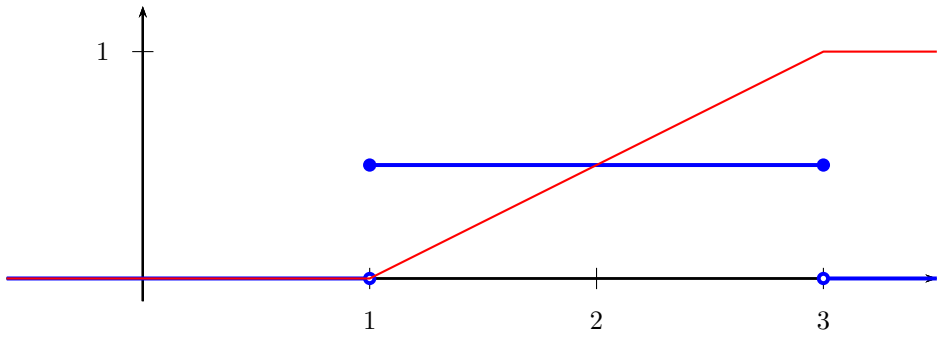


Abbildung 4.4: Dichte und Verteilungsfunktion der Gleichverteilung auf dem Intervall $(1,3)$.

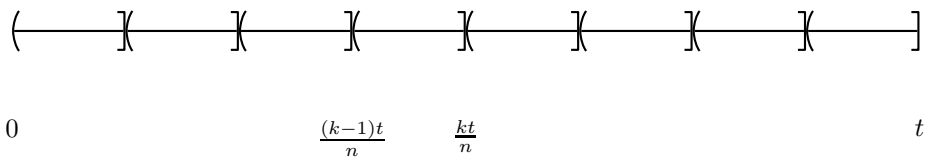
Die Dichte und Verteilungsfunktion der Verteilung $\mathrm{Unif}_{(a,b)}$ sind gegeben durch

$$f(x) = \frac{1}{b-a} I_{(a,b)}(x), \qquad F(c) = \begin{cases} 0 & \text{für } c \le a, \\ \frac{c-a}{b-a} & \text{für } a \le c \le b, \\ 1 & \text{für } c \ge b. \end{cases}$$

Affine Funktionen von gleichverteilten Zufallsvariablen sind wieder gleichverteilt.

Exponentialverteilung

Das kontinuierliche Analogon zur geometrischen Verteilung ist die Exponentialverteilung. Angenommen, wir wollen die Wartezeit auf das erste Eintreten eines unvorhersehbaren Ereignisses (z.B. radioaktiver Zerfall) mithilfe einer Zufallsvariable $T : \Omega \to (0,\infty)$ beschreiben. Wir überlegen uns zunächst, welche Verteilung zur Modellierung einer solchen Situation angemessen sein könnte. Um die Wahrscheinlichkeit $P[T > t]$ zu approximieren, unterteilen wir das Zeitintervall $(0,t]$ in eine große Anzahl $n \in \mathbb{N}$ von gleich großen Intervallen $(\frac{(k-1)t}{n}, \frac{kt}{n}], 1 \le k \le n$.



Sei A_k das Ereignis, dass das unvorhersehbare Geschehen im Zeitraum $(\frac{(k-1)t}{n}, \frac{kt}{n}]$ eintritt. Ein naheliegender Modellierungsansatz ist anzunehmen, dass die Ereignisse A_k unabhängig sind mit Wahrscheinlichkeit

$$P[A_k] \approx \lambda \frac{t}{n},$$

wobei $\lambda > 0$ die „Intensität“, d.h. die mittlere Häufigkeit des Geschehens pro Zeiteinheit, beschreibt, und die Approximation für $n \to \infty$ immer genauer wird. Damit erhalten wir:

$$P[T > t] = P[A_1^C \cap \ldots \cap A_n^C] \approx \left(1 - \frac{\lambda t}{n}\right)^n \quad \text{für großes } n.$$

Für $n \to \infty$ konvergiert die rechte Seite gegen $e^{-\lambda t}$. Daher liegt folgende Definition nahe:

Definition 4.13 (Exponentialverteilung). Eine Zufallsvariable $T : \Omega \to [0, \infty)$ heißt *exponentialverteilt zum Parameter* λ>0, falls

$$P[T > t] \;=\; e^{-\lambda t} \qquad \text{für alle } t \geq 0 \text{ gilt.}$$

Die *Exponentialverteilung zum Parameter* λ ist dementsprechend die Wahrscheinlichkeitsverteilung $\mu = \mathrm{Exp}(\lambda)$ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit

$$\mu[(t, \infty)] = e^{-\lambda t} \qquad \text{für alle } t \geq 0,$$

bzw. mit Verteilungsfunktion

$$F(t) = \mu[(-\infty, t]] = \begin{cases} 1 - e^{-\lambda t} & \text{für } t \geq 0, \\ 0 & \text{für } t < 0. \end{cases} \tag{4.11}$$

Nach dem Eindeutigkeitssatz ist die $\mathrm{Exp}(\lambda)$-Verteilung durch (4.11) eindeutig festgelegt.

Die Dichte der Exponentialverteilung mit Parameter $\lambda > 0$ ist gegeben durch

$$f(t) \;=\; \lambda e^{-\lambda t} I_{(0,\infty)}(t).$$

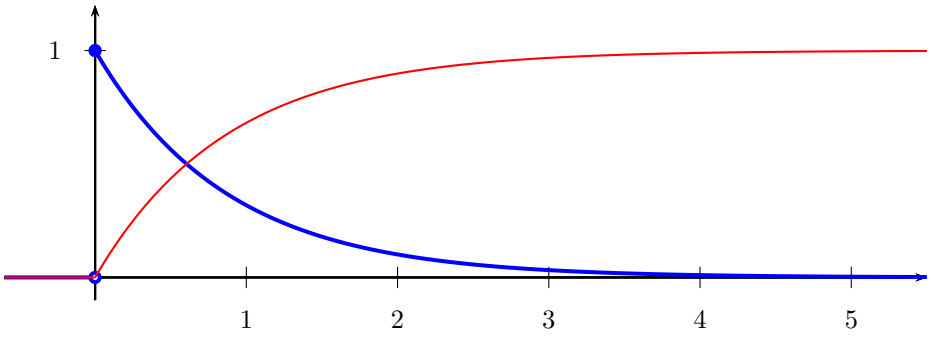


Abbildung 4.5: Dichte und Verteilungsfunktion der Exponentialverteilung Exp(1).

Ist T eine exponentialverteilte Zufallsvariable zum Parameter λ, und $a > 0$, dann ist aT exponentialverteilt zum Parameter λ/a, denn

$$P[aT > c] \;=\; P[T > c/a] \;=\; \exp(-\lambda c/a) \qquad \text{für alle } c \geq 0.$$

Eine bemerkenswerte Eigenschaft exponentialverteilter Zufallsvariablen ist die „Gedächtnislosigkeit”:

Satz 4.14 (Gedächtnislosigkeit der Exponentialverteilung). Ist T exponentialverteilt, dann gilt für alle $s, t \geq 0$:

$$P[T - s > t | T > s] \;=\; P[T > t].$$

Hierbei ist $T - s$ die verbleibende Wartezeit auf das erste Eintreten des Ereignisses. Also: *Auch wenn man schon sehr lange vergeblich gewartet hat, liegt das nächste Ereignis nicht näher als am Anfang!*

Beweis. Für $s, t \geq 0$ gilt

$$P[T - s > t | T > s] = \frac{P[T - s > t \text{ und } T > s]}{P[T > s]} = \frac{P[T > s + t]}{T > s} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P[T > t].$$

Wir konstruieren nun explizit eine exponentialverteilte Zufallsvariable aus einer gleichverteilten Zufallsvariable. Dazu bemerken wir, dass $T : \Omega \to \mathbb{R}$ genau dann exponentialverteilt mit Parameter λ ist, wenn

$$P[e^{-\lambda T} < u] \;=\; P\left[T > -\frac{1}{\lambda} \log u\right] \;=\; e^{\frac{\lambda}{\lambda} \log u} \;=\; u$$

für alle $u \in (0, 1)$ gilt, d.h. wenn $e^{-\lambda T}$ auf $(0, 1)$ gleichverteilt ist. Also können wir eine exponentialverteilte Zufallsvariable erhalten, indem wir umgekehrt

$$T \;:=\; -\frac{1}{\lambda} \log U \qquad \text{mit} \qquad U \sim \text{Unif}_{(0,1)}$$

setzen. Insbesondere ergibt sich die folgende Methode zur Simulation einer exponentialverteilten Zufallsvariable:

Algorithmus 1: Simulation einer Stichprobe t von $\text{Exp}(\lambda)$

Input : Intensität $\lambda > 0$
Output Stichprobe t von $\text{Exp}(\lambda)$
:
1 Erzeuge $u \sim \text{Unif}_{(0,1)}$.;
2 Setze $t := -\frac{1}{\lambda} \log u$.;
3 **return** t;

Wir werden in Abschnitt 4.5 zeigen, dass mit einem ähnlichen Verfahren beliebige reelle Zufallsvariablen konstruiert und simuliert werden können.

Normalverteilungen

Wegen $\int_{-\infty}^{\infty} e^{-z^2/2} dz = \sqrt{2\pi}$ ist die „Gaußsche Glockenkurve“

$$f(z) \;=\; \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad z \in \mathbb{R},$$

eine Wahrscheinlichkeitsdichte. Eine stetige Zufallsvariable Z mit Dichtefunktion f heißt *standardnormalverteilt*. Die Verteilungsfunktion

$$\Phi(c) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz$$

der Standardnormalverteilung ist im Allgemeinen nicht explizit berechenbar. Ist Z standardnormalverteilt, und

$$X(\omega) = \sigma Z(\omega) + m$$

mit $\sigma > 0, m \in \mathbb{R}$, dann ist X eine Zufallsvariable mit Verteilungsfunktion

$$F_X(c) \;=\; P[X \leq c] \;=\; P\left[Z \leq \frac{c - m}{\sigma}\right] \;=\; \Phi\left(\frac{c - m}{\sigma}\right).$$

Mithilfe der Substitution $z = \frac{x-m}{\sigma}$ erhalten wir

$$F_X(c) \;=\; \int_{-\infty}^{\frac{c-m}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz \;=\; \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} \, dx.$$

Definition 4.15 (Normalverteilung). Die Wahrscheinlichkeitsverteilung $N(m, \sigma^2)$ auf $\mathbb{R}$ mit Dichtefunktion

$$f_{m,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

heißt *Normalverteilung mit Mittel m und Varianz* σ^2. Die Verteilung $N(0,1)$ heißt *Standardnormalverteilung*.

Wir werden im nächsten Abschnitt sehen, dass die Binomialverteilung (also die Verteilung der Anzahl der Erfolge bei unabhängigen 0-1-Experimenten mit Erfolgswahrscheinlichkeit p) für große n näherungsweise durch eine Normalverteilung beschrieben werden kann. Entsprechendes gilt viel allgemeiner für die Verteilungen von Summen vieler kleiner unabhängiger Zufallsvariablen (*Zentraler Grenzwertsatz*, s.u.).

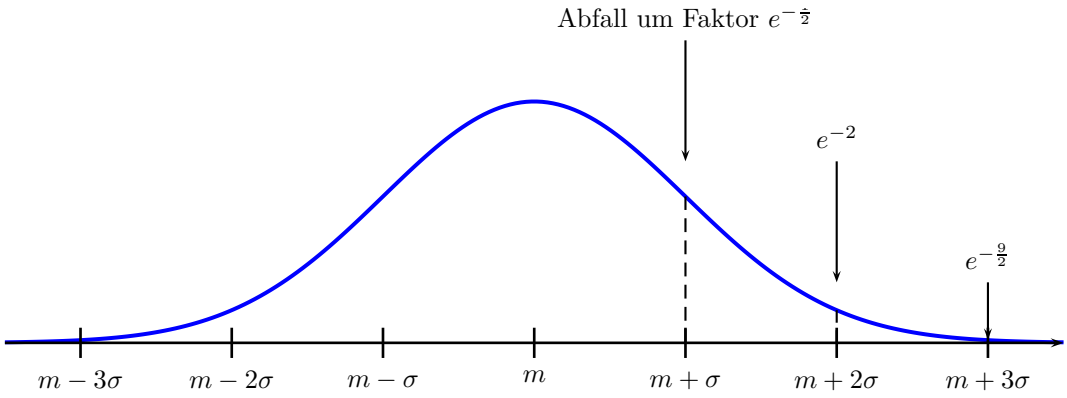


Abbildung 4.6: Dichte der Normalverteilung mit Mittelwert m und Varianz σ^2.

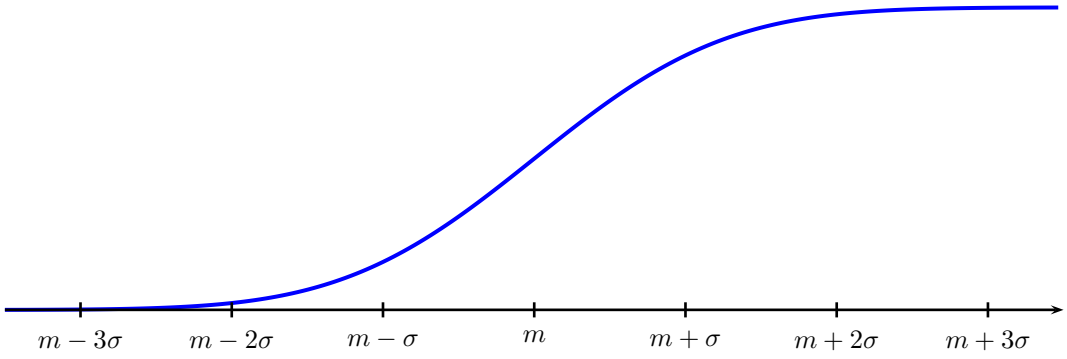


Abbildung 4.7: Verteilungsfunktion der Normalverteilung mit Mittelwert m und Varianz σ^2.

Die Dichte der Normalverteilung ist an der Stelle m maximal, und klingt außerhalb einer σ-Umgebung von

m rasch ab. Beispielsweise gilt

$$f_{m,\sigma}(m \pm \sigma) = \frac{f_{m,\sigma}(m)}{\sqrt{e}}, \quad f_{m,\sigma}(m \pm 2\sigma) = \frac{f_{m,\sigma}(m)}{e^2}, \quad f_{m,\sigma}(m \pm 3\sigma) = \frac{f_{m,\sigma}(m)}{e^{9/2}}.$$

Für die Wahrscheinlichkeit, dass eine normalverteilte Zufallsvariable Werte außerhalb der σ-, 2σ- und 3σ-Umgebungen annimmt, erhält man:

$$\begin{aligned} P[|X-m| > k\sigma] &= P\left[\left|\frac{X-m}{\sigma}\right| > k\right] = P[|Z| > k] = 2P[Z > k] = 2(1-\Phi(k)) \\ &\approx \begin{cases} 31.7\% & \text{für } k = 1, \\ 4.6\% & \text{für } k = 2, \\ 0.26\% & \text{für } k = 3. \end{cases} \end{aligned}$$

Eine Abweichung der Größe σ vom Mittelwert m ist also für eine normalverteilte Zufallsvariable relativ typisch, eine Abweichung der Größe 3σ dagegen schon sehr selten.

Die folgenden expliziten Abschätzungen für die Wahrscheinlichkeiten großer Werte sind oft nützlich:

Lemma 4.16. *Für eine standardnormalverteilte Zufallsvariable Z gilt:*

$$(2\pi)^{-1/2} \cdot \left(\frac{1}{y} - \frac{1}{y^3}\right) \cdot e^{-y^2/2} \quad \leq \quad P[Z \geq y] \quad \leq \quad (2\pi)^{-1/2} \cdot \frac{1}{y} \cdot e^{-y^2/2} \qquad \forall y > 0.$$

Beweis. Es gilt:

$$P[Z \geq y] = (2\pi)^{-1/2} \int_y^\infty e^{-z^2/2}\, dz$$

Um das Integral abzuschätzen, versuchen wir approximative Stammfunktionen zu finden. Zunächst gilt:

$$\frac{d}{dz}\left(-\frac{1}{z}e^{-z^2/2}\right) = \left(1 + \frac{1}{z^2}\right) \cdot e^{-z^2/2} \geq e^{-z^2/2} \qquad \forall z \geq 0,$$

also

$$\frac{1}{y}e^{-y^2/2} = \int_y^\infty \frac{d}{dz}\left(-\frac{1}{z}e^{-z^2/2}\right) \geq \int_y^\infty e^{-z^2/2}\, dz,$$

woraus die obere Schranke für $P[Z \geq y]$ folgt.

Für die untere Schranke approximieren wir die Stammfunktion noch etwas genauer. Es gilt

$$\frac{d}{dz}\left(\left(-\frac{1}{z} + \frac{1}{z^3}\right)e^{-z^2/2}\right) = \left(1 + \frac{1}{z^2} - \frac{1}{z^2} - \frac{3}{z^4}\right)e^{-z^2/2} \leq e^{-z^2/2},$$

und damit

$$\left(\frac{1}{y} - \frac{1}{y^3}\right)e^{-y^2/2} \leq \int_y^\infty e^{-z^2/2}\, dz.$$

Hieraus folgt die untere Schranke für $P[Z \geq y]$. ∎

Für eine $N(m, \sigma^2)$-verteilte Zufallsvariable X mit $\sigma > 0$ ist $Z = \frac{X-m}{\sigma}$ standardnormalverteilt. Also erhalten wir für $y \geq m$:

$$P[X \geq y] = P\left[\frac{X-m}{\sigma} \geq \frac{y-m}{\sigma}\right] \leq \frac{\sigma}{y-m} \cdot (2\pi)^{-1/2} \cdot e^{-\frac{(y-m)^2}{2\sigma^2}}, \tag{4.12}$$

sowie eine entsprechende Abschätzung nach unten.

4.4 Erwartungswert

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Wir wollen nun den Erwartungswert einer allgemeinen nicht negativen Zufallsvariable $X : \Omega \to \mathbb{R}$ definieren. Ist X diskret, also der Wertebereich $X(\Omega)$ abzählbar, dann gilt

$$E[X] \;=\; \sum_{a \in X(\Omega)} a\, P[X = a] \;=\; \sum_{a \in X(\Omega)} a\, p_X(a)\,. \tag{4.13}$$

Für Zufallsvariablen mit absolutstetiger Verteilung macht diese Definition offensichtlich keinen Sinn, da überabzählbar viele Werte angenommen werden, und die Wahrscheinlichkeit $P[X = a]$ für jeden einzelnen Wert a gleich Null ist. Stattdessen approximieren wir X durch eine *monoton wachsende* Folge von diskreten Zufallsvariablen X_n $(n \in \mathbb{N})$ mit

$$X(\omega) \;=\; \lim_{n\to\infty} X_n(\omega) \qquad \text{für alle } \omega \in \Omega\,,$$

und definieren den Erwartungswert $E[X]$ dann als Grenzwert der Erwartungswerte $E[X_n]$. Da die Folge der Zufallsvariablen X_n monoton wachsend ist, ist auch die Folge $E[X_n]$ der Erwartungswerte monoton wachsend, und somit existiert der Grenzwert in $[0, \infty]$.

Konkret unterteilen wir den Wertebereich $[0, \infty)$ beispielsweise in Intervalle der Länge 2^{-n}, und runden Werte von X, die in einem dieser Intervalle liegen, auf die untere Intervallgrenze ab, d.h. wir setzen

$$X_n(\omega) \;:=\; k \cdot 2^{-n} \qquad \text{falls } \; k \cdot 2^{-n} \le X(\omega) < (k+1) \cdot 2^{-n} \;\text{ für } k \in \mathbb{N}_0.$$

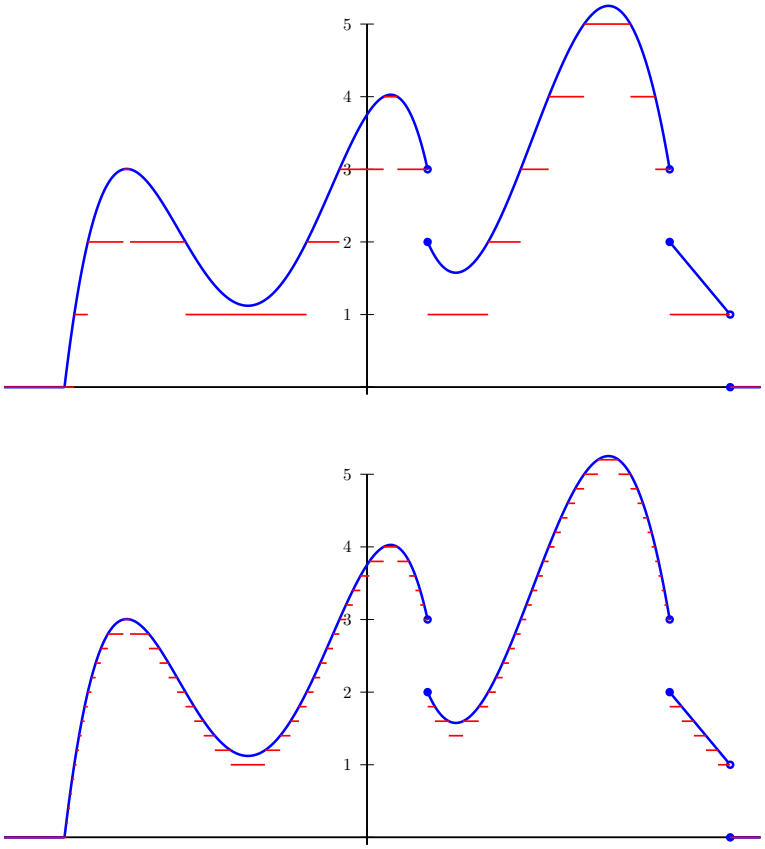


Abbildung 4.8: Approximation einer allgemeinen Zufallsvariable durch diskrete Zufallsvariablen.

Offensichtlich ist X_n eine Zufallsvariable mit abzählbarem Wertebereich, und es gilt

$$X_n \quad = \quad \sum_{k=0}^{\infty} \frac{k}{2^n} I_{\{\frac{k}{2^n} \le X < \frac{k+1}{2^n}\}}. \tag{4.14}$$

Die Folge $X_n(\omega)$ ist für jedes ω monoton wachsend, da die Unterteilung immer feiner wird, und

$$\lim_{n\to\infty} X_n(\omega) \quad = \quad \sup_{n\in\mathbb{N}} X_n(\omega) \quad = \quad X(\omega) \qquad \text{für alle } \omega \in \Omega.$$

Definition 4.17 (Erwartungswert einer reellwertigen Zufallsvariable).

(i) Der **Erwartungswert** einer Zufallsvariable $X : \Omega \to [0,\infty]$ bzgl. P ist definiert als

$$E[X] \;:=\; \lim_{n\to\infty} E[X_n] \;=\; \sup_{n\in\mathbb{N}} E[X_n] \;\in\; [0,\infty], \tag{4.15}$$

wobei $(X_n)_{n\in\mathbb{N}}$ eine beliebige monoton wachsende Folge von nichtnegativen diskreten Zufallsvariablen mit $X = \lim X_n$ ist.

(ii) Für eine allgemeine Zufallsvariable $X : \Omega \to \mathbb{R}$ sei $X = X^+ - X^-$ mit

$$X^+ := \max(X,0) \quad \text{und} \quad X^- := -\min(X,0)$$

die Zerlegung in den Positivteil X^+ und den Negativteil X^-. Ist mindestens einer der beiden Erwartungswerte $E[X^+]$ bzw. $E[X^-]$ endlich, dann ist der **Erwartungswert** von X bzgl. P definiert als

$$E[X] \quad := \quad E[X^+] - E[X^-] \quad \in \quad [-\infty,\infty].$$

Es ist zunächst nicht klar, ob die Definition des Erwartungswerts einer nichtnegativen Zufallsvariable in (4.15) von der Wahl der approximierenden Folge (X_n) abhängt. Tatsächlich kann man zeigen, dass dies nicht der Fall ist:

Lemma 4.18 (Wohldefiniertheit). *Die Definition in 4.17 (i) ist unabhängig von der Wahl einer monoton wachsenden Folge (X_n) von nichtnegativen diskreten Zufallsvariablen mit $X = \lim_{n\in\mathbb{N}} X_n$.*

Aus dem Lemma folgt auch, dass Definition 4.17 konsistent mit der oben gegebenen Definition des Erwartungswerts einer diskreten Zufallsvariable ist: Ist X selbst diskret, dann können wir nämlich insbesondere die konstante approximierende Folge $X_n = X$ in (4.15) verwenden.

Für den Beweis von Lemma 4.18 verweisen wir auf die Vorlesungen Analysis III und Einführung in die Wahrscheinlichkeitstheorie, sowie auf die Literatur, siehe z.B. Klenke „Wahrscheinlichkeitstheorie“ [1] oder Williams „Probability with martingales“ [4, Appendix A5]. Tatsächlich ist der Erwartungswert einer reellwertigen Zufallsvariable X nichts anderes als das Lebesgue-Integral $\int X\, dP$ der Funktion X bezüglich des Wahrscheinlichkeitsmaßes P. Die Lebesguesche Integrationstheorie wird im Detail in der Vorlesung Analysis III behandelt. Auch der Beweis der folgenden Aussagen findet sich in den obigen Referenzen und in vielen anderen Lehrbüchern.

Satz 4.19 (Eigenschaften des Erwartungswerts). Für reellwertige Zufallsvariablen X, Y auf $(\Omega, \mathcal{A}, P)$ mit $E[|X|] < \infty$ und $E[|Y|] < \infty$, sowie für $a, b \in \mathbb{R}$ gilt:

(i) *Linearität:* $E[aX + bY] \;=\; a \cdot E[X] + b \cdot E[Y]$.

(ii) *Monotonie:* Gilt $P[X \le Y] = 1$, dann folgt $E[X] \le E[Y]$.

(iii) *Monotone Konvergenz:* Ist $(X_n)_{n\in\mathbb{N}}$ eine *monoton wachsende* Folge von Zufallsvariablen mit $E[X_1^-] < \infty$ und gilt $P[X = \lim X_n] = 1$, dann folgt $E[X] = \lim_{n\to\infty} E[X_n]$.

Die Aussagen sind von grundlegender Bedeutung für die Lebesguesche Integrationstheorie, siehe Analysis III. Insbesondere der Beweis der letzten Teilaussage ist nicht trivial (Satz von Beppo Levi).

Erwartungswerte von Zufallsvariablen mit absolutstetiger Verteilung

Wir wollen nun in Analogie zu (4.13) eine Formel zur Berechnung des Erwartungswerts einer Zufallsvariable $X : \Omega \to \mathbb{R}$ mit absolutstetiger Verteilung herleiten. Sei $f : \mathbb{R} \to [0,\infty)$ die Dichtefunktion, d.h.

$$P[X \le c] \;=\; F(c) \;=\; \int_{-\infty}^{c} f(x)\,dx.$$

Insbesondere ist die Verteilungsfunktion stetig, und daher gilt für alle $c \in \mathbb{R}$ nach Satz 4.11

$$P[X = c] \;=\; F(c) - \lim_{y\uparrow c} F(y) \;=\; 0\,, \quad \text{und somit} \quad P[X < c] = P[X \le c] = F(c).$$

Satz 4.20. Sei X eine reellwertige Zufallsvariable mit Verteilungsdichte f.

(i) Ist $X \ge 0$, dann gilt

$$E[X] \;=\; \int_0^\infty x\, f(x)\, dx. \tag{4.16}$$

(ii) Allgemein gilt

$$E[X] \;=\; \int_{-\infty}^\infty x\, f(x)\, dx, \tag{4.17}$$

vorausgesetzt $\int_0^\infty x\, f(x)\,dx < \infty$ oder $\int_{-\infty}^0 |x|\, f(x)\,dx < \infty$.

Beweis. (i) Sei zunächst X nichtnegativ. Wir approximieren X durch die in (4.14) definierte monoton wachsende Folge diskreter Zufallsvariablen. Für $n \in \mathbb{N}$ erhalten wir dann

$$\begin{aligned} E[X_n] \;&=\; \sum_{k=0}^\infty k2^{-n}\, P\left[k2^{-n} \le X < (k+1)2^{-n}\right] \;=\; \sum_{k=0}^\infty k2^{-n}\, \left(F((k+1)2^{-n}) - F(k2^{-n})\right) \\ &=\; \sum_{k=0}^\infty \int_{k2^{-n}}^{(k+1)2^{-n}} k2^{-n}\, f(x)\,dx \;=\; \int_0^\infty \lfloor x \rfloor_n\; f(x)\,dx, \end{aligned}$$

wobei wir $\lfloor x \rfloor_n := k2^{-n}$ für $k \in \mathbb{N}_0$ mit $x \in [k2^{-n}, (k+1)2^{-n})$ setzen. Nach Definition 4.17 folgt dann

$$E[X] \;=\; \lim_{n\to\infty} E[X_n] \;=\; \lim_{n\to\infty} \int_0^\infty \lfloor x \rfloor_n\; f(x)\,dx \;=\; \int_0^\infty x\, f(x)\,dx\,.$$

Hierbei folgt die Konvergenz der Integrale im letzten Schritt wegen

$$\left| \int_0^\infty x\, f(x)\,dx \;-\; \int_0^\infty \lfloor x \rfloor_n\; f(x)\,dx \right| \;\le\; \int_0^\infty |x - \lfloor x \rfloor_n|\; f(x)\,dx \;\le\; 2^{-n} \int_0^\infty f(x)\,dx \;=\; 2^{-n} \;\longrightarrow\; 0$$

für $n \to \infty$.

(ii) Im allgemeinen Fall überzeugt man sich leicht, dass X^+ und X^- nichtnegative Zufallsvariablen mit Dichtefunktionen $f_{X^+}(x) = f(x) \cdot I_{(0,\infty)}(x)$ und $f_{X^-}(x) = f(-x) \cdot I_{(-\infty,0)}(-x)$ sind. Nach (i) folgt

$$E\left[X^+\right] \;=\; \int_0^\infty x\, f(x)\, dx \quad \text{und} \quad E\left[X^-\right] \;=\; \int_{-\infty}^0 |x|\, f(x)\, dx,$$

und damit erhalten wir

$$E[X] \;=\; E[X^+] - E[X^-] \;=\; \int_{-\infty}^\infty x\, f(x)\, dx\,,$$

vorausgesetzt mindestens einer der beiden Erwartungswerte $E[X^+]$ oder $E[X^-]$ ist endlich. ■

Bemerkung (Erwartungswert hängt nur von der Verteilung ab). Satz 4.20 zeigt insbesondere, dass genau wie im diskreten Fall auch der Erwartungswert einer absolutstetigen Zufallsvariable X nur von der Verteilung μ_X von X abhängt. Dies gilt auch allgemein. Tatsächlich gilt für eine beliebige reellwertige Zufallsvariable

$$E[X] \;=\; \int_{\mathbb{R}} x\; \mu_X(dx)\,, \tag{4.18}$$

sofern der Erwartungswert definiert, d.h. $E[X^+]$ oder $E[X^-]$ endlich ist. Hierbei ist das Integral auf der rechten Seite ein Lebesgue-Integral bezüglich des Wahrscheinlichkeitsmaßes μ_X, siehe die Vorlesung Einführung in die Wahrscheinlichkeitstheorie. Die Identität (4.18) verallgemeinert sowohl die Berechnungsformel (4.13) für den Erwartungswert diskreter Zufallsvariablen, als auch die Formel (4.17) für den Erwartungswert absolutstetiger Zufallsvariablen.

Beispiele

Beispiel (Kontinuierliche Gleichverteilung). Für eine auf einem endlichen nichtleeren Intervall (a, b) gleichverteilte Zufallsvariable U erhalten wir

$$E[U] \;=\; \int x\, \frac{1}{b-a}\, I_{(a,b)}(x)\; dx \;=\; \frac{1}{b-a} \int_a^b x\; dx \;=\; \frac{1}{2}\, \frac{b^2 - a^2}{b-a} \;=\; \frac{a+b}{2}\,.$$

Beispiel (Exponentialverteilung). Der Erwartungswert einer zum Parameter $\lambda > 0$ exponentialverteilten Zufallsvariable T ist

$$E[T] \;=\; \int x\, \lambda\, e^{-\lambda x}\, I_{(0,\infty)}(x)\; dx \;=\; \int_0^\infty \lambda\, x\, e^{-\lambda x}\; dx \;=\; \int_0^\infty e^{-\lambda x}\; dx \;=\; \frac{1}{\lambda}.$$

Hierbei haben wir im vorletzten Schritt partielle Integration benutzt.

Beispiel (Normalverteilung). Für eine standardnormalverteilte Zufallsvariable Z gilt

$$E[Z] \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty x\, e^{-x^2/2}\; dx \;=\; 0\,.$$

Eine mit Parametern $m \in \mathbb{R}$ und $\sigma^2 \in (0, \infty)$ normalverteilte Zufallsvariable X können wir darstellen als $X = m + \sigma Z$ mit $Z \sim N(0, 1)$. Aufgrund der Linearität des Erwartungswerts erhalten wir

$$E[X] \;=\; m + \sigma\, E[Z] \;=\; m$$

als Erwartungswert der Verteilung $N(m, \sigma^2)$.

Beispiel (Eine Zufallsvariable, die weder diskret noch absolutstetig ist). Sei $Y = B \cdot X$ mit unabhängigen Zufallsvariablen $B \sim \text{Bernoulli}(p)$ und $Z \sim N(m, \sigma^2)$. Nach dem Satz von der totalen Wahrscheinlichkeit gilt für jede Menge $A \in \mathcal{B}(\mathbb{R})$:

$$\begin{aligned} \mu_X[A] \;&=\; P[X \in A] \;=\; P[X \in A \,|\, B = 0] \cdot P[B = 0] \;+\; P[X \in A \,|\, B = 1] \cdot P[B = 1] \\ &=\; (1-p) \cdot \delta_0[A] \;+\; p \cdot N(0,1)[A]\,. \end{aligned}$$

Die Verteilung von X ist also das Wahrscheinlichkeitsmaß $\mu_X = (1-p)\, \delta_0 + p\, N(m, \sigma^2)$, und

$$E[X] \;=\; (1-p) \cdot 0 + p \cdot m \;=\; p \cdot m\,.$$

4.5 Transformationen von reellwertigen Zufallsvariablen

In diesem Abschnitt überlegen wir uns zunächst, wie man Verteilungsfunktionen, Dichten und Erwartungswerte von transformierten Zufallsvariablen, d.h. von Funktionen $g(X)$ einer Zufallsvariable X berechnen kann. Mithilfe einer geeigneten Transformation können wir auch reelle Zufallsvariablen mit einer vorgegebenen Verteilungsfunktion F aus gleichverteilten Zufallsvariablen erzeugen. Ist beispielsweise die Verteilungsfunktion $F : \mathbb{R} \to [0,1]$ streng monoton wachsend und stetig, also eine Bijektion von $\mathbb{R}$ nach $(0,1)$, und ist $U : \Omega \to (0,1)$ auf $(0,1)$ gleichverteilt, dann hat die Zufallsvariable $F^{-1}(U)$ die Verteilung μ, denn es gilt

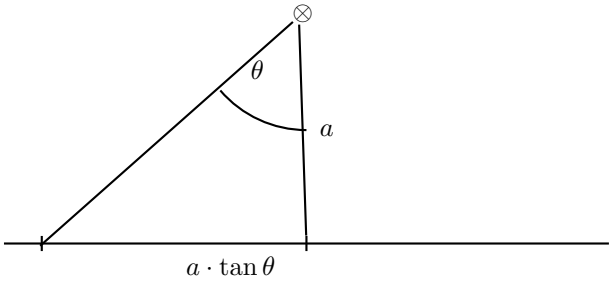
$$P[F^{-1}(U) \le c] \;=\; P[U \le F(c)] \;=\; F(c) \qquad \text{für alle } c \in \mathbb{R}.$$

Das beschriebene Inversionsverfahren werden wir in Satz 4.25 erweitern, um reelle Zufallsvariablen mit beliebigen Verteilungen zu konstruieren. Da die Verteilungsfunktion dann im Allgemeinen keine Bijektion ist, verwenden wir statt der Inversen F^{-1} eine verallgemeinerte (linksstetige) Inverse, die durch die Quantile der zugrundeliegenden Verteilung bestimmt ist.

Transformation von Verteilungsfunktionen und Dichten

Wir beginnen mit einem Beispiel.

Beispiel. Eine Lichtquelle strahlt gleichmäßig in alle Richtungen. Im Abstand a von der Lichtquelle befindet sich eine Gerade. Wir wollen nun die Intensitätsverteilung der Lichtstrahlung auf der Geraden berechnen. Ist $\theta \in (-\pi/2, \pi/2)$ der Winkel eines Lichtstrahls zur Normalen, dann können wir den Auftreffpunkt des Lichtstrahls auf der Geraden durch $a \cdot \tan\theta$ parametrisieren. Da die Lichtquelle gleichmäßig in alle Richtungen strahlt, nehmen wir an, dass der Winkel θ gleichverteilt ist. Die gesuchte Intensitätsverteilung ist dann (bis auf eine Normierungskonstante) die Verteilung von $a \cdot \tan\theta$ für $\theta \sim \mathrm{Unif}(-\pi/2, \pi/2)$.



Sei nun allgemein $I = (a,b) \subseteq \mathbb{R}$ ein offenes Intervall, und sei $X : \Omega \to I$ eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Ferner sei $g : I \to g(I) \subseteq \mathbb{R}$ eine streng monoton wachsende Funktion. Insbesondere ist g bijektiv, die Umkehrfunktion bezeichnen wir mit g^{-1}.

Satz 4.21 (Transformation von Verteilungsfunktionen und Dichten).

(i) $g(X)$ ist eine Zufallsvariable mit Verteilungsfunktion

$$F_{g(X)}(c) \;=\; F_X(g^{-1}(c)) \quad \text{für } c \in g(I). \tag{4.19}$$

(ii) Ist die Verteilung von X absolutstetig mit Dichte f_X, und ist g stetig differenzierbar mit $g'(x) > 0$ für alle $x \in I$, dann ist auch die Verteilung von $g(X)$ absolutstetig mit Dichte

$$f_{g(X)}(y) \quad = \quad \begin{cases} f_X(g^{-1}(y)) \cdot |(g^{-1})'(y)| & \text{für } y \in g(I), \\ 0 & \text{sonst.} \end{cases} \tag{4.20}$$

Der zweite Teil der Aussage gilt auch im Fall $g' < 0$. Man beachte, dass die Ableitung der Umkehrfunktion in beiden Fällen durch $(g^{-1})'(y) = 1/g'(g^{-1}(y))$ gegeben ist.

Beweis. (i) Da g streng monoton wachsend ist, gilt

$$\{g(X) \le c\} \;=\; \{X \le g^{-1}(c)\} \;\in\; \mathcal{A}$$

für alle $c \in g(I)$. Also ist $g(X)$ eine Zufallsvariable mit Verteilungsfunktion

$$F_{g(X)}(c) \;=\; P[g(X) \le c] \;=\; P[X \le g^{-1}(c)] \;=\; F_X(g^{-1}(c)).$$

(ii) Ist g stetig differenzierbar mit $g' > 0$, dann ist auch die Umkehrfunktion g^{-1} stetig differenzierbar. Nach (i) erhalten wir dann mit der Substitution $y = g(x)$:

$$F_{g(X)}(c) \;=\; F_X(g^{-1}(c)) \;=\; \int_a^{g^{-1}(c)} f_X(x)\,dx \;=\; \int_{g(a)}^{c} f_X(g^{-1}(y))\,(g^{-1})'(y)\,dy$$

für alle $c \in g(I)$. Die Behauptung folgt wegen $P[g(X) \notin g(I)] = 0$. ■

Beispiel (Geometrische Wahrscheinlichkeiten I). Im Beipiel von oben nehmen wir an, dass θ auf $(-\pi/2, \pi/2)$ gleichverteilt ist. Dann erhalten wir für $a > 0$:

$$f_{a\tan\theta}(x) \quad = \quad \frac{1}{\pi a} \cdot \frac{1}{1 + (x/a)^2}, \qquad x \in \mathbb{R}.$$

Die Verteilung mit dieser Dichte heißt *Cauchyverteilung* zum Parameter a.

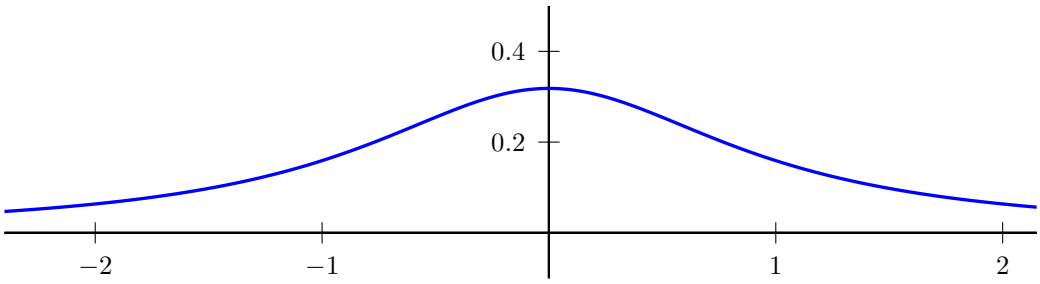


Abbildung 4.9: Graph der Dichtefunktion $f_{\tan\theta}$

Beispiel (Geometrische Wahrscheinlichkeiten II). Sei $\theta : \Omega \to [0, 2\pi)$ ein zufälliger, auf $[0, 2\pi)$ gleichverteilter, Winkel. Wir wollen die Verteilung von $\cos\theta$ berechnen. Da die Kosinusfunktion auf $[0, 2\pi)$ nicht streng monoton ist, ist (4.20) nicht direkt anwendbar. Wir können aber das Intervall $[0, 2\pi)$ in die Teile $[0, \pi)$ und $[\pi, 2\pi)$ zerlegen, und dann die Verteilung ähnlich wie im Beweis von Satz 4.21 berechnen. Wegen

$$\begin{aligned} P[\cos\theta > c] \;&=\; P[\cos\theta > c \text{ und } \theta \in [0,\pi)] \;+\; P[\cos\theta > c \text{ und } \theta \in [\pi, 2\pi)] \\ &=\; P[\theta \in [0, \arccos c)] \;+\; P[\theta \in [2\pi - \arccos c, 2\pi)] \\ &=\; \frac{2}{2\pi} \cdot \arccos c \end{aligned}$$

erhalten wir, dass $\cos\theta$ eine absolutstetige Verteilung mit Dichte

$$f_{\cos\theta}(x) \quad = \quad F'_{\cos\theta}(x) \quad = \quad \frac{1}{\pi} \cdot \frac{1}{\sqrt{1 - x^2}}; \qquad x \in (-1, 1)$$

hat. Anstelle von (4.20) gilt in diesem Fall

$$f_{\cos\theta}(x) \quad = \quad f_X(\psi_1(x)) \cdot |\psi_1'(x)| + f_X(\psi_2(x)) \cdot |\psi_2'(x)|,$$

wobei $\psi_1(x) = \arccos x$ und $\psi_2(x) = 2\pi - \arccos x$ die Umkehrfunktionen auf den Teilintervallen sind. Entsprechende Formeln erhält man auch allgemein, wenn die Transformation nur stückweise bijektiv ist.

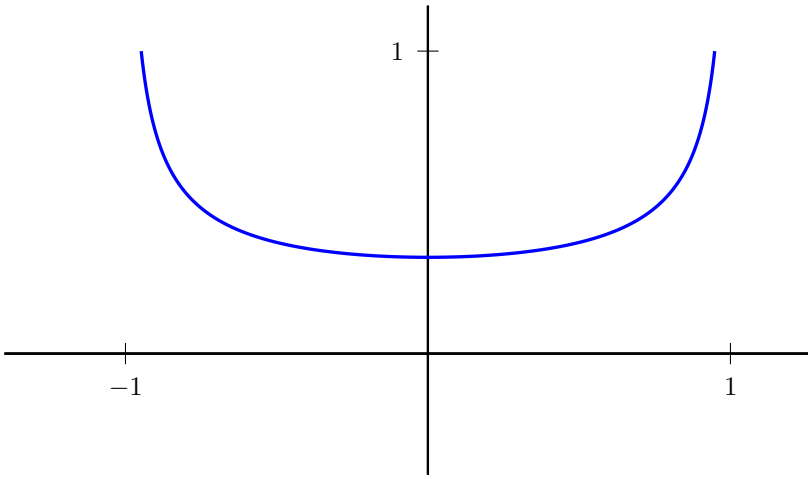


Abbildung 4.10: Graph der Dichtefunktion $f_{\cos\theta}$

Transformationsformel für Erwartungswerte

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Für diskrete Zufallsvariablen $X : \Omega \to \mathbb{R}$ gilt

$$E[g(X)] \;=\; \sum_{a \in X(\Omega)} g(a)\, p_X(a)$$

für beliebige Funktionen $g : \mathbb{R} \to \mathbb{R}$ mit $E[|g(X)|] < \infty$. Eine entsprechende Aussage gilt auch im absolutstetigen Fall.

Satz 4.22. Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable mit Verteilungsdichte f, und sei $g : \mathbb{R} \to \mathbb{R}$ eine Funktion, für die $g \cdot f$ Lebesgue-integrierbar ist. Dann gilt

$$E[g(X)] \;=\; \int_{\mathbb{R}} g(x)\, f(x)\, dx\,. \tag{4.21}$$

Insbesondere gilt

$$P[X \in B] \;=\; E[I_B(X)] \;=\; \int_{\mathbb{R}} I_B(x)\, f(x)\, dx \;=\; \int_B f(x)\, dx \tag{4.22}$$

für alle Mengen $B \in \mathcal{B}(R)$.

Ist g stetig differenzierbar mit $g' > 0$, dann können wir den Satz mithilfe der Dichtetransformationsformel beweisen. Nach Satz 4.21 erhalten wir in diesem Fall

$$E[g(X)] \;=\; \int_{\mathbb{R}} y\, f_{g(X)}(y)\, dy \;=\; \int_{g(\mathbb{R})} y\, f(g^{-1}(y))\, (g^{-1})'(y)\, dy \;=\; \int_{\mathbb{R}} g(x)\, f(x)\, dx\,.$$

Hierbei haben wir im letzten Schritt die Substitution $y = g(x)$ bzw. $x = g^{-1}(y)$ verwendet. Der Beweis im allgemeinen Fall erfordert Grundlagen aus der Maßtheorie. Wir skizzieren die wesentlichen Schritte.

Beweis (Skizze). (i) Wir bemerken zunächst, dass die Aussage für $g = I_{(a,b]}$ mit $-\infty < a \le b < \infty$ erfüllt ist, denn

$$E[I_{(a,b]}(X)] \;=\; P[a < X \le b] \;=\; F(b) - F(a) \;=\; \int_a^b f(x)\,dx \;=\; \int_{\mathbb{R}} I_{(a,b]}(x)\, f(x)\,dx\,.$$

(ii) Als nächstes zeigen wir die Aussage für $g = I_B$, wobei B eine beliebige Borel-Menge ist. Zu zeigen ist (4.22), also

$$\mu_X[B] \;=\; \int I_B(x)\, f(x)\,dx \;:=\; J[B] \quad \text{für alle } B \in \mathcal{B}(R). \tag{4.23}$$

Man kann zeigen, dass sowohl μ_X als auch J Wahrscheinlichkeitsverteilungen auf $\mathcal{B}(R)$ sind. Nach Schritt (i) wissen wir bereits, dass diese Wahrscheinlichkeitsverteilungen für Intervalle $B = (a,b]$ übereinstimmen. Da diese Intervalle einen durchschnittsstabilen Erzeuger der Borelschen σ-Algebra bilden, folgt (4.23) nach dem Eindeutigkeitssatz 4.2.

(iii) Im nächsten Schritt zeigen wir, dass die Aussage für Funktionen g gilt, die nur endlich viele Werte annehmen. Diese Funktionen sind Linearkombinationen von Indikatorfunktionen, d.h. $g = \sum_{i=1}^n c_i\, I_{B_i}$ mit $n \in \mathbb{N}$, $c_i \in \mathbb{R}$ und $B_i \in \mathcal{B}(\mathbb{R})$. Daher folgt die Aussage aus (ii) und der Linearität des Erwartungswerts.

(iv) Ist $g : \mathbb{R} \to [0,\infty)$ eine beliebige *messbare* Funktion, d.h. $\{x \in \mathbb{R} : g(x) \le c\} \in \mathcal{B}(\mathbb{R})$ für alle $c \in \mathbb{R}$, dann können wir g ähnlich wie in (4.14) als Grenzwert einer monoton wachsenden Folge von messbaren Funktionen g_n schreiben, die nur endlich viele Werte annehmen. Nach (iii) und dem Satz von der monotonen Konvergenz (siehe Satz 4.19 (iii)) erhalten wir dann

$$E[g(X)] \;=\; E[\lim_{n\to\infty} g_n(X)] \;=\; \lim_{n\to\infty} E[g(X_n)] \;=\; \lim_{n\to\infty} \int_{\mathbb{R}} g_n(x)\, f(x)\,dx \;=\; \int_{\mathbb{R}} g(x)\, f(x)\,dx.$$

(v) Im allgemeinen Fall folgt die Aussage schließlich durch die Zerlegung $g = g^+ - g^-$. ■

Das obige Beweisverfahren wird sehr häufig verwendet, und manchmal auch als *"maßtheoretische Induktion"* bezeichnet. Für Zufallsvariablen mit einer beliebigen Verteilung gilt in Analogie zu (4.21)

$$E[g(X)] \;=\; \int g(x)\; \mu_X(dx)\,,$$

wobei das Integral auf der rechten Seite ein Lebesgue-Integral bezüglich des Wahrscheinlichkeitsmaßes μ_X ist, siehe die Vorlesung Einführung in die Wahrscheinlichkeitstheorie.

Quantile und Inversion der Verteilungsfunktion

Quantile sind Stellen, an denen die Verteilungsfunktion einen bestimmten Wert überschreitet. Solche Stellen müssen häufig in praktischen Anwendungen (z.B. Qualitätskontrolle) berechnet werden. Mithilfe von Quantilen kann man verallgemeinerte Umkehrfunktionen der im Allgemeinen nicht bijektiven Verteilungsfunktion definieren. Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit Verteilungsfunktion F.

Definition 4.23 (Quantile). Sei $u \in [0,1]$. Dann heißt $q \in \mathbb{R}$ ein *u-Quantil* der Verteilung von X, falls

$$P[X < q] \le u \qquad \text{und} \qquad P[X > q] \le 1 - u$$

gilt. Ein *Median* ist ein $\frac{1}{2}$-Quantil. Quantile zu $u = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$ bezeichnet man als *Quartile*.

Ist die Verteilungsfunktion streng monoton wachsend, dann ist $q = F^{-1}(u)$ für $u \in (0,1)$ das einzige u-Quantil. Im Allgemeinen kann es hingegen mehrere u-Quantile zu einem Wert u geben.

Beispiel (Empirische Quantile). Wir betrachten eine Stichprobe, die aus n reellwertigen Daten / Messwerten $x_1, \ldots, x_n$ mit $x_1 \le x_2 \le \ldots \le x_n$ besteht. Die empirische Verteilung der Stichprobe ist die Wahrscheinlichkeitsverteilung

$$\mu = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

auf $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$, d.h. für $B \subseteq \mathbb{R}$ ist

$$\mu[B] = \frac{1}{n} \left| \{ x_i \in B, 1 \le i \le n \} \right|$$

die relative Häufigkeit des Bereichs B unter den Messwerten x_i. Die empirische Verteilung ergibt sich, wenn wir zufällig ein $i \in \{1, \ldots, n\}$ wählen, und den entsprechenden Messwert betrachten. Die Quantile der empirischen Verteilung bezeichnet man als *Stichprobenquantile*. Ist $n \cdot u$ nicht ganzzahlig, dann ist $x_{\lceil n \cdot u \rceil}$ das eindeutige u-Quantil der Stichprobe. Ist hingegen $u = k/n$ mit $k \in \{1, \ldots, n-1\}$, dann ist jedes $q \in [x_k, x_{k+1}]$ ein u-Quantil der empirischen Verteilung.

Wir definieren nun zwei verallgemeinerte Inverse einer Verteilungsfunktion F, die ja im Allgemeinen nicht bijektiv ist. Für $u \in (0,1)$ sei

$$\begin{aligned} \underline{G}(u) &:= \inf\{x \in \mathbb{R} : F(x) \ge u\} = \sup\{x \in \mathbb{R} : F(x) < u\}, \qquad \text{und} \\ \overline{G}(u) &:= \inf\{x \in \mathbb{R} : F(x) > u\} = \sup\{x \in \mathbb{R} : F(x) \le u\}. \end{aligned}$$

Offensichtlich gilt $\underline{G}(u) \le \overline{G}(u)$. Die Funktionen $\underline{G}$ bzw. $\overline{G}$ sind links- bzw. rechtsstetig. Ist F stetig und streng monoton wachsend, also eine Bijektion von $\mathbb{R}$ nach $(0,1)$, dann ist $\underline{G}(u) = \overline{G}(u) = F^{-1}(u)$. Die Funktion $\underline{G}$ heißt daher auch die *linksstetige verallgemeinerte Inverse* von F. Das folgende Lemma zeigt, dass $\underline{G}(u)$ das kleinste und $\overline{G}(u)$ das größte u-Quantil ist:

Lemma 4.24. *Für $u \in (0,1)$ und $q \in \mathbb{R}$ sind die folgenden Aussagen äquivalent:*

(i) *q ist ein u-Quantil.*

(ii) $F(q-) \le u \le F(q)$.

(iii) $\underline{G}(u) \le q \le \overline{G}(u)$.

Hierbei ist $F(q-) := \lim_{y \nearrow q} F(y)$ der linksseitige Limes von F an der Stelle q.

Beweis. Nach Definition ist q genau dann ein u-Quantil, wenn

$$P[X < q] \le u \le 1 - P[X > q] = P[X \le q]$$

gilt. Hieraus folgt die Äquivalenz von (i) und (ii).

Um zu beweisen, dass (3) äquivalent zu diesen Bedingungen ist, müssen wir zeigen, dass $\underline{G}(u)$ das kleinste und $\overline{G}(u)$ das größte u-Quantil ist. Wir bemerken zunächst, dass $\underline{G}(u)$ ein u-Quantil ist, da

$$F(\underline{G}(u)-) = \lim_{x \nearrow \underline{G}(u)} \underbrace{F(x)}_{\substack{<u \\ \text{für } x<\underline{G}(u)}} \le u, \qquad \text{und} \qquad F(\underline{G}(u)) = \lim_{x \searrow \underline{G}(u)} \underbrace{F(x)}_{\substack{\ge u \\ \text{für } x>\underline{G}(u)}} \ge u.$$

Andererseits ist $x < \underline{G}(u)$ kein u-Quantil, denn es gilt $F(x) < u$. Somit ist $\underline{G}(u)$ das kleinste u-Quantil. Auf ähnliche Weise folgt, dass $\overline{G}(u)$ das größte u-Quantil ist. ■

Konstruktion und Simulation reellwertiger Zufallsvariablen

Wie erzeugt man ausgehend von auf $(0,1)$ gleichverteilten Zufallszahlen Stichproben von anderen Verteilungen μ auf $\mathbb{R}^1$?

Beispiel (Endlicher Fall). Gilt $\mu[S] = 1$ für eine endliche Teilmenge $S \subseteq \mathbb{R}$, dann können wir die Frage leicht beantworten: Sei $S = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}$ mit $n \in \mathbb{N}$ und $x_1 < x_2 < \ldots < x_n$. Die Verteilungsfunktion einer Wahrscheinlichkeitsverteilung μ auf S ist gegeben durch

$$F(c) \;=\; \mu[(-\infty, c]] \;=\; \sum_{i:x_i \le c} \mu[\{x_i\}].$$

Ist U eine auf $(0,1)$ gleichverteilte Zufallsvariable, dann wird durch

$$X(\omega) := x_k \qquad \text{falls } F(x_{k-1}) < U(\omega) \le F(x_k), \quad x_0 := -\infty,$$

eine Zufallsvariable mit Verteilung μ definiert, denn für $k = 1, \ldots, n$ gilt

$$P[X = x_k] = F(x_k) - F(x_{k-1}) = \mu[\{x_k\}].$$

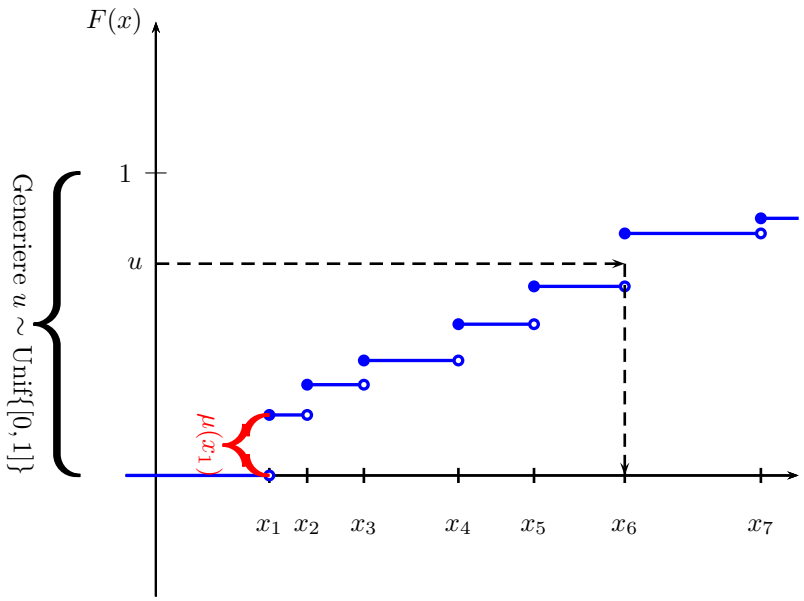


Abbildung 4.11: Erzeugen einer Stichprobe von einer Wahrscheinlichkeitsverteilung auf einer endlichen Menge $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}$.

Wir wollen das Vorgehen aus dem Beispiel nun verallgemeinern. Sei $F : \mathbb{R} \to [0,1]$ eine Funktion mit den folgenden Eigenschaften: F ist

(i) monoton wachsend: $F(x) \le F(y) \quad \forall\, x \le y,$

(ii) rechtsstetig: $\lim_{x \downarrow c} F(x) = F(c) \quad \forall\, c \in \mathbb{R},$

(iii) normiert: $\lim_{x \searrow -\infty} F(x) = 0 \quad , \quad \lim_{x \nearrow +\infty} F(x) = 1.$

Das folgende Resultat liefert eine explizite Konstruktion einer Zufallsvariable mit Verteilungsfunktion F:

Satz 4.25 (Quantiltransformation). Ist $F : \mathbb{R} \to [0,1]$ eine Funktion mit (i)-(iii), und $\underline{G} : (0,1) \to \mathbb{R}$ die linksstetige verallgemeinerte Inverse, dann ist das Bild $\mu := \mathrm{Unif}_{(0,1)} \circ \underline{G}^{-1}$ der Gleichverteilung auf $(0,1)$ unter $\underline{G}$ ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ mit Verteilungsfunktion F.

Insbesondere gilt: Ist $U : \Omega \to (0,1)$ eine unter P gleichverteilte Zufallsvariable, dann hat die Zufallsvariable

$$X(\omega) \ := \ \underline{G}(U(\omega))$$

unter P die Verteilungsfunktion F.

Beweis. Da $\underline{G}(u)$ ein u-Quantil ist, gilt $F(\underline{G}(u)) \geq u$, also

$$\underline{G}(u) \ = \ \min\{x \in \mathbb{R} : F(x) \geq u\},$$

und somit für $c \in \mathbb{R}$:

$$\underline{G}(u) \leq c \iff F(x) \geq u \text{ für ein } x \leq c \iff F(c) \geq u.$$

Es folgt:

$$P[\underline{G}(U) \leq c] \ = \ \mathrm{Unif}_{(0,1)}[\{u \in (0,1) : \underbrace{\underline{G}(u) \leq c}_{\iff F(c) \geq u} \}] \ = \ F(c).$$

Also ist F die Verteilungsfunktion von $G(U)$ bzw. von μ. ■

Bemerkung. Nimmt X nur endlich viele Werte $x_1 < x_2 < \ldots < x_n$ an, dann ist F stückweise konstant, und es gilt:

$$\underline{G}(u) = x_k \text{ für } F(x_{k-1}) < u \leq F(x_k), \quad x_0 := -\infty,$$

d.h. $\underline{G}$ ist genau die oben im endlichen Fall verwendete Transformation.

Als Konsequenz aus Satz 4.25 ergibt sich unmittelbar:

Korollar 4.26 (Existenzsatz). Zu jeder Funktion F mit (i)-(iii) existiert eine reelle Zufallsvariable X bzw. eine Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}$ mit Verteilungsfunktion F.

Zudem erhalten wir einen expliziten Algorithmus zur Simulation einer Stichprobe von μ:

Algorithmus 2: Direktes Verfahren zur Simulation einer Stichprobe x von μ

Input : Linksstetige Inverse $\underline{G}$ der Verteilungsfunktion von μ.
Output Stichprobe x von μ.
:
1 Erzeuge (Pseudo)-Zufallszahl $u \in (0,1)$;
2 Setze $x := \underline{G}(u)$;
3 **return** x;

Dieser Algorithmus funktioniert theoretisch immer. Er ist aber oft nicht praktikabel, da man $\underline{G}$ nicht immer berechnen kann, oder da das Anwenden der Transformation $\underline{G}$ (zunächst unwesentliche) Schwachstellen des verwendeten Zufallsgenerators verstärkt. Man greift daher oft selbst im eindimensionalen Fall auf andere Simulationsverfahren wie z.B. „Acceptance Rejection“ Methoden zurück.

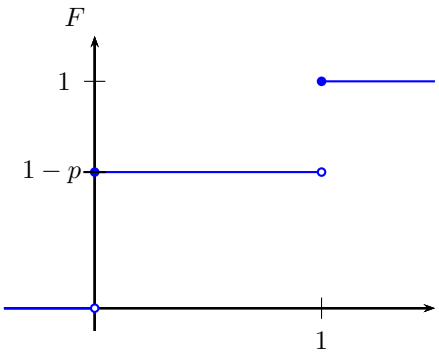


Abbildung 4.12: $\underline{G}(U) = I_{\{U>1-p\}}$ ist Bernoulli(p)-verteilt.

Beispiele. (i) Bernoulli(p)-Verteilung auf $\{0,1\}$. Hier gilt

$$F = (1-p)\cdot I_{[0,1)} + 1\cdot I_{[1,\infty)} \qquad \text{und} \qquad \underline{G} = I_{(1-p,1)}.$$

Also ist die Zufallsvariable $\underline{G}(U) = I_{\{U<1-p\}}$ für $U \sim \mathrm{Unif}_{(0,1)}$ Bernoulli(p)-verteilt.

(ii) Gleichverteilung auf (a,b). Hier ist $F(c) = \frac{c-a}{b-a}$ für $c \in [a,b]$, und damit

$$\underline{G}(u) = a + (b-a)u,$$

siehe Abbildung 4.13. Also ist $a + (b-a)U$ für $U \sim \mathrm{Unif}_{(0,1)}$ gleichverteilt auf (a,b).

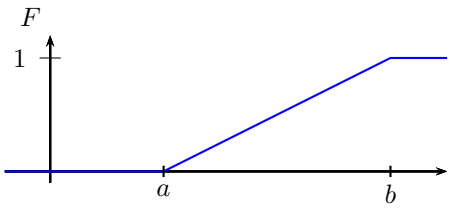


Abbildung 4.13: $\underline{G}(U) = a + (b-a)U$ ist auf (a,b) gleichverteilt.

(iii) Exponentialverteilung mit Parameter $\lambda > 0$:

$$F(c) = 1 - e^{-\lambda c}, \quad \underline{G}(u) = F^{-1}(u) = -\frac{1}{\lambda}\log(1-u).$$

Anwenden des negativen Logarithmus transformiert also die gleichverteilte Zufallsvariable $1-U$ in eine exponentialverteilte Zufallsvariable.

4.6 Mehrdimensionale Verteilungen

Um Wahrscheinlichkeiten und Erwartungswerte zu berechnen, die von mehreren reellwertigen Zufallsvariablen $X_1, \dots, X_d$ abhängen, benötigen wir wie schon im diskreten Fall deren gemeinsame Verteilung. Diese ist eine Wahrscheinlichkeitsverteilung auf $\mathbb{R}^d$.

Wahrscheinlichkeitsverteilungen auf $\mathbb{R}^d$

Ähnlich wie im eindimensionalen Fall können wir die *Borelsche σ-Algebra* $\mathcal{B}(\mathbb{R}^d)$ als die von den Quadern

$$(a_1,b_1]\times(a_2,b_2]\times\cdots\times(a_d,b_d], \qquad -\infty < a_i \le b_i < \infty \quad \text{für } i = 1,\dots,d,$$

erzeugte σ-Algebra definieren. Ein anderes durchschnittsstabiles Erzeugendensystem der Borelschen σ-Algebra bilden die Mengen

$$(-\infty, c_1] \times (-\infty, c_2] \times \cdots \times (-\infty, c_d], \qquad c_1, \ldots, c_d \in \mathbb{R}.$$

Außerdem wird $\mathcal{B}(\mathbb{R}^d)$ auch von der Kollektion aller offenen, sowie von der Kollektion aller abgeschlossenen Teilmengen des $\mathbb{R}^d$ erzeugt. Die nächste Aussage folgt aus dem Eindeutigkeitssatz 4.2.

Korollar 4.27 (Mehrdimensionale Verteilungsfunktion). Eine Wahrscheinlichkeitsverteilung μ auf $\mathcal{B}(\mathbb{R}^d)$ ist eindeutig festgelegt durch die Wahrscheinlichkeiten

$$F(c_1, \ldots, c_d) \;:=\; \mu\left[\,(-\infty, c_1] \times \cdots \times (-\infty, c_d]\,\right], \qquad c_1, \ldots c_d \in \mathbb{R}.$$

In Analogie zum eindimensionalen Fall definieren wir:

Definition 4.28 (Absolutstetige Verteilung; Dichtefunktion). Eine Wahrscheinlichkeitsverteilung μ auf $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ heißt *absolutstetig*, falls eine integrierbare Funktion $f : \mathbb{R}^d \to [0, \infty)$ existiert mit

$$\mu\left[(a_1, b_1] \times \cdots \times (a_d, b_d]\right] \;=\; \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} f(x_1, \ldots, x_d)\, dx_d \cdots dx_1 \tag{4.24}$$

für alle $-\infty < a_i \le b_i < \infty$ $(i = 1, \ldots, d)$. Eine integrierbare Funktion $f : \mathbb{R}^d \to [0, \infty)$ mit (4.25) heißt *Dichtefunktion* von μ.

Hierbei sind die Integrale im Allgemeinen als Lebesgue-Integrale zu interpretieren. Nach dem Satz von Fubini ist das Mehrfachintegral unabhängig von der Integrationsreihenfolge, siehe Analysis III. Mithilfe des Eindeutigkeitssatzes 4.2 kann man zeigen, dass im absolutstetigen Fall

$$\mu[B] \;=\; \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} I_B(x_1, \ldots, x_d)\, f(x_1, \ldots, x_d)\, dx_d \cdots dx_1 \tag{4.25}$$

für alle Mengen $B \in \mathcal{B}(\mathbb{R}^d)$ gilt.

Beispiel (Gleichverteilung). Sei $A \in B(\mathbb{R}^d)$ mit

$$\mathrm{vol}[A] \;=\; \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} I_A(x_1, \ldots, x_d)\, dx_d \cdots dx_1 \;\in\; (0, \infty).$$

Dann heißt die Wahrscheinlichkeitsverteilung auf $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ mit Dichtefunktion

$$f(x) \;=\; \frac{1}{\mathrm{vol}[A]}\, I_A(x)$$

Gleichverteilung auf der Menge A.

Produktmodelle

Sind $f_1, \ldots, f_d : \mathbb{R} \to [0, \infty)$ Dichten von Wahrscheinlichkeitsverteilungen $\mu_1, \ldots, \mu_d$ auf $\mathbb{R}$, dann ist

$$f(x_1, \ldots, x_d) \;=\; f_1(x_1) \cdot f_2(x_2) \cdot \cdots \cdot f_d(x_d) \tag{4.26}$$

die Dichte einer Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}^d$. Die Verteilung μ ist das Produkt $\mu_1 \otimes \cdots \otimes \mu_d$ der Wahrscheinlichkeitsverteilungen $\mu_1, \mu_2, \ldots, \mu_d$, d.h. es gilt

$$\mu\left[B_1 \times \cdots B_d\right] \;=\; \mu_1[B_1] \cdot \cdots \cdot \mu_d[B_d] \qquad \text{für alle } B_1, \ldots B_d \in \mathcal{B}(\mathbb{R}).$$

Beispiel (Gleichverteilung auf d-dimensionalem Quader). Ist $\mu_i = \mathrm{Unif}_{(a_i,b_i)}$ die Gleichverteilung auf einem endlichen Intervall $(a_i, b_i), -\infty < a_i < b_i < \infty$, dann ist $\mu = \mu_1 \otimes \ldots \otimes \mu_d$ die Gleichverteilung auf dem Quader $S = (a_1, b_1) \times \ldots \times (a_d, b_d)$, denn für die Dichtefunktion gilt:

$$f(x_1, \ldots, x_d) \quad = \quad \prod_{i=1}^{d} \frac{I_{(a_i,b_i)}(x_i)}{b_i - a_i} \quad = \quad \frac{I_S(x)}{\mathrm{vol}[S]}.$$

Ein anderes Produktmaß von fundamentaler Bedeutung für die Wahrscheinlichkeitstheorie ist die mehrdimensionale Standardnormalverteilung:

Beispiel (Standardnormalverteilung im $\mathbb{R}^d$). Das Produkt $\mu = \bigotimes_{i=1}^{d} N(0,1)$ von d eindimensionalen Standardnormalverteilungen ist eine absolutstetige Wahrscheinlichkeitsverteilung auf $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ mit Dichte

$$f(x_1, \ldots, x_d) \quad = \quad \prod_{i=1}^{d} \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2} \right) \right) \quad = \quad (2\pi)^{-d/2} \exp\left(-\frac{\|x\|^2}{2} \right), \qquad x \in \mathbb{R}^d.$$

Die Wahrscheinlichkeitsverteilung μ heißt *d-dimensionale Standardnormalverteilung*.

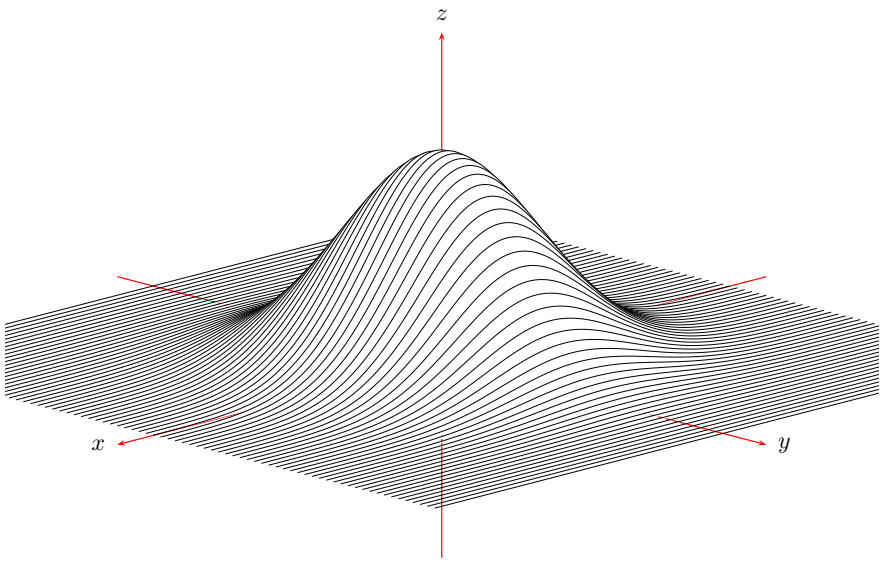


Abbildung 4.14: Dichte der Standardnormalverteilung im $\mathbb{R}^2$.

Allgemeinere mehrstufige Modelle kann man im absolutstetigen Fall ähnlich wie im diskreten Fall konstruieren. Die Dichte hat dann die Form

$$f(x_1, \ldots, x_d) \;=\; f_1(x_1) \cdot f_2(x_2|x_1) \cdot f_3(x_3|x_1, x_2) \cdot \;\cdots\; \cdot f_d(x_d|x_1, \ldots x_{d-1})$$

wobei f_1 die Marginaldichte der ersten Komponente ist, und $f_2, \ldots, f_d$ *bedingte Dichten* sind.

Gemeinsame Verteilung von Zufallsvariablen

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, und seien $X_1, X_2, \ldots, X_d$ Abbildungen von $\Omega \to \mathbb{R}$. Wir versehen $\mathbb{R}^d$ mit der Borelschen σ-Algebra $\mathcal{B}(\mathbb{R}^d)$. Dann ist die Abbildung $X = (X_1, \ldots, X_d)$ mit Werten im $\mathbb{R}^d$ eine *Zufallsvariable* (oder ein *Zufallsvektor*), falls

$$X^{-1}(B) \;=\; \{X \in B\} \;\in\; \mathcal{A} \qquad \text{für alle } B \in \mathcal{B}(\mathbb{R}^d)\,.$$

Die *Verteilung* μ_X von X bezüglich P ist in diesem Fall die durch

$$\mu_X[B] \;=\; P[X \in B] \;=\; P[(X_1, \dots X_d) \in B], \qquad B \in \mathcal{B}(\mathbb{R}^d),$$

definierte Wahrscheinlichkeitsverteilung.

Lemma 4.29. *Eine Abbildung $X = (X_1, \dots, X_d) : \Omega \to \mathbb{R}^d$ ist genau dann eine Zufallsvariable, wenn die Komponentenabbildungen $X_1, \dots, X_d$ reellwertige Zufallsvariablen sind.*

Beweis. Die Borelsche σ-Algebra auf $\mathbb{R}^d$ wird erzeugt von den Mengen $(-\infty, c_1] \times \cdots \times (-\infty, c_d]$ mit $c_1, \dots, c_d \in \mathbb{R}$. Daher ist X genau dann eine Zufallsvariable mit Werten im $\mathbb{R}^d$, wenn die Mengen

$$\{X_1 \le c_1, \dots, X_d \le c_d\} \;=\; \{X \in (-\infty, c_1] \times \cdots \times (-\infty, c_d]\}$$

alle in $\mathcal{A}$ enthalten sind. Dies ist aber genau dann der Fall, wenn $\{X_i \le c_i\} \in \mathcal{A}$ für alle $i = 1, \dots, d$ und $c_1, \dots, c_d \in \mathbb{R}$ gilt, also wenn die Komponenten $X_1, \dots, X_d$ Zufallsvariablen sind. ∎

Wie im diskreten Fall heißt die Verteilung μ_X des Zufallsvektors $X = (X_1, \dots, X_d)$ *gemeinsame Verteilung* von $X_1, \dots, X_d$. Aus der gemeinsamen Verteilung können wir auch die *Randverteilungen* berechnen, d.h. die Verteilungen μ_{X_i} der einzelnen Zufallsvariablen $X_1, \dots, X_d$. Beispielsweise gilt

$$\mu_{X_1}[B] \;=\; P[X_1 \in B] \;=\; P[X \in B \times \mathbb{R} \times \cdots \times \mathbb{R}] \;=\; \mu_X[B \times \mathbb{R} \times \cdots \times \mathbb{R}] \quad \text{für alle } B \in \mathcal{B}(\mathbb{R}).$$

Lemma 4.30. *Ist die gemeinsame Verteilung μ_X der Zufallsvariablen $X_1, \dots, X_d$ absolutstetig mit Dichte $f(x_1, \dots, x_d)$, dann sind auch die Randverteilungen $\mu_{X_1}, \dots, \mu_{X_d}$ absolutstetig mit Dichten*

$$f_{X_i}(x_i) \;=\; \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(x_1, \dots, x_d)\, dx_1\, dx_2 \cdots dx_{i-1}\, dx_{i+1} \cdots dx_d. \tag{4.27}$$

Beweis. Ohne Beschränkung der Allgemeinheit zeigen wir die Aussage im Fall $i = 1$. Hier gilt nach dem Satz von Fubini

$$\begin{aligned} \mu_{X_1}[B] \;&=\; \int \cdots \int I_{B \times \mathbb{R}^{d-1}}(x_1, \dots, x_d)\, f(x_1, \dots, x_d)\, d_{x_d} \cdots d_{x_1} \\ &=\; \int_B \left(\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(x_1, \dots, x_d)\, dx_2 \cdots dx_d \right) dx_1 \end{aligned}$$

für alle $B \in \mathcal{B}(\mathbb{R})$, woraus die Behauptung folgt. ∎

Beispiel. Die gemeinsame Verteilung der Zufallsvariablen X_1 und X_2 sei die Gleichverteilung auf dem Dreieck

$$A \;=\; \left\{ x \in \mathbb{R}^2 \,:\, x_1 \ge 0,\; x_2 \le 1,\; x_1 \le x_2 \right\}.$$

Da das Dreieck den Flächeninhalt $1/2$ hat, ist

$$f(x_1, x_2) \;=\; \frac{1}{1/2}\, I_A(x_1, x_2)$$

die Dichte der gemeinsamen Verteilung. Damit erhalten wir als Verteilungsdichte von X_1:

$$\begin{aligned} f_{X_1}(x_1) \;&=\; \int_{\mathbb{R}} f(x_1, x_2)\, dx_2 \;=\; 2 \int I_{x_1 \ge 0,\, x_1 \le x_2 \le 1}\, dx_2 \\ &=\; 2 \cdot I_{x_1 \ge 0} \cdot \int I_{x_1 \le x_2 \le 1}\, dx_2 \;=\; 2 \cdot (1 - x_1) \cdot I_{0 \le x_1 \le 1}\,. \end{aligned}$$

Unabhängigkeit von Zufallsvariablen

Unabhängigkeit von allgemeinen reellwertigen Zufallsvariablen können wir ähnlich wie im diskreten Fall definieren. Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, und I eine beliebige Menge.

Definition 4.31. Eine Familie $X_i : \Omega \to \mathbb{R}$ $(i \in I)$ von reellwertigen Zufallsvariablen auf $(\Omega, \mathcal{A}, P)$ heißt **unabhängig**, falls die Ereignisse $\{X_i \in B_i\}$ $(i \in I)$ für alle Mengen $B_i \in \mathcal{B}(\mathbb{R})$ unabhängig sind.

Aus der Definition folgt unmittelbar, dass die Zufallsvariablen X_i $(i \in I)$ genau dann unabhängig sind, wenn jede endliche Teilkollektion unabhängig ist. Daher beschränken wir uns im folgenden wieder auf den Fall $I = \{1, \dots, d\}$ mit $d \in \mathbb{N}$.

Satz 4.32. Die folgenden Aussagen sind äquivalent:

(i) $X_1, \dots, X_d$ sind unabhängig.

(ii) Die Ereignisse $\{X_1 \le c_1\}, \dots, \{X_d \le c_d\}$ sind unabhängig für alle $c_1, \dots, c_d \in \mathbb{R}$.

(iii) $F_{X_1,\dots,X_d}(c_1, \dots, c_d) = \prod_{i=1}^d F_{X_i}(c_i)$ für alle $c_1, \dots, c_d \in \mathbb{R}$.

(iv) $\mu_{X_1,\dots,X_d} = \bigotimes_{i=1}^d \mu_{X_i}$.

Hierbei bezeichnet $F_{X_1,\dots,X_d}$ die Verteilungsfunktion der gemeinsamen Verteilung $\mu_{X_1,\dots,X_d}$, siehe Korollar 4.27.

Beweis. Da die Borelsche σ-Algebra auf $\mathbb{R}^d$ von den Mengen $(-\infty, c_1] \times \cdots \times (-\infty, c_d]$ erzeugt wird, folgt die Äquivalenz der ersten beiden Aussagen mithilfe des Eindeutigkeitssatzes 4.2. Die Äquivalenz der übrigen Aussagen zeigt man ähnlich wie im Beweis von Satz 2.14. ■

Im absolutstetigen Fall ergibt sich die folgende Aussage als Konsequenz aus Satz 4.32.

Korollar 4.33. Absolutstetige Zufallsvariablen $X_i : \Omega \to \mathbb{R}$ $(i = 1, \dots, d)$ mit Dichtefunktionen f_i sind genau dann unabhängig, wenn die gemeinsame Verteilung absolutstetig ist, und die Dichte der gemeinsamen Verteilung eine Darstellung in Produktform

$$f_{X_1,\dots,X_d}(x_1, \dots, x_d) = c \cdot \prod_{i=1}^d g_i(x_i) \qquad \forall\, (x_1, \dots, x_d) \in \mathbb{R}^d \tag{4.28}$$

mit Funktionen $g_i : \mathbb{R} \to [0, \infty)$ und einer Proportionalitätskonstanten $c \in \mathbb{R}$ hat. In diesem Fall sind die Marginaldichten f_i proportional zu g_i, und es gilt

$$f_{X_1,\dots,X_d}(x_1, \dots, x_d) = \prod_{i=1}^d f_i(x_i), \qquad x \in \mathbb{R}^d.$$

Summen unabhängiger Zufallsvariablen

Seien $X, Y : \Omega \to \mathbb{R}$ unabhängige Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Im diskreten Fall können wir die Massenfunktion der Verteilung von $X + Y$ mithilfe der Faltungsformel (2.14) aus den Massenfunktionen der Verteilungen von X und Y berechnen. Eine entsprechende Aussage gilt auch im absolutstetigen Fall.

Satz 4.34 (Faltung von absolutstetigen Wahrscheinlichkeitsverteilungen). Sind X und Y unabhängige Zufallsvariablen mit absolutstetigen Verteilungen mit Dichten f_X und f_Y, dann ist auch die Verteilung von $X+Y$ absolutstetig mit Dichte

$$f_{X+Y}(z) \;=\; (f_X * f_Y)(z) \;:=\; \int_{-\infty}^{\infty} f_X(x)\, f_Y(z-x)\, dx\,. \tag{4.29}$$

Beweis. Da X und Y unabhängig sind, hat die gemeinsame Verteilung die Produktdichte $f_X(x) f_Y(y)$. Mithilfe der Substitution $z = x + y$ erhalten wir

$$\begin{aligned} P[X+Y \le c] &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{c-x} f_X(x)\, f_Y(y)\, dy\, dx \;=\; \int_{x=-\infty}^{\infty} \int_{z=-\infty}^{c} f_X(x)\, f_Y(z-x)\, dz\, dx \\ &= \int_{z=-\infty}^{c} \int_{x=-\infty}^{\infty} f_X(x)\, f_Y(z-x)\, dx\, dz \;=\; \int_{s=-\infty}^{c} (f_X * f_Y)(z)\, dz \end{aligned}$$

für alle $c \in \mathbb{R}$. Dabei haben wir den Satz von Fubini benutzt, um die Integrationsreihenfolge zu vertauschen. ■

Beispiel (Faltung von Gleichverteilungen). Sind X und Y unabhängige, auf dem Intervall $(0,1)$ gleichverteilte Zufallsvariablen, dann hat $X+Y$ die Verteilungsdichte

$$f_{X+Y}(z) \;=\; \int_{-\infty}^{\infty} I_{(0,1)}(x)\, I_{(0,1)}(z-x)\, dx \;=\; \int_{\max(0,z-1)}^{\min(z,1)} 1\, dx \;=\; \begin{cases} z & \text{für } 0 \le z \le 1, \\ 2-z & \text{für } 1 < z \le 2, \\ 0 & \text{sonst.} \end{cases}$$

Werte nahe 1 sind also für $X+Y$ wahrscheinlicher als Werte nahe 0 oder 2.

Beispiel (Faltung von Normalverteilungen). Sind X und Y unabhängig und normalverteilt mit Parametern (m, v) und $(\widetilde{m}, \widetilde{v})$, dann ist $X+Y$ normalverteilt mit Parametern $(m+\widetilde{m}, v+\widetilde{v})$. Dies kann man zum Beispiel mithilfe der Faltungsformel nachrechnen.

Berechnung von Erwartungswerten und Kovarianzen

Erwartungswerte der Form $E[g(X_1, X_2, \ldots, X_d)]$ kann man für reellwertige Zufallsvariablen $X_1, X_2, \ldots, X_d$, deren gemeinsame Verteilung absolutstetig mit Dichte $f(x_1, x_2, \ldots, x_d)$ ist, ähnlich wie im eindimensionalen Fall berechnen.

Satz 4.35. Ist die Funktion $g \cdot f : \mathbb{R}^d \to \mathbb{R}$ integrierbar, dann gilt

$$E[g(X_1, X_2, \ldots, X_d)] \;=\; \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, x_2, \ldots, x_d)\, f(x_1, x_2, \ldots, x_d)\, dx_d \cdots dx_2\, dx_1\,. \tag{4.30}$$

Insbesondere gilt für alle Mengen $B \in \mathcal{B}(\mathbb{R}^d)$:

$$P[(X_1, X_2, \ldots, X_d) \in B] \;=\; \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} I_B(x_1, x_2, \ldots, x_d)\, f(x_1, x_2, \ldots, x_d)\, dx_d \cdots dx_2\, dx_1\,. \tag{4.31}$$

Der Beweis verläuft ähnlich wie der oben skizzierte Beweis von Satz 4.22. Beispielsweise erhalten wir für zwei reellwertige Zufallsvariablen X und Y mit $E[X^2] < \infty$ und $E[Y^2] < \infty$:

$$\begin{aligned} \mathrm{Cov}[X,Y] &= E[XY] - E[X]\,E[Y] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, y\, f(x,y)\, dy\, dx \;-\; \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f(x,y)\, dy\, dx \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, f(x,y)\, dy\, dx\,. \end{aligned} \tag{4.32}$$

Beispiel (Zweidimensionale Normalverteilung mit Korrelation). Ist die gemeinsame Verteilung von X und Y die zweidimensionale Standardnormalverteilung mit Dichte $f(x,y) = (2\pi)^{-1} \exp\left(-(x^2+y^2)/2\right)$, dann sind X und Y unkorreliert. Allgemeiner gilt für $\varrho \in (-1,1)$: Ist die gemeinsame Verteilung von X und Y die zweidimensionale Normalverteilung mit Dichtefunktion

$$f_\varrho(x,y) \;=\; \frac{1}{2\pi\sqrt{1-\varrho^2}} \exp\left[-\frac{x^2 - 2\varrho xy + y^2}{2(1-\varrho^2)}\right],$$

dann sind X und Y standardnormalverteilt mit Korrelation

$$\varrho[X,Y] \;=\; \mathrm{Cov}[X,Y] \;=\; \varrho\,.$$

Der Beweis ist eine Übungsaufgabe.

5 Grenzwertsätze und Statistik

5.1 Grenzwertsätze

Von der Binomialverteilung zur Normalverteilung

Die Binomialverteilung mit Parametern n und p beschreibt die Verteilung der Anzahl derjenigen unter n unabhängigen Ereignissen mit Wahrscheinlichkeit p, die in einem Zufallsexperiment eintreten. Viele Anwendungsprobleme führen daher auf die Berechnung von Wahrscheinlichkeiten bzgl. der Binomialverteilung. Für große n ist eine exakte Berechnung dieser Wahrscheinlichkeiten aber in der Regel nicht mehr möglich. Bei seltenen Ereignissen kann man die Poissonapproximation zur näherungsweisen Berechnung nutzen:

Konvergiert $n \to \infty$, und konvergiert gleichzeitig der Erwartungswert $n \cdot p_n$ gegen eine positive reelle Zahl $\lambda > 0$, dann nähern sich die Gewichte $b_{n,p_n}(k)$ der Binomialverteilung denen einer Poissonverteilung mit Parameter λ an:

$$b_{n,p_n}(k) \;=\; \binom{n}{k} p_n^k (1-p_n)^{n-k} \;\to\; \frac{\lambda^k}{k!} e^{-\lambda} \qquad (k = 0, 1, 2, \ldots),$$

siehe Satz 1.11. Geht die Wahrscheinlichkeit p_n für $n \to \infty$ nicht gegen 0, sondern hat zum Beispiel einen festen Wert $p \in (0,1)$, dann kann die Poissonapproximation nicht verwendet werden. Stattdessen scheinen sich die Gewichte der Binomialverteilung einer Gaußschen Glockenkurve anzunähern, siehe z.B. Abbildung 4.3 oder die Mathematica-Demonstrationen auf der Vorlesungshomepage. Wir wollen diese Aussage nun mathematisch präzisieren und beweisen.

Wir analysieren zunächst das asymptotische Verhalten von Binomialkoeffizienten mithilfe der Stirlingschen Formel.

Definition 5.1 (Asymptotische Äquivalenz von Folgen). Zwei Folgen $a_n, b_n \in \mathbb{R}_+$ ($n \in \mathbb{N}$), heißen *asymptotisch äquivalent* ($a_n \sim b_n$), falls

$$\lim_{n\to\infty} \frac{a_n}{b_n} \;=\; 1 \qquad \text{gilt.}$$

Bemerkung. Für Folgen mit Werten $a_n, b_n, c_n, d_n \in \mathbb{R}_+$ gilt :

(i) $a_n \sim b_n \iff \exists\, \varepsilon_n \to 0 : a_n = b_n(1+\varepsilon_n) \iff \log a_n - \log b_n \to 0,$

(ii) $a_n \sim b_n \iff b_n \sim a_n \iff \frac{1}{a_n} \sim \frac{1}{b_n},$

(iii) $a_n \sim b_n, c_n \sim d_n \implies a_n \cdot c_n \sim b_n \cdot d_n.$

Satz 5.2 (Stirlingsche Formel).

$$n! \;\sim\; \sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n$$

Einen Beweis der Stirlingschen Formel findet man in vielen Analysis-Büchern, siehe z.B. Forster:„Analysis 1". Dass der Quotient der beiden Terme in der Stirlingschen Formel beschränkt ist, sieht man rasch durch eine Integralapproximation von $\log n!$:

$$\begin{aligned}\log n! &= \sum_{k=1}^{n} \log k = \int_0^n \log x\, dx + \sum_{k=1}^{n} \int_{k-1}^{k} (\log k - \log x)\, dx \\ &= n\log n - n + \frac{1}{2}\sum_{k=1}^{n} \frac{1}{k} + O(1) \\ &= n\log n - n + \frac{1}{2}\log n + O(1),\end{aligned}$$

wobei wir im vorletzten Schritt die Taylor-Approximationen $\log k - \log x = \frac{1}{k}(k-x) + O(k^{-2})$ auf den Intervallen $[k-1,k]$, $k \geq 2$, verwendet haben. Ein alternativer, sehr kurzer vollständiger Beweis der Stirling-Formel mit Identifikation des korrekten asymptotischen Vorfaktors $\sqrt{2\pi}$ wird in J.M. Patin, The American Mathematical Monthly 96 (1989) gegeben.

Mithilfe der Stirlingschen Formel können wir die Gewichte

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

der Binomialverteilung für große n und k approximieren. Sei dazu S_n eine $\mathrm{Bin}(n,p)$-verteilte Zufallsvariable auf $(\Omega, \mathcal{A}, P)$. Für den Erwartungswert und die Standardabweichung von S_n gilt:

$$E[S_n] = np \quad \text{und} \quad \sigma[S_n] = \sqrt{\mathrm{Var}[S_n]} = \sqrt{np(1-p)}.$$

Dies deutet darauf hin, dass sich die Masse der Binomialverteilung für große n überwiegend in einer Umgebung der Größenordnung $O(\sqrt{n})$ um np konzentriert.

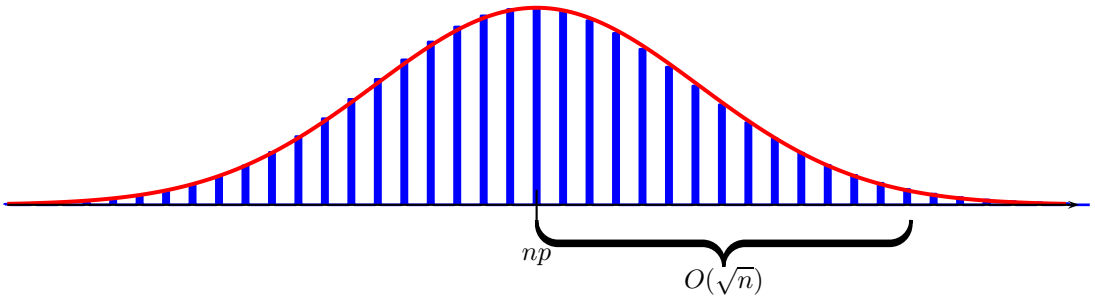


Abbildung 5.1: Die Gewichte der Binomialverteilung liegen für große n näherungsweise auf einer Glockenkurve mit Mittel np und Standardabweichung $\sqrt{np(1-p)}$.

Wir werden nun die Gewichte

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

der Binomialverteilung für große n und k in Umgebungen der Größenordnung $O(\sqrt{n})$ von np ausgehend von der Stirlingschen Formel approximieren, und die vermutete asymptotische Darstellung präzisieren und beweisen. Dazu führen wir noch folgende Notation ein: Wir schreiben

$$a_n(k) \approx b_n(k) \qquad (\text{„lokal gleichmäßig asymptotisch äquivalent"}),$$

falls

$$\lim_{n\to\infty} \sup_{k\in U_{n,r}} \left| \frac{a_n(k)}{b_n(k)} - 1 \right| \;=\; 0 \qquad \text{für alle } r \in \mathbb{R}_+ \text{ gilt,}$$

wobei

$$U_{n,r} \quad = \quad \{0 \le k \le n : |k - np| \le r \cdot \sqrt{n}\}.$$

Die Aussagen aus der Bemerkung oben gelten analog für diese Art der lokal gleichmäßigen asymptotischen Äquivalenz von $a_n(k)$ und $b_n(k)$.

Satz 5.3 (Grenzwertsatz von De Moivre (1733) und Laplace (1819)). Sei S_n binomialverteilt mit Parametern $n \in \mathbb{N}$ und $p \in (0,1)$, und sei $\sigma^2 = p(1-p)$. Dann gilt:

(i) $$P[S_n = k] \quad = \quad b_{n,p}(k) \quad \approx \quad \widetilde{b}_{n,p}(k) \quad := \quad \frac{1}{\sqrt{2\pi n\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(\frac{k-np}{\sqrt{n}} \right)^2 \right).$$

(ii) $$P\left[a \le \frac{S_n - np}{\sqrt{n}} \le b \right] \quad \overset{n\nearrow\infty}{\longrightarrow} \quad \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{x^2}{2\sigma^2} \right) dx \qquad \text{für alle } a, b \in \mathbb{R} \text{ mit } a \le b.$$

Beweis. (i). Wir beweisen die Aussage in zwei Schritten:

(a) Wir zeigen zunächst mithilfe der *Stirlingschen Formel*, dass

$$b_{n,p}(k) \quad \approx \quad \overline{b}_{n,p}(k) \quad := \quad \frac{1}{\sqrt{2\pi n \frac{k}{n}(1-\frac{k}{n})}} \cdot \left(\frac{p}{k/n} \right)^k \cdot \left(\frac{1-p}{1-k/n} \right)^{n-k} \tag{5.1}$$

gilt. Nach der Stirlingschen Formel ist

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\,(n/e)^n} \quad = \quad 1.$$

Wegen $k \ge np - r \cdot \sqrt{n}$ für $k \in U_{n,r}$ folgt

$$\sup_{k\in U_{n,r}} \left| \frac{k!}{\sqrt{2\pi k}\,(k/e)^k} - 1 \right| \quad \longrightarrow \quad 0 \qquad \text{für } n \to \infty,$$

d.h.

$$k! \quad \approx \quad \sqrt{2\pi k}\,(k/e)^k\,.$$

Analog erhält man

$$(n-k)! \quad \approx \quad \sqrt{2\pi(n-k)}\,((n-k)/e)^{n-k}\,,$$

und damit

$$\begin{aligned} b_{n,p}(k) \quad &= \quad \frac{n!}{k!\cdot(n-k)!}\, p^k (1-p)^{n-k} \quad \approx \quad \frac{\sqrt{2\pi n}\cdot n^n \cdot p^k \cdot (1-p)^{n-k}}{2\pi\sqrt{k(n-k)}\cdot k^k \cdot (n-k)^{n-k}} \\ &= \quad \sqrt{\frac{n}{2\pi k(n-k)}} \left(\frac{np}{k} \right)^k \left(\frac{n(1-p)}{n-k} \right)^{n-k} \quad = \quad \overline{b}_{n,p}(k). \end{aligned}$$

(b) Wir zeigen nun weiterhin mithilfe einer *Taylorapproximation*, dass

$$\overline{b}_{n,p}(k) \quad \approx \quad \widetilde{b}_{n,p}(k) \tag{5.2}$$

gilt. Dazu benutzen wir mehrfach die Abschätzung

$$\left|\frac{k}{n} - p\right| \quad \leq \quad r \cdot n^{-\frac{1}{2}} \qquad \text{für } k \in U_{n,r}.$$

Hieraus folgt zunächst unmittelbar

$$\sqrt{2\pi n \frac{k}{n}\left(1 - \frac{k}{n}\right)} \quad \approx \quad \sqrt{2\pi np(1-p)} \quad = \quad \sqrt{2\pi n\sigma^2}. \tag{5.3}$$

Um die Asymptotik der übrigen Faktoren von $\overline{b}_{n,p}(k)$ zu erhalten, nehmen wir den Logarithmus, und verwenden die *Taylorapproximation*

$$x \log \frac{x}{p} + (1-x) \log \frac{1-x}{1-p} = \frac{1}{2p(1-p)}(x-p)^2 + O(|x-p|^3),$$

die man durch Berechnen der Ableitungen der Funktion auf der linken Seite an der Stelle $x = p$ verifiziert. Wir erhalten

$$\begin{aligned} \frac{1}{n} \log \left[\left(\frac{p}{k/n}\right)^k \left(\frac{1-p}{1-k/n}\right)^{n-k}\right] &= -\frac{k}{n} \log\left(\frac{k/n}{p}\right) - \left(1 - \frac{k}{n}\right) \log\left(\frac{1-k/n}{1-p}\right) \\ &= -\frac{1}{2p(1-p)}\left(\frac{k}{n} - p\right)^2 + O\left(|\frac{k}{n} - p|^3\right). \end{aligned}$$

Wegen $\left|\frac{k}{n} - p\right|^3 \quad \leq \quad r^3 \cdot n^{-\frac{3}{2}}$ für $k \in U_{n,r}$ folgt

$$\log\left(\left(\frac{p}{k/n}\right)^k \left(\frac{1-p}{1-k/n}\right)^{n-k}\right) \quad = \quad -\frac{n}{2\sigma^2}\left(\frac{k}{n} - p\right)^2 + R_{k,n},$$

wobei $|R_{k,n}| \leq \text{const.} \cdot r^3 n^{-\frac{1}{2}}$ für alle $k \in U_{n,r}$, d.h.

$$\left(\frac{p}{k/n}\right)^k \left(\frac{1-p}{1-k/n}\right)^{n-k} \quad \approx \quad \exp\left(-\frac{n}{2\sigma^2}\left(\frac{k}{n} - p\right)^2\right). \tag{5.4}$$

Aussage (5.2) folgt dann aus (5.3) und (5.4).

(c) Aus (a) und (b) folgt nun Behauptung (i).

(ii). Aufgrund von (i) erhalten wir für $a, b \in \mathbb{R}$ mit $a < b$:

$$P\left[a \leq \frac{S_n - np}{\sqrt{n}} \leq b\right] = \sum_{\substack{k \in \{0,1,\ldots,n\} \\ a \leq \frac{k-np}{\sqrt{n}} \leq b}} b_{n,p}(k) = \sum_{\substack{k \in \{0,1,\ldots,n\} \\ a \leq \frac{k-np}{\sqrt{n}} \leq b}} \widetilde{b}_{n,p}(k)(1 + \varepsilon_{n,p}(k)),$$

wobei

$$\overline{\varepsilon}_{n,p} \quad := \quad \sup_{a \leq \frac{k-np}{\sqrt{n}} \leq b} |\varepsilon_{n,p}(k)| \quad \longrightarrow \quad 0 \qquad \text{für } n \to \infty. \tag{5.5}$$

Wir zeigen nun

$$\lim_{n\to\infty} \sum_{\substack{k\in\{0,1,\ldots,n\}\\ a\le\frac{k-np}{\sqrt{n}}\le b}} \widetilde{b}_{n,p}(k) \quad = \quad \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \tag{5.6}$$

Zum Beweis von (5.6) bemerken wir, dass

$$\Gamma_n \ := \ \left\{\frac{k-np}{\sqrt{n}} : k = 0,1,\ldots,n\right\} \ \subseteq \ \mathbb{R}$$

ein äquidistantes Gitter mit Maschenweite $1/\sqrt{n}$ ist. Es gilt

$$\sum_{\substack{k\in\{0,1,\ldots,n\}\\ a\le\frac{k-np}{\sqrt{n}}\le b}} \widetilde{b}_{n,p}(k) \quad = \quad \sum_{\substack{x\in\Gamma_n\\ a\le x\le b}} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{x^2}{2\sigma^2}}\, \frac{1}{\sqrt{n}}\,. \tag{5.7}$$

Für $n \to \infty$ folgt dann (5.6), da die rechte Seite in (5.7) eine Riemannsummenapproximation des Integrals in (5.6) ist, und der Integrand stetig ist. Die Behauptung folgt nun aus (5.5) und (5.6). ■

Der Satz von De Moivre/Laplace impliziert, dass die Zufallsvariablen $\frac{S_n-np}{\sqrt{n}}$ für $n \to \infty$ *in Verteilung gegen eine $N(0,\sigma^2)$-verteilte Zufallsvariable mit Varianz $\sigma^2 = p(1-p)$ konvergieren*:

$$\frac{S_n - np}{\sqrt{n}} \quad \xrightarrow{\mathcal{D}} \quad \sigma Z \qquad \text{mit } Z \sim N(0,1).$$

Hierbei bedeutet *Konvergenz in Verteilung (convergence in distribution; Notation „$\xrightarrow{\mathcal{D}}$“)*, dass die Verteilungen der Zufallsvariablen *schwach konvergieren*, d.h. für die Verteilungsfunktionen gilt

$$\lim_{n\to\infty} F_{\frac{S_n-np}{\sqrt{n}}}(c) \ = \ \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{x^2}{2\sigma^2}}\, dx \ = \ F_{\sigma Z}(c) \qquad \text{für alle } c \in \mathbb{R}. \tag{5.8}$$

Konvergenz in Verteilung ist nicht wirklich ein Konvergenzbegriff für Zufallsvariablen, sondern „nur” eine Konvergenz der Verteilungen. Für die standardisierten (d.h. auf Erwartungswert 0 und Varianz 1 normierten) Zufallsvariablen gilt entsprechend

$$\frac{S_n - E[S_n]}{\sigma(S_n)} \quad = \quad \frac{S_n - np}{\sigma\sqrt{n}} \quad \xrightarrow{\mathcal{D}} \quad Z. \tag{5.9}$$

Bemerkung. (i) Die Aussage (5.9) ist ein Spezialfall eines viel allgemeineren *zentralen Grenzwertsatzes*: Sind $X_1, X_2, \ldots$ unabhängige, identisch verteilte Zufallsvariablen mit endlicher Varianz, und ist $S_n = X_1 + \ldots + X_n$, dann konvergieren die Verteilungen der standardisierten Summen

$$\frac{S_n - E[S_n]}{\sigma(S_n)}$$

schwach gegen eine Standardnormalverteilung. Normalverteilungen treten also als universelle Skalierungslimiten der Verteilungen von Summen unabhängiger Zufallsvariablen auf.

(ii) Heuristisch gilt für große n nach (5.9)

$$„ \quad S_n \quad \overset{\mathcal{D}}{\approx} \quad np + \sqrt{np(1-p)} \cdot Z, \quad “ \tag{5.10}$$

wobei „$\overset{\mathcal{D}}{\approx}$“ dafür steht, dass die Verteilungen der Zufallsvariablen sich einander in einem gewissen Sinn annähern. In diesem Sinne wäre für große n

$$\text{„Bin}(n,p) \quad \approx \quad N(np, np(1-p))\text{.“}$$

Entsprechende „Approximationen“ werden häufig in Anwendungen benutzt, sollten aber hinterfragt werden, da beim Übergang von (5.9) zu (5.10) mit dem divergierende Faktor $\sqrt{n}$ multipliziert wird. Die mathematische Präzisierung entsprechender heuristischer Argumentationen erfolgt üblicherweise über den Satz von De Moivre/Laplace.

Beispiel (Normalapproximation der Binomialverteilung). Seien $X_1, X_2, \ldots$ unabhängige Zufallsvariablen mit $P[X_i = 0] = P[X_i = 1] = \frac{1}{2}$, und sei $S_n = X_1 + \ldots + X_n$ (z.B. Häufigkeit von „Zahl“ bei n fairen Münzwürfen). In diesem Fall ist $p = 1/2$ und $\sigma = \sqrt{p(1-p)} = 1/2$.

(i) *100 faire Münzwürfe:* Für die Wahrscheinlichkeit, dass mehr als 60 mal Zahl fällt, gilt

$$P[S_{100} > 60] \quad = \quad P[S_{100} - E[S_{100}] > 10] \quad = \quad P\left[\frac{S_{100} - E[S_{100}]}{\sigma(S_{100})} > \frac{10}{\sigma\sqrt{100}}\right].$$

Da $\frac{S_{100}-E[S_{100}]}{\sigma(S_{100})}$ nach (5.9) näherungsweise $N(0,1)$-verteilt ist, und $\frac{10}{\sigma\sqrt{100}} = 2$ gilt, folgt

$$P[S_{100} > 60] \quad \approx \quad P[Z > 2] \quad = \quad 1 - \Phi(2) \quad \approx \quad 0.0227 \quad = \quad 2.27\%.$$

(ii) *16 faire Münzwürfe*: Für die Wahrscheinlichkeit, dass genau 8 mal Zahl fällt, ergibt sich

$$P[S_{16} = 8] \quad = \quad P[7.5 \le S_{16} \le 8.5] \; = \; P\left[\left|\frac{S_{16} - E[S_{16}]}{\sigma(S_{16})}\right| \le \frac{0.5}{\sigma\sqrt{16}}\right]$$

Mit $\frac{0.5}{\sigma\sqrt{16}} = \frac{1}{4}$ folgt näherungsweise

$$P[S_{16} = 8] \quad \approx \quad P[|Z| \le 1/4] \quad = \quad 0.1974\ldots.$$

Der exakte Wert beträgt $P[S_{16} = 8] = 0.1964\ldots$. Bei geschickter Anwendung ist die Normalapproximation oft schon für eine kleine Anzahl von Summanden relativ genau!

Der zentrale Grenzwertsatz

Seien $X_1, X_2, \ldots$ unabhängige, identisch verteilte reellwertige Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit Erwartungswert $m \in \mathbb{R}$ und endlicher Varianz σ^2, und sei $S_n = X_1 + \ldots + X_n$. Nach dem Gesetz der großen Zahlen gilt

$$\frac{S_n}{n} \quad \to \quad m \qquad P\text{-fast sicher und } P\text{-stochastisch.}$$

Wie sieht die Verteilung von S_n für große n aus?

Um eine asymptotische Darstellung zu erhalten, reskalieren wir zunächst wieder so, dass die Erwartungswerte gleich 0 und die Varianzen konstant sind. Es gilt

$$E[S_n] \; = \; n \cdot m \qquad \text{und} \qquad \mathrm{Var}[S_n] \; = \; n \cdot \sigma^2,$$

also hat die Zufallsvariable

$$\widetilde{S}_n \; := \; \frac{S_n - n \cdot m}{\sqrt{n}} \; = \; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - m)$$

unabhängig von n den Erwartungswert $E\left[\widetilde{S}_n\right] = 0$ und die Varianz

$$\mathrm{Var}\left[\widetilde{S}_n\right] \; = \; \mathrm{Var}\left[\frac{S_n}{\sqrt{n}}\right] \; = \; \frac{1}{n} \cdot \mathrm{Var}[S_n] \; = \; \sigma^2.$$

.

Satz 5.4 (Zentraler Grenzwertsatz). Seien $X_1, X_2, \ldots$ unabhängige, identisch verteilte reellwertige Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ mit Erwartungswert $m \in \mathbb{R}$ und endlicher Varianz $\sigma^2 > 0$, und sei

$$S_n \;=\; X_1 + \ldots + X_n \;.$$

Dann konvergieren die Verteilungen der standardisierten Summen $\widetilde{S_n}$ für $n \to \infty$ schwach gegen $N(0, \sigma^2)$, d.h., für alle $a, b \in \mathbb{R}$ mit $a \le b$ gilt

$$P\left[a \le \frac{S_n - np}{\sqrt{n}} \le b\right] \quad \overset{n\to\infty}{\longrightarrow} \quad \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \;.$$

Bemerkung. (i) Alternativ kann man die standardisierten Summen auf Varianz 1 normieren, und erhält

$$\text{Verteilung}\left(\frac{S_n - E[S_n]}{\sigma \cdot \sqrt{n}}\right) \quad \overset{n\to\infty}{\longrightarrow} \quad \text{Verteilung}\,(Z)\,,$$

wobei Z eine standardnormalverteilte Zufallsvariable ist.

(ii) Die Limesverteilung im zentralen Grenzwertsatz ist unabhängig von der Verteilung von X_1, vorausgesetzt, es gilt $E[X_1^2] < \infty$. Die Normalverteilung tritt also als *universeller Skalierungslimes* der Verteilungen von Summen unabhängiger identisch verteilter Zufallsvariablen auf. Dies erklärt, warum die Normalverteilungen in der Stochastik von so großer Bedeutung sind.

Beispiele. (i) Sind $X_1, X_2, \ldots$ unabhängige Zufallsvariablen mit $P[X_i = 1] = p$ und $P[X_i = 0] = 1 - p$, dann ist $S_n = \sum_{i=1}^n X_i$ binomialverteilt mit Parametern n und p. Die Aussage des Zentralen Grenzwertsatzes folgt in diesem Fall aus dem Satz von de Moivre/Laplace.

(ii) Sind die Zufallsvariablen X_i unabhängig und Poisson-verteilt mit Parameter $\lambda > 0$, dann ist $S_n = \sum_{i=1}^n X_i$ Poisson-verteilt mit Parameter $n\lambda$. Der Zentrale Grenzwertsatz liefert in diesem Fall eine Normalapproximation für Poisson-Verteilungen mit großer Intensität.

(iii) Sind $X_1, X_2, \ldots$ unabhängige, $N(m, \sigma^2)$-verteilte Zufallsvariablen, dann gilt

$$\widetilde{S_n} \quad = \quad \frac{X_1 + X_2 + \ldots + X_n - nm}{\sqrt{n}} \quad \sim \quad N(0, \sigma^2)$$

für alle $n \in \mathbb{N}$ (und nicht nur asymptotisch!).

Ein vollständiger Beweis des zentralen Grenzwertsatzes wird in der Vorlesung Einführung in die Wahrscheinlichkeitstheorie gegeben. An dieser Stelle skizzieren wir nur die Idee eines Beweises mithilfe von charakteristischen Funktionen. Wie schon im Beweis des Satzes von De Moivre/Laplace spielt eine Taylor-Approximation bis zur zweiten Ordnung eine entscheidende Rolle.

Beweis (Skizze). Wir setzen ohne Beschränkung der Allgemeinheit voraus, dass die Zufallsvariablen X_j zentriert sind, d.h. $E[X_j] = 0$. Anderenfalls betrachten wir stattdessen die zentrierten Zufallsvariablen $\widetilde{X_j} := X_j - E[X_j]$. Gilt die Aussage des zentralen Grenzwertsatzes für diese Zufallsvariablen, dann gilt sie auch für die Zufallsvariablen X_j. Um nun die Asymptotik der Verteilungen der entsprechend standardisierten Summen $S_n/\sqrt{n}$ zu bestimmen, betrachten wir charakteristische Funktionen. Die charakteristische Funktion einer reellwertigen Zufallsvariable X ist die durch

$$\phi(t) \;:=\; E\left[e^{itX}\right] \;=\; E\left[\cos(tX)\right] + i \cdot E\left[\cos(tX)\right], \quad t \in \mathbb{R},$$

definierte komplexwertige Funktion $\phi : \mathbb{R} \to \mathbb{C}$. Hierbei ist $e^{ix} = \cos(x) + i\sin(x)$ die komplexe Exponentialfunktion, und der Erwartungswert einer komplexwertigen Zufallsvariable ist definiert, indem man separat die Erwartungswerte des Real- und Imaginärteils berechnet. Die charakteristische Funktion einer Zufallsvariable X ist die Fourier-Transformation der Verteilung von X. Der Grund, warum wir hier charakteristische Funktionen betrachten, ist, dass die Verteilung einer Zufallsvariable eindeutig durch ihre charakteristische Funktion bestimmt ist, und dass aus der punktweisen Konvergenz der charakteristischen Funktionen auch die Konvergenz der Verteilungen folgt. Diese Aussagen werden in der Vorlesung Einführung in die Wahrscheinlichkeitstheorie bewiesen, und hier ohne Beweis vorausgesetzt.

In unserem Fall kann man den Grenzwert der charakteristischen Funktionen der standardisierten Summen $S_n/\sqrt{n}$ relativ leicht berechnen. Da die Summanden X_j unabhängig und identisch verteilt sind, gilt

$$\phi_{\frac{S_n}{\sqrt{n}}}(t) \;=\; E\left[e^{itS_n/\sqrt{n}}\right] \;=\; E\left[\prod_{j=1}^{n} e^{itX_j/\sqrt{n}}\right] \;=\; \prod_{j=1}^{n} E\left[e^{itX_j/\sqrt{n}}\right] \;=\; E\left[e^{itX_1/\sqrt{n}}\right]^n .$$

Für $x \to 0$ hat die komplexe Exponentialfunktion die Taylorreihenentwicklung

$$e^{ix} \;=\; 1 + ix + \frac{1}{2}(ix)^2 + o(x^2) \;=\; 1 + ix - \frac{1}{2}x^2 + o(x^2) .$$

Hieraus ergibt sich für $t \in \mathbb{R}$ die Reihenentwicklung

$$E\left[e^{itX_1/\sqrt{n}}\right] \;=\; 1 + i \cdot E[X_1/\sqrt{n}] \cdot t - \frac{1}{2}\, E[X_1^2/n] \cdot t^2 + o\left(\frac{1}{n}\right) \;=\; 1 - \frac{\sigma^2 t^2}{2n} + o\left(\frac{1}{n}\right)$$

für die charakteristische Funktion der Zufallsvariable $X_1/\sqrt{n}$ im Grenzwert $n \to \infty$. Damit erhalten wir

$$\lim_{n\to\infty} \phi_{\frac{S_n}{\sqrt{n}}}(t) \;=\; \lim_{n\to\infty} \left(1 - \frac{\sigma^2 t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \;=\; \exp\left(-\frac{\sigma^2 t^2}{2}\right) .$$

Nachrechnen zeigt, dass die Funktion auf der rechten Seite die charakteristische Funktion der Normalverteilung $N(0, \sigma^2)$ ist. Also konvergieren die charakteristischen Funktionen der standardisierten Summen $S_n/\sqrt{n}$ in der Tat punktweise gegen die charakteristische Funktion der Normalverteilung, und somit konvergieren auch die Verteilungen der Zufallsvariablen. ■

5.2 Konfidenzintervalle

Konfidenzintervalle im Binomialmodell

Angenommen, wir wollen den Anteil p der Wähler einer Partei durch Befragung von n Wählern schätzen. Seien $X_1, \ldots, X_n$ unter P_p unabhängige und Bernoulli(p)-verteilte Zufallsvariablen, wobei $X_i = 1$ dafür steht, dass der i-te Wähler für die Partei A stimmen wird, und sei $S_n = X_1 + \cdots + X_n$ die Anzahl der Stimmen für Partei A in unserer Stichprobe. Ein naheliegender Schätzwert für p ist $\overline{X_n} := \frac{S_n}{n}$. Wie viele Wähler muss man befragen, damit der tatsächliche Stimmenanteil mit 95% Wahrscheinlichkeit um höchstens $\varepsilon = 1\%$ von diesem Schätzwert abweicht?

Definition 5.5 (Konfidenzintervall). Sei $\alpha \in (0, 1)$. Ein von den Werten $X_1, \ldots, X_n$ abhängendes (und damit zufälliges) Intervall $I(X_1, \ldots, X_n) \subseteq \mathbb{R}$ heißt *Konfidenzintervall zum Konfidenzniveau* $1 - \alpha$ *(bzw. zum Irrtumsniveau* α*) für den unbekannten Parameter* p, falls

$$P_p[p \notin I(X_1, \ldots, X_n)] \quad \leq \quad \alpha$$

für *alle* möglichen Parameterwerte $p \in [0, 1]$ gilt.

Wir betrachten nun zunächst einige einfache Methoden, mit denen wir in unserem Modell Konfidenzintervalle herleiten können.

Abschätzung mithilfe der Čebyšev-Ungleichung:

$$P_p\left[\left|\frac{S_n}{n}-p\right|\geq\varepsilon\right] \quad\leq\quad \frac{1}{\varepsilon^2}\cdot\mathrm{Var}\left(\frac{S_n}{n}\right) \quad=\quad \frac{p(1-p)}{n\varepsilon^2} \quad\leq\quad \frac{1}{4n\varepsilon^2} \quad\overset{!}{\leq}\quad \alpha \qquad \forall\, p\in[0,1]$$

Dies ist erfüllt für $n \geq \frac{1}{4\varepsilon^2\alpha}$, also im Beispiel für $n \geq 50.000$. In diesem Fall ist also das Intervall $\left(\overline{X_n}-\varepsilon, \overline{X_n}+\varepsilon\right)$ ein Konfidenzintervall für den unbekannten Parameter p zum Konfidenzniveau $1-\alpha$.

Abschätzung über die Bernstein-Ungleichung:

$$P_p\left[\left|\frac{S_n}{n}-p\right|\geq\varepsilon\right] \quad\leq\quad 2\cdot e^{-2\varepsilon^2 n} \quad\overset{!}{\leq}\quad \alpha \qquad \forall\, p\in[0,1].$$

Dies ist erfüllt für $n \geq \frac{1}{2\varepsilon^2}\log(\frac{2}{\alpha})$, also im Beispiel für $n \geq 18445$.

Die Abschätzung über die Bernstein-Ungleichung ist genauer - sie zeigt, dass bereits weniger als 20.000 Stichproben genügen, um die Niveaubedingung zu erfüllen. Können wir mit noch weniger Stichproben auskommen? Dazu berechnen wir die Wahrscheinlichkeit, dass der Parameter p im Konfidenzintervall liegt, näherungsweise mithilfe des zentralen Grenzwertsatzes:

Approximative Berechnung mithilfe der Normalapproximation:

$$\begin{aligned}
P_p\left[\left|\frac{S_n}{n}-p\right|\leq\varepsilon\right] \quad &= \quad P_p\left[\left|\frac{S_n-np}{\sqrt{np(1-p)}}\right|\leq\frac{n\varepsilon}{\sqrt{np(1-p)}}\right] \\
&\approx \quad N(0,1)\left(-\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\,,\,\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right) \\
&= \quad 2\left(\Phi\left(\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right)-\frac{1}{2}\right) \\
&\overset{p(1-p)\leq\frac{1}{4}}{\geq} \quad 2\Phi(2\sqrt{n}\varepsilon)-1 \quad\geq\quad 1-\alpha \qquad \forall\, p\in[0,1],
\end{aligned}$$

falls

$$n \quad\geq\quad \left(\frac{1}{2\varepsilon}\cdot\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2.$$

Im Beispiel gilt $\Phi^{-1}(1-\alpha/2) \approx 1.96$, und die Bedingung ist für $n \geq 9604$ erfüllt. Also sollten bereits ca. 10.000 Stichproben ausreichen! *Exakte* (also ohne Verwendung einer Näherung hergeleitete) Konfidenzintervalle sind in vielen Fällen zu konservativ. In Anwendungen werden daher meistens *approximative* Konfidenzintervalle angegeben, die mithilfe einer Normalapproximation hergeleitet wurden. Dabei ist aber folgendes zu beachten:

Warnung: Mithilfe der Normalapproximation hergeleitete approximative Konfidenzintervalle erfüllen die Niveaubedingung im Allgemeinen nicht (bzw. nur näherungsweise). Da die Qualität der Normalapproximation für $p \to 0$ bzw. $p \to 1$ degeneriert, ist die Niveaubedingung im Allgemeinen selbst für $n \to \infty$ nicht erfüllt. Beispielsweise beträgt das Niveau von approximativen 99% Konfidenzintervallen asymptotisch tatsächlich nur 96.8%!

Exakte Konfidenzintervalle im Binomialmodell: Im oben betrachteten einfachen Modell ist es sogar möglich, nahezu optimale exakte Konfidenzintervalle aus der zugrundeliegenden Verteilungsfunktion zu erhalten. Dazu bemerken wir, dass für Funktionen $a, b : \{0, 1, \dots, n\} \to [0, 1]$ mit $a \le b$ gilt:

$$P_p\left[p \notin (a(S_n), b(S_n))\right] \;\le\; P_p\left[p \ge b(S_n)\right] + P_p\left[p \le a(S_n)\right] \;\le\; 2\alpha\,, \tag{5.11}$$

falls die Wahrscheinlichkeiten auf der rechten Seite beide kleiner oder gleich α sind. Sei nun

$$F_{n,p}(k) \;=\; \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

die Verteilungsfunktion von S_n unter P_p, ausgewertet an einer festen Stelle $k \in \{0, 1, \dots, n-1\}$. Durch Ableiten kann man zeigen, dass die Abbildung $p \mapsto F_{n,p}(k)$ stetig und streng monoton fallend mit $F_{n,0}(k) = 1$ und $F_{n,1}(k) = 0$, und daher eine Bijektion von $[0, 1]$ nach $[0, 1]$ ist. Genauer erhält man mithilfe des Hauptsatzes der Differential- und Integralrechnung die Darstellung

$$F_{n,p}(k) \;=\; \int_p^1 n \binom{n-1}{k} u^k\,(1-u)^{n-1-k}\,du \quad \text{für alle } p \in [0, 1]. \tag{5.12}$$

Sei nun $b(k)$ für $k < n$ die eindeutige Lösung p der Gleichung $F_{n,p}(k) = \alpha$, und sei $b(n) := 1$. Dann gilt wegen der Monotonie für alle $p \in (0, 1)$ die Äquivalenz

$$F_{n,p}(k) \le \alpha \quad \Longleftrightarrow \quad p \ge b(k)\,,$$

und damit

$$P_p\left[p \ge b(S_n)\right] \;=\; P_p\left[F_{n,p}(S_n) \le \alpha\right] \;\le\; \alpha\,, \tag{5.13}$$

siehe Lemma 5.7 für den Beweis der letzten Abschätzung. Auf ähnliche Weise zeigt man

$$P_p\left[p \le a(S_n)\right] \;=\; P_p\left[F_{n,p}(S_n - 1) \ge 1 - \alpha\right] \;\le\; \alpha\,, \tag{5.14}$$

wobei $a(k)$ für $k \in \{1, \dots, n\}$ die eindeutige Lösung p der Gleichung $F_{n,p}(k-1) = F_{n,p}(k-) = 1 - \alpha$ ist, und wir $a(0) := 0$ setzen. Aus (5.14) und (5.13) folgt, dass die Bedingung (5.11) für die oben definierten Funktionen a und b erfüllt ist. Damit haben wir die folgende Aussage bewiesen.

Satz 5.6 (Exakte Konfidenzintervalle im Binomialmodell). Sei $\alpha \in (0, 1/2)$, und seien $a(k)$ für $k \in \{1, \dots, n\}$ und $b(k)$ für $k \in \{0, 1, \dots, n-1\}$ die eindeutigen Lösungen p der Gleichungen

$$F_{n,p}(k-1) \;=\; 1 - \alpha\,, \quad \text{bzw.} \quad F_{n,p}(k) = \alpha.$$

Zudem sei $a(0) := 0$ und $b(n) := 1$. Dann ist das Intervall $(a(S_n), b(S_n))$ ein Konfidenzintervall für p zum Irrtumsniveau 2α bzw. zum Konfidenzniveau $1 - 2\alpha$.

Im oben dargestellten Beweis haben wir das folgende Lemma verwendet, dessen Beweis wir jetzt nachtragen.

Lemma 5.7. *Sei X eine reellwertige Zufallsvariable mit Verteilungsfunktion F. Dann gilt für alle $\alpha \in (0, 1)$:*

$$P[F(X) \le \alpha] \;\le\; \alpha\,, \qquad \textit{und} \qquad P[F(X-) \ge 1 - \alpha] \;\le\; \alpha\,. \tag{5.15}$$

Beweis. Sei $G(u) = \inf\{c \in \mathbb{R} : F(c) \geq u\}$ die entsprechende untere Quantilfunktion. Nach Satz 4.25 hat X dieselbe Verteilung wie $G(U)$, wobei U eine auf $(0,1)$ gleichverteilte Zufallsvariable ist. Daher gilt

$$\begin{aligned} P[F(X) \leq \alpha] &= P[F(G(U)) \leq \alpha] \leq P[U \leq \alpha] = \alpha, \qquad \text{und} \\ P[F(X-) \geq 1-\alpha] &= P[F(G(U)-) \geq 1-\alpha] \leq P[U \geq 1-\alpha] = \alpha. \end{aligned}$$

Hierbei haben wir für die hinteren Abschätzungen benutzt, dass $F(G(U)-) \leq U \leq F(G(U))$ gilt. ■

Ein Problem bei dem oben beschriebenen Vorgehen zur Konstruktion eines exakten Konfidenzintervalls ist, dass die Intervallgrenzen a und b nicht explizit berechnet werden können. Stattdessen werden die entsprechenden Gleichungen mithilfe von Statistiksoftware numerisch gelöst.

Konfidenzintervalle im Gauß-Modell

Angenommen, wir beobachten reellwertige Messwerte (Stichproben, Daten), die von einer unbekannten Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}$ stammen. Ziel der Statistik ist es, Rückschlüsse auf die zugrundeliegende Verteilung aus den Daten zu erhalten. Im Gauß-Modell nimmt man an, dass die Daten unabhängige Stichproben von einer Normalverteilung mit unbekanntem Mittelwert und/oder Varianz sind:

$$\mu = N(m, v), \quad m, v \text{ unbekannt.}$$

Eine partielle Rechtfertigung für die Normalverteilungsannahme liefert der zentrale Grenzwertsatz. Letztendlich muss man aber in jedem Fall überprüfen, ob eine solche Annahme gerechtfertigt ist. Ein erstes Ziel ist es nun, den Wert von m auf der Basis von n unabhängigen Stichproben $X_1(\omega) = x_1, \ldots, X_n(\omega) = x_n$ zu schätzen, und den Schätzfehler zu quantifizieren.

Problemstellung: Schätzung des Erwartungswerts

- Schätze m auf der Basis von n unabhängigen Stichproben $X_1(\omega), ..., X_n(\omega)$ von μ.
- Herleitung von Konfidenzintervallen.

Im mathematischen Modell interpretieren wir die Beobachtungswerte als Realisierungen von unabhängigen Zufallsvariablen $X_1, \ldots, X_n$. Da wir die tatsächliche Verteilung nicht kennen, untersuchen wir alle in Betracht gezogenen Verteilungen simultan:

$$X_1, \ldots, X_n \sim N(m, v) \quad \text{unabhängig unter } P_{m,v}. \tag{5.16}$$

Ein naheliegender Schätzer für m ist der *empirische Mittelwert*

$$\overline{X}_n(\omega) := \frac{X_1(\omega) + ... + X_n(\omega)}{n}.$$

Wir haben oben bereits gezeigt, dass dieser Schätzer *erwartungstreu (unbiassed)* und *konsistent* ist, d.h. für alle m, v gilt:

$$E_{m,v}[\overline{X}_n] = m$$

und

$$\overline{X}_n \to m \quad P_{m,v}\text{-stochastisch für } n \to \infty.$$

Wie wir den Schätzfehler quantifizieren hängt davon ab, ob wir die Varianz kennen.

Schätzung von m bei bekannter Varianz v.
Um den Schätzfehler zu kontrollieren, berechnen wir die Verteilung von $\overline{X}_n$:

$$\begin{aligned} X_i \sim N(m,v) \text{ unabhängig} &\Rightarrow X_1 + \ldots + X_n \sim N(nm, nv) \\ &\Rightarrow \overline{X}_n \sim N(m, v/n) \\ &\Rightarrow \frac{\overline{X}_n - m}{\sqrt{v/n}} \sim N(0,1) \end{aligned}$$

Bezeichnet Φ die Verteilungsfunktion der Standardnormalverteilung, dann erhalten wir

$$P_{m,v}\left[|\overline{X}_n - m| < q\sqrt{\frac{v}{n}}\right] = N(0,1)(-q,q) = 2\left(\Phi(q) - \frac{1}{2}\right) \quad \text{für alle } m \in \mathbb{R}.$$

Satz 5.8 (Konfidenzintervalle bei bekannter Varianz). Im Gaußmodell (5.16) mit bekannter Varianz v ist das zufällige Intervall

$$\left(\overline{X}_n - \Phi^{-1}(\alpha)\sqrt{\frac{v}{n}},\ \overline{X}_n + \Phi^{-1}(\alpha)\sqrt{\frac{v}{n}}\right)$$

ein $(2\alpha - 1)\cdot 100\%$ **Konfidenzintervall** für m, d.h.

$$P_{m,v}[m \in \text{Intervall}] \geq 2\alpha - 1 \qquad \text{für alle } m \in \mathbb{R}.$$

Man beachte, dass die Länge des Konfidenzintervalls in diesem Fall nicht von den beobachteten Stichproben abhängt!

Schätzung von m bei unbekannter Varianz v. In Anwendungen ist meistens die Varianz unbekannt. In diesem Fall können wir das Intervall oben nicht verwenden, da es von der unbekannten Varianz v abhängt. Stattdessen schätzen wir m und v simultan, und konstruieren ein Konfidenzintervall für m mithilfe beider Schätzwerte. Erwartungstreue Schätzer für m und v sind

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{und} \quad V_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2 .$$

Bei bekannter Varianz v hatten wir Konfidenzintervalle für m vom Typ $\overline{X}_n \pm q\cdot\sqrt{v/n}$ erhalten, wobei q ein geeignetes Quantil der Standardnormalverteilung ist. Daher liegt es nahe, zu versuchen, bei unbekannter Varianz Konfidenzintervalle vom Typ $\overline{X}_n \pm q\cdot\sqrt{V_n/n}$ herzuleiten. Es gilt

$$P_{m,v}\left[|\overline{X}_n - m| \geq q\sqrt{V_n/n}\right] = P_{m,v}[|T_{n-1}| \geq q] \qquad \text{mit}$$

$$T_{n-1} := \frac{\sqrt{n}\cdot(\overline{X}_n - m)}{\sqrt{V_n}} .$$

Die Zufallsvariable T_{n-1} heißt **Studentsche t-Statistik mit $n-1$ Freiheitsgraden**.[1] Es stellt sich heraus, dass die Verteilung von T_{n-1} unter $P_{m,v}$ nicht von den unbekannten Parametern m und v abhängt. Daher können wir aus Quantilen der Studentschen t-Statistik Konfidenzintervalle für das Gauß-Modell herleiten.

[1] In der Statistik bezeichnet man eine messbare Funktion der Beobachtungsdaten als Statistik - ein (Punkt-) Schätzer ist eine Statistik, die zum Schätzen eines unbekannten Parameters verwendet wird, ein Konfidenzintervall nennt man auch Intervallschätzer.

Satz 5.9 (Student). Die Verteilung von T_{n-1} ist absolutstetig mit Dichte

$$f_{T_{n-1}}(t) \quad = \quad B\left(\frac{1}{2}, \frac{n-1}{2}\right)^{-1} \cdot (n-1)^{-1/2} \cdot \left(1 + \frac{t^2}{n-1}\right)^{-n/2} \qquad (t \in \mathbb{R}). \tag{5.17}$$

Hierbei ist $B(a,b) = \int_0^1 s^{a-1}(1-s)^{b-1}\,ds$ die *Eulersche Beta-Funktion*, die als Normierungsfaktor auftritt. Insbesondere ist das zufällige Intervall $\left(\overline{X}_n - q \cdot \sqrt{V_n/n},\, \overline{X}_n + q \cdot \sqrt{V_n/n}\right)$ ein $100 \cdot (1-2\alpha)\%$ Konfidenzintervall für m, falls

$$q = F^{-1}_{T_{n-1}}(1-\alpha)$$

ein $(1-\alpha)$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden ist.

Der Beweis basiert auf dem Transformationssatz für Wahrscheinlichkeitsverteilungen auf $\mathbb{R}^n$, siehe zum Beispiel Abschnitt 5.4 des Skripts zur Vorlesung Einführung in die Wahrscheinlichkeitstheorie.

Definition 5.10 (Studentsche t-Verteilung). Die Wahrscheinlichkeitsverteilung auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Dichtefunktion (5.17) nennt man »**Studentsche t-Verteilung mit $n-1$ Freiheitsgraden**«.

Bemerkung (Nichtparametrische und verteilungsunabhängige Konfidenzintervalle). In Anwendungen ist es oft unklar, ob eine Normalverteilungsannahme an die Beobachtungswerte gerechtfertigt ist. Zudem können einzelne größere Ausreißer in den Daten (z.B. aufgrund von Messfehlern) das Stichprobenmittel relativ stark beeinflussen. Der Stichprobenmedian ist dagegen in den meisten Fällen ein deutlich stabilerer Schätzwert für den Median der zugrundeliegenden Verteilung. Im folgenden Abschnitt zeigen wir allgemein, wie wir Konfidenzintervalle für die Quantile einer Verteilung erhalten können, welche weniger stark durch Ausreißer beeinflusst werden. Zudem gelten diese Konfidenzintervalle simultan für alle stetigen Verteilungen. Ist man sich daher nicht sicher, ob eine Normalverteilungsannahme aufgrund der Daten gerechtfertigt ist, empfiehlt es sich, auf diese stabileren Ordnungsintervalle zurückzugreifen.

Konfidenzintervalle für Quantile

Sei $(x_1, \ldots, x_n)$ eine n-elementige Stichprobe von einer unbekannten Wahrscheinlichkeitsverteilung μ auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Wir nehmen an, dass $x_1, \ldots, x_n$ Realisierungen von unabhängigen Zufallsvariablen mit Verteilung μ sind:

Annahme: $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ unabhängig unter P_μ mit Verteilung μ.

Wir wollen nun die Quantile (z.B. den Median) der zugrundeliegenden Verteilung auf Basis der Stichprobe schätzen. Eine Funktion $T(X_1, \ldots, X_n)$, $T : \mathbb{R}^n \to \mathbb{R}$ messbar, nennt man in diesem Zusammenhang auch eine *Statistik* der Stichprobe $(X_1, \ldots, X_n)$. Eine Statistik, deren Wert als Schätzwert für eine Kenngröße $q(\mu)$ der unbekannten Verteilung verwendet wird, nennt man auch einen *(Punkt-) Schätzer* für q. Nahe liegende Schätzer für Quantile von μ sind die entsprechenden Stichprobenquantile. Unser Ziel ist es nun, *Konfidenzintervalle* für die Quantile anzugeben, d.h. von den Werten $X_1, \ldots, X_n$ abhängende Intervalle, in denen die Quantile *unabhängig von der tatsächlichen Verteilung* mit hoher Wahrscheinlichkeit enthalten sind. Seien dazu

$$X_{(1)} \quad \leq \quad X_{(2)} \quad \leq \quad \ldots \quad \leq \quad X_{(n)}$$

die der Größe nach geordneten Werte $X_1, \ldots, X_n$; diese nennt man auch *Ordnungsstatistiken* der Stichprobe. Die Verteilung der Ordnungsstatistiken können wir explizit berechnen:

Satz 5.11 (Verteilung der Ordnungsstatistiken). Sei $u \in (0,1)$ und $1 \le k < l \le n$.

(i) Ist F die Verteilungsfunktion von μ, dann hat $X_{(k)}$ die Verteilungsfunktion

$$F_{(k)}(c) \;=\; \mathrm{Bin}(n, F(c))[\{k, k+1, \ldots, n\}] \;=\; \sum_{j=k}^{n} \binom{n}{j} F(c)^j \, (1 - F(c))^{n-j}. \tag{5.18}$$

(ii) Ist F stetig und q ein u-Quantil von μ, dann gilt $F(q) = u$ und

$$P_\mu \left[q \in (X_{(k)}, X_{(l)}) \right] \;=\; \mathrm{Bin}(n, u)[\{k, k+1, \ldots, l-1\}].$$

Beweis. Die Ereignisse $\{X_i \le c\}$ $(1 \le i \le n)$ sind unabhängig mit Wahrscheinlichkeit $F(c)$. Also gilt

$$\begin{aligned} F_{(k)}(c) \;=\; P_\mu \left[X_{(k)} \le c \right] \;&=\; P_\mu \left[X_i \le c \text{ für mindestens } k \text{ verschiedene } i \in \{1, \ldots, n\} \right] \\ &=\; \mathrm{Bin}(n, F(c))[\{k, k+1, \ldots, n\}]. \end{aligned}$$

Ist F stetig und q ein u-Quantil von F, dann gilt $F(q) = u$ nach Lemma 4.24. Zudem ist $F_{(k)}$ nach (i) ebenfalls stetig. Damit erhalten wir

$$\begin{aligned} P_\mu[X_{(k)} < q < X(l)] \;&=\; F_{(k)}(q-) - F_{(l)}(q) \;=\; F_{(k)}(q) - F_{(l)}(q) \\ &=\; \mathrm{Bin}(n, u)[\{k, k+1, \ldots, n\}] - \mathrm{Bin}(n, u)[\{l, l+1, \ldots, n\}] \\ &=\; \mathrm{Bin}(n, u)[\{k, k+1, \ldots, l-1\}]. \end{aligned}$$

■

Nach Satz 5.11 ist die Wahrscheinlichkeit, dass der Wert eines u-Quantils der zugrundeliegenden Verteilung μ zwischen $X_{(k)}$ und $X_{(l)}$ liegt, *für alle stetigen Verteilungen gleich!* Damit folgt unmittelbar:

Korollar 5.12 (Ordnungsintervalle). Sei $u \in (0,1)$ und $1 \le k < l \le n$. Dann ist das zufällige Intervall $(X_{(k)}, X_{(l)})$ ein Konfidenzintervall für das u-Quantil der zugrundeliegenden Verteilung μ zum Konfidenzniveau

$$\beta \;:=\; \mathrm{Bin}(n, u)[\{k, k+1, \ldots, l-1\}],$$

d.h. für jede Wahrscheinlichkeitsverteilung μ auf $\mathbb{R}$ mit stetiger Verteilungsfunktion, und für jedes u-Quantil q von μ gilt:

$$P_\mu \left[q \in (X_{(k)}, X_{(l)}) \right] \;\ge\; \beta.$$

Für große n kann man die Quantile der Binomialverteilung näherungsweise mithilfe der Normalapproximation berechnen, und erhält daraus entsprechende Konfidenzintervalle für die Quantile von Verteilungen auf $\mathbb{R}$. Bemerkenswert ist, dass diese Konfidenzintervalle nicht nur für Verteilungen aus einer bestimmten parametrischen Familie (z.B. der Familie der Normalverteilungen) gelten, sondern für *alle* Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$ mit stetiger Verteilungsfunktion (*nichtparametrisches statistisches Modell*).

Beispiel (Konfidenzintervalle für den Median). Die Binomialverteilung $\mathrm{Bin}(n, 1/2)$ hat den Mittelwert $m = n/2$ und die Standardabweichung $\sigma = \sqrt{n}/2$. Nach dem Satz von De Moivre/Laplace ist für große n ca. 95 % der Masse in der Menge $\{\lfloor m - 2\sigma \rfloor, \ldots, \lceil m + 2\sigma \rceil\}$ enthalten. Also ist das Intervall $\left(X_{(\lfloor n/2 - \sqrt{n} \rfloor)}, X_{(\lceil n/2 + \sqrt{n} \rceil)} \right)$ ein approximatives 95 % Konfidenzintervall für den Median einer beliebigen Verteilung mit stetiger Verteilungsfunktion. Beispielsweise können wir bei Zufallsstichproben der Größe 100 mit hoher Konfidenz erwarten, dass der Median zwischen dem 40. und 61. Wert liegt.

5.3 Hypothesentests

In Anwendungen werden statistische Aussagen häufig nicht über Konfidenzintervalle, sondern als Hypothesentest formuliert. Zum Teil handelt es sich dabei nur um eine durch praktische Erwägungen motivierte Umformulierung derselben mathematischen Resultate.

Hypothesentest im Binomialmodell

Wir beginnen mit einem typischen Beispiel.

Beispiel. Die folgende Aufgabe stammt aus dem Schulbuch “Mathematik Neue Wege":

Christine trinkt gern stilles Mineralwasser. Leon ärgert sie und sagt: „Da kannst du gleich Leitungswasser trinken”. Christine behauptet, sie könne recht zuverlässig Leitungswasser von stillem Mineralwasser unterscheiden. Leon schlägt vor: „Wir können gleich einen Test machen.” Er füllt fünf Gläser mit Leitungswasser und fünf Gläser mit stillem Mineralwasser. Er stellt die Gläser, ohne dass Christine es sehen kann, in beliebiger Reihenfolge auf. Christine nimmt jeweils eine Geschmacksprobe. Bei acht Gläsern ordnet sie den Inhalt richtig zu. Spricht das Testergebnis nun für Christines Behauptung oder hat sie nur Glück gehabt?

Das Problem lässt sich als Hypothesentest formulieren. Sei X eine Zufallsvariable, die die Anzahl der erfolgreichen Zuordnungen bei n Versuchen beschreibt. Im Beispiel ist $n = 10$. Wenn Christine stilles Mineralwasser gar nicht von Leitungswasser unterscheiden kann, dann sollte X binomialverteilt sein mit Parametern n und $p = 1/2$.

$$\textit{Nullhypothese } H_0\colon \qquad »X \sim \mathrm{Bin}(n, 1/2)«$$

Wenn Christine dagegen mit Wahrscheinlichkeit $p > 1/2$ stilles Mineralwasser von Leitungswasser unterscheiden kann, dann sollte gelten:

$$\textit{Alternative } H_1\colon \qquad »X \sim \mathrm{Bin}(n, p) \quad \text{für ein } p \in (1/2, 1]«$$

In der Interpretation statistischer Modelle sind die Nullhypothese und die Alternative nicht gleichwertig. Die Nullhypothese beschreibt meist den „Normalfall”, die Alternative das Vorhandensein eines bestimmten „Effekts”.

Wir beobachten nun eine einzige Realisierung $x = X(\omega)$ der Zufallsvariable X. Was können wir basierend auf dieser Stichprobe aussagen? Könnte ein Erfolg in acht Fällen eine typische zufällige Fluktuation sein, oder sollten wir davon ausgehen, dass Christine tatsächlich stilles Mineralwasser von Leitungswasser unterscheiden kann? In einem Hypothesentest verwerfen wir die Nullhypothese und akzeptieren die Alternative, falls der realisierte Wert x der Zufallsvariable X oberhalb eines vorgegebenen *Schwellenwertes* $c \in \{0, 1, \dots, n\}$ liegt. Andernfalls verwerfen wir die Nullhypothese nicht (allerdings können wir sie auch nicht beweisen). Wie groß sollten wir c wählen? Bei unserer Entscheidung können zwei Arten von Fehlern auftreten:

Fehler 1. Art: *H_0 wird verworfen, obwohl wahr.* Die Wahrscheinlichkeit dafür beträgt

$$P_{1/2}[X \geq c] \;=\; \sum_{k=c}^{n} \binom{n}{k} 2^{-n}.$$

Fehler 2. Art: *H_0 wird nicht verworfen, obwohl falsch.* Die Wahrscheinlichkeit beträgt

$$P_p[X < c] \;=\; 1 - \sum_{k=c}^{n} \binom{n}{k} p^k (1-p)^{n-k},$$

wobei $p > 1/2$ die tatsächliche Erfolgswahrscheinlichkeit ist.

Der Fehler 1. Art bedeutet, dass wir einen Effekt behaupten, obwohl er nicht vorliegt. Diese Art von Fehler möchte man mit hoher Sicherheit vermeiden. Deshalb fordert man, dass die Wahrscheinlichkeit

für den Fehler 1. Art auf jeden Fall unterhalb eines vorgegebenen *Signifikanzniveaus* α liegt, z.B. $\alpha = 5\%$. In unserem Beispiel gilt für den Schwellenwert $c = 8$:

$$P_{1/2}[X \geq c] \;=\; \binom{10}{8}2^{-10} + \binom{10}{9}2^{-10} + \binom{10}{10}2^{-10} \;=\; \frac{56}{1024} \;\approx\; 5,5\%\,.$$

Verwerfen wir die Nullhypothese also für $X \geq 8$, dann wird unser angestrebtes Signifikanzniveau von 5% nicht unterschritten. Für den Schwellenwert $c = 9$ gilt dagegen

$$P_{1/2}[X \geq c] \;=\; \binom{10}{9}2^{-10} + \binom{10}{10}2^{-10} \;=\; \frac{11}{1024} \;\approx\; 1,1\%\,.$$

In diesem Fall wird das Signifikanzniveau deutlich unterschritten. Um zu vermeiden, dass die Wahrscheinlichkeit für den Fehler 1. Art größer als 5% ist, müssen wir also $c = 9$ wählen. In diesem Fall können wir die Nullhypothese aufgrund des beobachteten Wertes $x = 8$ aber nicht verwerfen. Das bedeutet aber nicht, dass wir nun davon ausgehen sollten, dass die Nullhypothese wahr ist, also Christine stilles Mineralwasser nicht von Leitungswasser unterscheiden kann! Wir können lediglich schließen, dass wir aufgrund der Beobachtung nicht in der Lage sind, die Nullhypothese zum Niveau 5% signifikant zu verwerfen. Wir können also aufgrund der wenigen vorliegenden Daten zu keiner signifikanten Entscheidung kommen, und sollten den Test eventuell mit einer größeren Anzahl n von Versuchen wiederholen.

Das Beispiel zeigt, dass wir mit einem Hypothesentest die Nullhypothese niemals bestätigen, sondern nur signifikant verwerfen können. Die kleinstmögliche Wahl des Schwellenwertes c ist durch das Signifikanzniveau bestimmt. Andererseits sollte auch die Wahrscheinlichkeit für den Fehler 2. Art möglichst klein sein, und diese wächst mit c. Daher sollten wir wenn möglich den kleinsten Schwellenwert wählen, für den das Signifikanzniveau unterschritten wird. Im Beispiel war dies $c = 9$. Man beachte, dass die Wahrscheinlichkeit für den Fehler 2. Art trotzdem sehr groß sein kann. Dies ist unvermeidbar, da p beliebig nahe bei $1/2$ liegen kann.

Allgemeiner Rahmen für Hypothesentests

Angenommen, wir haben n unabhängige reellwertige Stichproben $X_1, ..., X_n$ von einer unbekannten Verteilung vorliegen und wir gehen davon aus, dass die zugrundeliegende Verteilung aus einer Familie μ_θ ($\theta \in \Theta$) von Wahrscheinlichkeitsverteilungen kommt, z.B. der Familie aller Normalverteilungen $\mu_\theta = N(m, v)$, $\theta = (m, v) \in \mathbb{R} \times \mathbb{R}_+$. Die gemeinsame Verteilung von $X_1, \ldots, X_n$ ist dann das Produktmaß $\mu_\theta^n = \bigotimes_{i=1}^n \mu_\theta$. Seien nun Θ_0 und Θ_1 disjunkte Teilmengen des Parameterbereichs. Wir wollen entscheiden zwischen der

$$\textit{Nullhypothese } H_0\text{:} \qquad \text{»}\theta \in \Theta_0\text{«}$$

und der

$$\textit{Alternative } H_1\text{:} \qquad \text{»}\theta \in \Theta_1\text{«}$$

Ein **Hypothesentest** für ein solches Problem ist bestimmt durch eine messbare Teilmenge $C \subseteq \mathbb{R}^n$ (den **Verwerfungsbereich**) mit zugehöriger Entscheidungsregel:

$$\text{Verwerfe } H_0 \quad \Longleftrightarrow \quad (X_1, ..., X_n) \in C.$$

Wenn wir die Nullhypothese verwerfen, dann entscheiden wir uns für die Alternative. Wenn die Nullhypothese hingegen nicht verworfen wird, dann bedeutet das nicht, dass wir davon ausgehen, dass die Nullhypothese wahr ist. Wir können lediglich sagen, dass die gegebenen Daten nicht ausreichen, um die Nullhypothese zu einem bestimmten Signifikanzniveau zu verwerfen.

Beispiel (t-Test). Seien $X_1, X_2, \ldots, X_n$ unabhängige Stichproben von einer Normalverteilung mit unbekanntem Parameter $\theta = (m, v) \in \Theta := \mathbb{R} \times \mathbb{R}_+$. Wir wollen testen, ob der Mittelwert der Verteilung einen bestimmten Wert m_0 hat:

$$\textit{Nullhypothese } H_0\colon \quad »m = m_0« \,, \qquad \Theta_0 = \{m_0\} \times \mathbb{R}_+ \,.$$

$$\textit{Alternative } H_1\colon \quad »m \neq m_0« \,, \qquad \Theta_1 = \Theta \setminus \Theta_0.$$

Ein solches Problem tritt beispielsweise in der Qualitätskontrolle auf, wenn man überprüfen möchte, ob ein Sollwert m_0 angenommen wird. Eine andere Anwendung ist der Vergleich zweier Verfahren, wobei X_i die Differenz der mit beiden Verfahren erhaltenen Messwerte ist. Die Nullhypothese mit $m_0 = 0$ besagt hier, dass kein signifikanter Unterschied zwischen den Verfahren besteht. Im *t–Test* für obiges Testproblem wird die Nullhypothese verworfen, falls der Betrag der *Studentschen t-Statistik* oberhalb eines angemessen zu wählenden Schwellenwerts c liegt, also falls

$$|T_{n-1}| \;=\; \left| \frac{\sqrt{n} \cdot (\overline{X}_n - m_0)}{\sqrt{V_n}} \right| \;>\; c \,.$$

Seien nun allgemein $X_1, X_2, \ldots$ unter P_θ unabhängige Zufallsvariablen mit Verteilung μ_θ. Bei einem Hypothesentest können zwei Arten von Fehlern auftreten:

Fehler 1. Art: *H_0 wird verworfen, obwohl wahr.* Die Wahrscheinlichkeit dafür beträgt:

$$P_\theta[(X_1, \ldots, X_n) \in C] \;=\; \mu_\theta^n[C] \,, \qquad \theta \in \Theta_0.$$

Fehler 2. Art: *H_0 wird nicht verworfen, obwohl falsch.* Die Wahrscheinlichkeit beträgt:

$$P_\theta[(X_1, \ldots, X_n) \notin C] \;=\; \mu_\theta^n[C^C] \,, \qquad \theta \in \Theta_1.$$

Obwohl das allgemeine Testproblem im Prinzip symmetrisch in H_0 und H_1 ist, interpretiert man beide Fehler unterschiedlich. Die Nullhypothese beschreibt in der Regel den Normalfall, die Alternative eine Abweichung oder einen zu beobachtenden Effekt. Da ein Test Kritiker überzeugen soll, sollte die Wahrscheinlichkeit für den Fehler 1. Art (Effekt prognostiziert, obgleich nicht vorhanden) unterhalb einer vorgegebenen (kleinen) Schranke α liegen. Die Wahrscheinlichkeit, dass kein Fehler 2. Art auftritt, sollte unter dieser Voraussetzung möglichst groß sein.

Definition 5.13 (Gütefunktion und Niveau eines Hypothesentests). Die Funktion

$$G(\theta) \;=\; P_\theta[(X_1, \ldots, X_n) \in C] \;=\; \mu_\theta^n[C]$$

heißt *Gütefunktion* des Tests. Der Test hat das *Signifikanzniveau* α, falls

$$G(\theta) \;\leq\; \alpha \quad \text{ für alle } \theta \in \Theta_0 \text{ gilt.}$$

Hypothesentest im Gauß-Modell

Aus Satz 5.9 und der Symmetrie der Studentschen t-Verteilung folgt unmittelbar:

Korollar 5.14. Der Studentsche t-Test hat Niveau α falls c ein $(1-\frac{\alpha}{2})$-Quantil der Studentschen t-Verteilung mit $n-1$ Freiheitsgraden ist.

Allgemeiner gilt:

Satz 5.15 (Korrespondenz Konfidenzintervalle ↔ Hypothesentests). Für einen reellwertigen Parameter $\gamma = c(\theta)$, ein Irrtumsniveau $\alpha \in (0,1)$, und messbare Abbildungen (Statistiken) $\hat{\gamma}, \varepsilon : \mathbb{R}^n \to \mathbb{R}$ sind äquivalent:

(i) Das Intervall
$$[\hat{\gamma}(X_1,\dots,X_n) - \varepsilon(X_1,\dots,X_n),\ \hat{\gamma}(X_1,\dots,X_n) + \varepsilon(X_1,\dots,X_n)]$$
ist ein Konfidenzintervall für γ zum Konfidenzniveau $1-\alpha$ bzw. zum Irrtumsniveau α.

(ii) Für jedes $\gamma_0 \in \mathbb{R}$ ist
$$C = \{(x_1,...,x_n) : |\hat{\gamma}(x_1,\dots,x_n) - \gamma_0| > \varepsilon(x_1,\dots,x_n)\}$$
der Verwerfungsbereich eines Tests der Nullhypothese $\gamma = \gamma_0$ zum Signifikanzniveau α.

Beweis. Das Intervall ist genau dann ein Konfidenzintervall für γ zum Irrtumsniveau α, wenn
$$P_\theta\ [|\hat{\gamma}(X_1,\dots,X_n) - c(\theta)| > \varepsilon(X_1,...,X_n)] \ \le\ \alpha \qquad \forall\, \theta \in \Theta$$
gilt, also wenn der entsprechende Test der Nullhypothese $c(\theta) = \gamma_0$ für jedes γ_0 Niveau α hat. ■

Überprüfen von Verteilungsannahmen

In vielen Fällen wollen wir überprüfen, ob die vorliegenden Stichproben von einer bestimmten Verteilung kommen, zum Beispiel der Gleichverteilung oder einer Normalverteilung. Dafür gibt es verschiedene visuelle Verfahren und Hypothesentests.

Diskrete Verteilungen: Histogramme und Anpassungstest

Wir betrachten zunächst den Fall einer Wahrscheinlichkeitsverteilung μ mit Massenfunktion $p(k) = \mu[\{k\}]$ auf der endlichen Menge $S = \{1,2,\dots,l\}$, $l \in \mathbb{N}$. Seien $X_1, X_2, \dots : \Omega \to S$ unabhängige identisch verteilte Zufallsvariablen mit einer unbekannten Verteilung. Wir wollen auf der Basis von Stichproben $x_1 = X_1(\omega), \dots, x_n = X_n(\omega)$ testen, ob die zugrundeliegende Verteilung μ ist.

$$\textit{Nullhypothese } H_0\text{:} \qquad \text{»Verteilung}(X_i) \ =\ \mu\text{«}$$

$$\textit{Alternative } H_1\text{:} \qquad \text{»Verteilung}(X_i) \ \neq\ \mu\text{«}$$

Dazu betrachten wir die empirische Verteilung $\hat{p}_n$ der Stichproben, deren Gewichtsfunktion durch die relativen Häufigkeiten
$$\hat{p}_n(k) \ =\ \frac{H_n(k)}{n}, \qquad H_n(k) \ =\ \sum_{i=1}^{n} I_{\{X_i = k\}},$$
gegeben ist. Unter der Nullhypothese ist $\hat{p}_n(k)$ ein erwartungstreuer und konsistenter Schätzer für $p(k)$, denn
$$E_0\,[\hat{p}_n(k)] \ =\ p(k) \qquad \mathrm{Var}_0\,[\hat{p}_n(k)] \ =\ \frac{p(k)(1-p(k))}{n} \ \to\ 0 \quad \text{für} \quad n \to \infty.$$

Den Vektor $H_n = (H_n(1),\dots,H_n(l))$ bezeichnet man als auch als *Histogrammvektor*, da er graphisch durch ein Histogramm dargestellt wird. Entsprechend beschreibt der Vektor $\hat{p}_n = \frac{1}{n} H_n$ das Histogramm der relativen Häufigkeiten. Unter der Nullhypothese gilt $\hat{p}_n \approx p$ für große n. Daher liegt es nahe, die Nullhypothese zu verwerfen, falls $\hat{p}_n$ zu stark von p abweicht. Aber wann ist die Abweichung „zu groß"? Eine visuelle

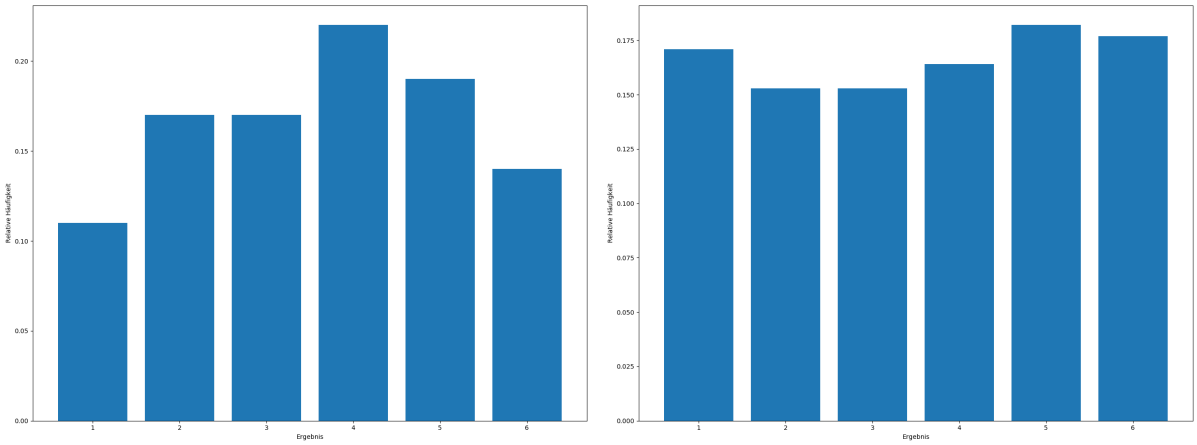


Abbildung 5.2: Histogramme der empirischen Verteilungen von 100 bzw. 1000 Würfen eines fairen Würfels.

Darstellung erhält man, indem man das Histogramm $\hat{p}_n$ der relativen Häufigkeiten plottet, und mit der Massenfunktion p vergleicht. Dabei sind sichtbare Abweichungen auch für große Werte von n durchaus normal, siehe Abildung 5.2. Es gibt verschiedene Arten, die Abweichung zu quantifizieren. Wir betrachten als Teststatistik die χ^2-*Divergenz* von $\hat{p}_n$ bezüglich p. Diese ist definiert als

$$T_n := n \sum_{k=1}^{n} \left(\frac{\hat{p}_n(k)}{p(k)} - 1 \right)^2 p(k) = \sum_{k=1}^{n} \frac{(H_n(k) - np(k))^2}{np(k)} .$$

Die Verwendung dieser speziellen Statistik hat verschiedene Gründe. Zum einen kann man damit gut rechnen, zum anderen handelt es sich um eine Approximation der relativen Entropie, welche eine natürliche Rolle in der Statistik und Informationstheorie spielt, siehe die Vorlesung Einführung in die Statistik. Mithilfe der folgenden Aussage können wir das asymptotische Niveau eines auf T_n basierenden Hypothesentests für $n \to \infty$ berechnen.

Satz 5.16 (Pearson (1900)). Unter der Nullhypothese konvergiert die Verteilung von T_n für $n \to \infty$ schwach gegen eine χ^2-Verteilung mit $l - 1$ Freiheitsgeraden

Hierbei ist die χ^2-Verteilung mit s Freiheitsgeraden definiert als die Verteilung von $\|Z\|^2 = Z_1^2 + \ldots + Z_s^2$, wobei $Z = (Z_1, \ldots, Z_s)$ ein standardnormalverteilter Zufallsvektor im $\mathbb{R}^s$ ist.

Für den Beweis des Satzes verweisen wir auf die Vorlesung Einführung in die Statistik. Wir skizzieren aber kurz die wesentlichen Schritte. Zunächst können wir die Verteilung des empirischen Histogrammvektors H_n unter der Nullhypothese explizit berechnen. Wir wissen bereits, dass die k-te Komponente $H_n(k)$ binomialverteilt ist mit Parametern n und $p(k)$. Mit einem ähnlichen kombinatorischen Argument überlegt man sich, dass der Vektor $H_n = (H_n(1), \ldots, H_n(l))$ die folgende **Multinomialverteilung** hat: Für alle $h_1, \ldots, h_l \in \mathbb{N}_0$ mit $h_1 + \cdots + h_l = n$ gilt

$$P_0 [H_n = (h_1, \ldots, h_l)] = \frac{n!}{h_1! h_2! \cdots h_l!} \prod_{k=1}^{l} p(k) .$$

Mit einer mehrdimensionalen Version des Satzes von De Moivre/Laplace folgt dann, dass die Verteilung des standardisierten Histogrammvektors

$$\widetilde{H}_n = \left(\frac{H_n(k) - np(k)}{\sqrt{np(k)}} \right)_{k=1,\ldots,l}$$

für $n \to \infty$ schwach gegen eine mehrdimensionale Normalverteilung konvergiert. Wegen $T_n = \|\widetilde{H}_n\|^2$ folgt dann auch die schwache Konvergenz der Verteilung von T_n gegen eine χ^2-Verteilung.

Wir betrachten nun den folgenden Hypothesentest für unser Testproblem:

Chiquadrat-Anpassungstest: Verwerfe H_0 falls $T_n \geq c$.

Hierbei ist c wie üblich ein zu spezifizierender Schwellenwert. Aufgrund des Satzes von Pearson gilt für die Wahrscheinlichkeit des Fehlers 1. Art:

$$P_0[T_n \geq c] \;\approx\; 1 - F_{\chi^2(l-1)}(c) \qquad \text{für große } n.$$

Also hat der Test mit Schwellenwert $c = q_{\chi^2(l-1),1-\alpha}$ asymptotisch das Niveau α.

Wahrscheinlichkeitsverteilungen auf $\mathbb{R}$: Kolmogorov-Smirnov-Test

Sei nun μ eine Wahrscheinlichkeitsverteilung auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Verteilungsfunktion F, und seien $X_1, X_2, \ldots$ unabhängige identisch verteilte, reellwertige Zufallsvariablen mit einer unbekannten Verteilung. Wir wollen erneut auf der Basis von Stichproben $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$ testen, ob die zugrundeliegende Verteilung μ ist.

$$\textit{Nullhypothese } H_0\text{:} \qquad \text{»Verteilung}(X_i) \;=\; \mu\text{«}$$

$$\textit{Alternative } H_1\text{:} \qquad \text{»Verteilung}(X_i) \;\neq\; \mu\text{«}$$

Dazu betrachten wir jetzt die empirische Verteilungsfunktion

$$\hat{F}_n(t) \;=\; \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq t\}}, \qquad t \in \mathbb{R}.$$

Unter der Nullhypothese ist die Zufallsvariable $\sum_{i=1}^{n} I_{\{X_i \leq t\}}$ für $t \in \mathbb{R}$ binomialverteilt mit Parametern n und $F(t)$, und es gilt $E_0\left[\hat{F}_n(t)\right] = F(t)$, sowie, nach der Bernstein-Ungleichung, gilt

$$P_0\left[\left|\hat{F}_n(t) - F(t)\right| \geq c\right] \;\leq\; 2e^{-2nc^2} \qquad \text{für alle } c > 0. \tag{5.19}$$

Für jedes feste t ist $\hat{F}_n(t)$ ein erwartungstreuer und konsistenter Schätzer für $F(t)$. Um zu testen, ob die gesamte empirische Verteilungsfunktion $\hat{F}_n$ sich stark von F unterscheidet, betrachten wir die Teststatistik

$$S_n \;:=\; \sup_{t \in \mathbb{R}} \left|\hat{F}_n(t) - F(t)\right| .$$

Die folgende bemerkenswerte Verschärfung der Bernstein-Ungleichung (5.19) wurde in dieser Form erst 1990 von Massart [**Massart**] bewiesen.

Satz 5.17 (Dvoretzky-Kiefer-Wolfowitz-Ungleichung). Unter der Nullhypothese gilt für jedes $c > 0$ die Abschätzung

$$P_0\left[S_n \geq c\right] \;\leq\; 2e^{-2nc^2} . \tag{5.20}$$

Auf den Beweis können wir hier nicht eingehen.

Mithilfe des Satzes können wir unmittelbar das Signifikanzniveau des folgenden Hypothesentests mit Schwellenwert c abschätzen:

Kolmogorov-Smirnov-Test: Verwerfe H_0 falls $S_n \geq c$.

Nach (5.20) erhalten wir $P_0[S_n \geq c] \leq \alpha$ für $c \geq \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}$.

Beispiel. Für die Stichprobengröße $n = 100$ erreichen wir das Niveau $\alpha = 5\%$ für den Schwellenwert $c = \sqrt{\frac{1}{200} \log(40)} = 0{,}018\ldots$. Unter der Nullhypothese weicht die empirische Verteilungsfunktion also in 95% der Fälle um weniger als $0{,}02$ von der tatsächlichen Verteilungsfunktion ab. Dies kann man zum Beispiel visuell überprüfen, indem man die Graphen der beiden Verteilungsfunktionen in ein Diagramm plottet.

Quantil-Quantil-Plots

Anstatt die Verteilungsfunktion der empirischen Verteilung mit der Verteilungsfunktion einer gegebenen Verteilung zu vergleichen, kann man auch die Quantile vergleichen. Dies geschieht visuell in einem Quantil-Quantil-Plot.

Test auf Gleichverteilung

Seien $U_1, U_2, \ldots U_n$ unabhängige, identisch verteilte Zufallsvariablen mit Werten im reellen Intervall $(0, 1)$. Wir wollen testen, ob die Zufallsvariablen gleichverteilt sind.

$$\textit{Nullhypothese } H_0\text{:} \qquad \text{»}U_i \sim \mathrm{Unif}(0,1)\text{«}$$

$$\textit{Alternative } H_1\text{:} \qquad \text{»}U_i \nsim \mathrm{Unif}(0,1)\text{«}$$

Für die Verteilungsfunktion der Gleichverteilung gilt $F(t) = t$ für alle $t \in [0, 1]$. Unter der Nullhypothese sollte die empirische Verteilungsfunktion ungefähr mit F übereinstimmen, falls n hinreichend groß ist. Entsprechend können wir erwarten, dass auch die Quantile der empirischen Verteilung, also die Ordnungsstatistiken $U_{(k)}$, $k = 1, \ldots, n$, ungefähr mit entsprechenden Quantilen der Gleichverteilung übereinstimmen. Tatsächlich gilt die folgende Aussage.

Lemma 5.18 (Verteilung der Ordnungsstatistiken von gleichverteilten Zufallsvariablen).
Unter der Nullhypothese ist die Verteilung von $U_{(k)}$ für $k \in \{1, 2, \ldots, n\}$ absolutstetig mit Dichte

$$f_{(k)}(u) \;=\; n\binom{n-1}{k-1} u^{k-1}(1-u)^{n-k}\; I_{(0,1)}(u).$$

Insbesondere gilt $E_0\left[U_{(k)}\right] = u_k$, und $\mathrm{Var}_0\left[U_{(k)}\right] = \frac{u_k(1-u_k)}{4(n+2)}$ mit $u_k := \frac{k}{n+1}$.

Beweis. Nach Satz 5.11 und Gleichung (5.12) gilt

$$P_0\left[U_{(k)} \leq c\right] \;=\; \mathrm{Bin}(n,c)[\{k, k+1, \ldots, n\}] \;=\; 1 - \mathrm{Bin}(n,c)[\{0, 1, \ldots, k-1\}] \;=\; 1 - \int_c^1 f_{(k)}(u)\, du$$

für alle $c \in [0, 1]$. Durch Ableiten folgt, dass $f_{(k)}$ die Dichte der Verteilung von $U_{(k)}$ unter P_0 ist. Den Erwartungswert und die Varianz erhält man nun durch explizite Berechnung der entsprechenden Integrale. ■

Das Lemma zeigt, dass die Varianz von $U_{(k)}$ von der Ordnung $O(1/n)$ ist. In diesem Sinne gilt unter der Nullhypothese

$$U_{(k)} \;\approx\; E_0\left[U_{(k)}\right] \;=\; u_k\,,$$

falls n „hinreichend groß" ist. Hierbei ist $u_k = k/(n+1)$ das $k/(n+1)$-Quantil der Gleichverteilung auf $(0, 1)$. Entsprechend überlegt man sich leicht, dass $U_{(k)}$ das $k/(n+1)$-Quantil der empirischen Verteilung ist. Beide Quantile sollten also für große n ungefähr gleich sein. Um dies visuell zu testen, plottet man in einem *Quantil-Quantil-Plot (QQ-Plot)* ein Streudiagramm, in dem die Punkte $(u_k, U_{(k)}) \in (0, 1)^2$ für $k = 1, \ldots, n$ aufgetragen sind. Ist n hinreichend groß und gilt die Nullhypothese, dann sollten die Punkte im QQ-Plot nahe der Winkelhalbierenden liegen.

Test auf allgemeine Verteilung

Sei nun μ eine allgemeine Wahrscheinlichkeitsverteilung auf $\mathbb{R}$ mit stetiger Verteilungsfunktion F und Quantilfunktion $\underline{G}(u) = \inf\{x \in \mathbb{R} : F(x) \geq c\}$. Um auf diese Verteilung zu testen, kann man ähnlich wie oben vorgehen. Seien $X_1, X_2, \ldots X_n$ unabhängige, identisch verteilte reelle Zufallsvariablen mit Werten im reellen Intervall $(0,1)$. Wir betrachten das Testproblem

$$\textit{Nullhypothese } H_0: \qquad »X_i \sim \mu«$$

$$\textit{Alternative } H_1: \qquad »X_i \nsim \mu«$$

Seien $U_1, U_2, \ldots U_n$ unabhängige, auf $(0,1)$ gleichverteilte Zufallsvariablen. Dann hat nach Satz 4.25 X_i unter der Nullhypothese für jedes i dieselbe Verteilung wie $\underline{G}(U_i)$, und wegen der Unabhängigkeit stimmt auch die gemeinsame Verteilung von $X_1, \ldots, X_n$ mit der gemeinsamen Verteilung von $\underline{G}(U_1), \ldots, \underline{G}(U_n)$ überein. Da $\underline{G}$ eine monoton wachsende Funktion ist, ist $\underline{G}(U_{(k)})$ die k-te Ordnungsstatistik der Zufallsvariablen $\underline{G}(U_1), \ldots, \underline{G}(U_n)$, und somit hat $X_{(k)}$ dieselbe Verteilung wie $\underline{G}(U_{(k)})$. Nach dem Lemma von oben folgt also unter der Nullhypothese

$$X_{(k)} \sim \underline{G}(U_{(k)}) \approx \underline{G}(u_k)$$

für n „hinreichend groß". Auch diese Aussage können wir mit einem QQ-Plot visuell testen. Dazu tragen wir jetzt die Punkte $(\underline{G}(u_k), U_{(k)}) \in (0,1)^2$ für $k = 1, \ldots, n$ in einem Streudiagramm auf. Ist n hinreichend groß und gilt die Nullhypothese, dann sollten diese Punkte wiederum nahe der Winkelhalbierenden liegen.

Test auf Normalverteilung

Viele statistische Verfahren basieren auf einer Normalverteilungsannahme. Um diese anwenden zu können möchte man testen, ob die zugrundeliegenden Stichproben (näherungsweise) von einer Normalverteilung stammen. Der Erwartungswert und die Varianz der zugrundeliegenden Verteilung sind dabei meist unbekannt. Wir erhalten also das folgende Testproblem:

$$\textit{Nullhypothese } H_0: \qquad »X_i \sim N(m, \sigma^2) \quad \text{für ein } m \in \mathbb{R} \text{ und } \sigma > 0.«$$

$$\textit{Alternative } H_1: \qquad »\text{Die Verteilung von } X_i \text{ ist keine Normalverteilung.}«$$

Hierbei setzen wir wieder voraus, dass $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen sind.

Unter der Nullhypothese gilt in diesem Fall

$$X_i = m + \sigma Z_i = m + \sigma \Phi^{-1}(U_i)$$

mit unabhängigen standardnormalverteilten Zufallsvariablen $Z_1, \ldots, Z_n$, sowie unabhängigen auf $(0,1)$ gleichverteilten Zufallsvariablen $U_1, \ldots, U_n$. Hierbei ist Φ die Verteilungsfunktion und Φ^{-1} die Quantilfunktion der Standardnormalverteilung. Ähnlich wie oben folgt für die Ordnungsstatistiken

$$X_{(k)} = m + \sigma \Phi^{-1}(U_{(k)}) \approx m + \sigma \Phi^{-1}(u_k)$$

für hinreichend große n. Um dies visuell zu testen, plottet man in einem *Normal-QQ-Plot* das Streudiagramm der Punkte

$$\left(\Phi^{-1}(u_k), X_{(k)}\right), \qquad k = 1, \ldots, n.$$

Gilt die Nullhypothese, dann sollten die Punkte näherungsweise auf der Geraden $y = m + \sigma x$ liegen. In diesem Fall entsprechen die Parameter m und σ dem Achsenabschnitt und der Steigung der Geraden.

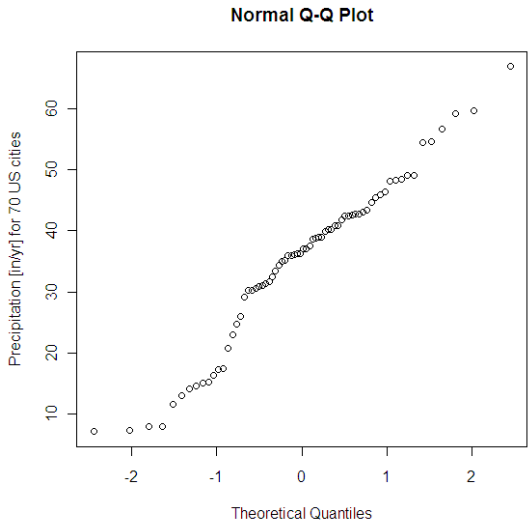


Abbildung 5.3: Normal-QQ-Plot eines empirischen Datensatzes

5.4 Pseudozufallszahlen und Simulationsverfahren

Zufallszahlengeneratoren

Simulationsverfahren für Stichproben von Wahrscheinlichkeitsverteilungen gehen in der Regel von der Existenz einer Folge von auf dem reellen Intervall $[0,1]$ gleichverteilten, unabhängigen Zufallszahlen aus, die durch einen Zufallszahlengenerator erzeugt werden. In Wirklichkeit simulieren Zufallszahlengeneratoren natürlich nur auf $\left\{k\, m^{-1} : k = 0, 1, \ldots, m-1\right\}$ gleichverteilte Zufallszahlen, wobei m^{-1} die Darstellungsgenauigkeit des Computers ist. Außerdem ist eine Folge von vom Computer erzeugten Pseudozufallszahlen eigentlich gar nicht zufällig, sondern deterministisch.

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

Konkret werden Pseudozufallszahlen üblicherweise über eine deterministische Rekurrenzrelation vom Typ

$$x_{n+1} = f\left(x_{n-k+1}, x_{n-k+2}, \ldots, x_n\right), \qquad n = k, k+1, k+2, \ldots,$$

aus *Saatwerten* $x_1, x_2, \ldots, x_k$ erzeugt. In vielen Fällen hängt die Funktion f nur von der letzten erzeugten Zufallszahl x_n ab. Beispiele von Pseudozufallszahlengeneratoren sind lineare Kongruenzgeneratoren und Shift-Register-Generatoren.

Lineare Kongruenzgeneratoren

Bei einem linearen Kongruenzgenerator (LCG) ist die Rekurrenzrelation vom Typ

$$x_{n+1} = (ax_n + c) \mod m, \qquad n = 0, 1, 2, \ldots.$$

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

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

Ein erstes Problem, dass bei linearen Kongruenzgeneratoren auftreten kann, ist, dass die Folge von Pseudozufallszahlen periodisch mit einer Periode ist, die im Allgemeinen deutlich kleiner als die maximal mögliche Periode m sein kann:

> **Beispiel (LCG mit kleiner Periode).** Wählen wir $m = 63$, $a = 11$ und $c = 0$, dann hat die Folge der vom linearen Kongruenzgenerator erzeugten Pseudozufallszahlen die Periode 6, siehe Abbildung 5.4.

Dieses erste Problem lässt sich leicht mithilfe der folgenden Charakterisierung aller linearen Kongruenzgeneratoren mit der maximal möglichen Periode m umgehen:

Satz 5.19 (Knuth). Die Periode eines LCG ist gleich m genau dann, wenn

(i) c und m teilerfremd sind,

(ii) jeder Primfaktor von m ein Teiler von $a - 1$ ist, und

(iii) falls 4 ein Teiler von m ist, dann auch von $a - 1$.

Der Beweis dieses zahlentheoretischen Satzes findet sich in [2].

Auch wenn ein linearer Kongruenzgenerator die maximal mögliche Periode hat, können weitere Probleme durch versteckte Strukturen und Symmetrien auftreten. Bei einigen einfachen Generatoren werden diese Probleme schon sichtbar, wenn man die Pseudozufallszahlen benutzt, um zwei- oder dreidimensionale Pseudozufallsvektoren zu erzeugen. Dies ist in den Abbildungen 5.5 und 5.6 demonstriert. Der Marsaglia-Generator besteht alle drei Tests; da in Wirklichkeit aber auch dieser deterministische Werte liefert, kann man auch hier einen Test konstruieren, der die Pseudozufallszahlen von echten Zufallszahlen unterscheidet.

Shift-Register-Generatoren

Eine andere Rekurrenzrelation wird zur Erzeugung von Pseudozufallszahlen mit Shift-Register-Generatoren verwendet. Hier interpretiert man eine Zahl $x_n \in \{0,1,\ldots,2^k - 1\}$ zunächst als Binärzahl bzw. als Vektor aus $\{0,1\}^k$, und wendet dann eine gegebene Matrix T darauf an, um x_{n+1} zu erhalten:

$$x_{n+1} = Tx_n, \qquad n = 0, 1, 2, \ldots.$$

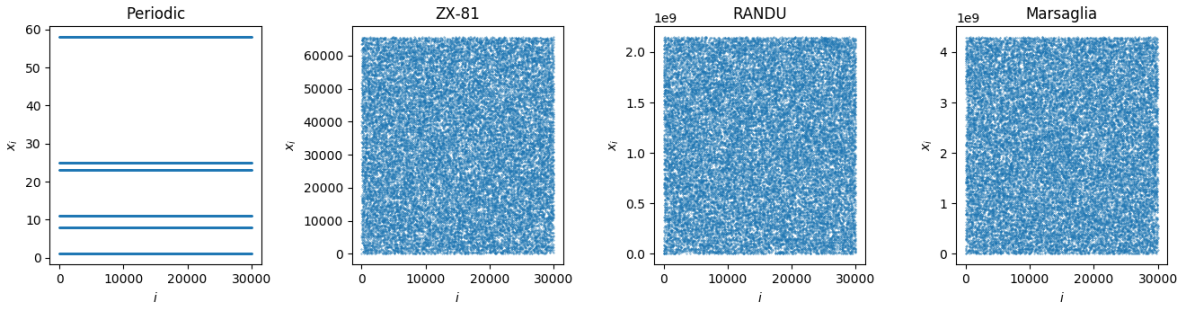


Abbildung 5.4: Plots der Folgen $x_1, \ldots, x_{30000}$ für den LCG mit Periode 6 aus dem Beispiel, sowie für den ZX81-Generator, RANDU, und den Marsaglia-Generator.

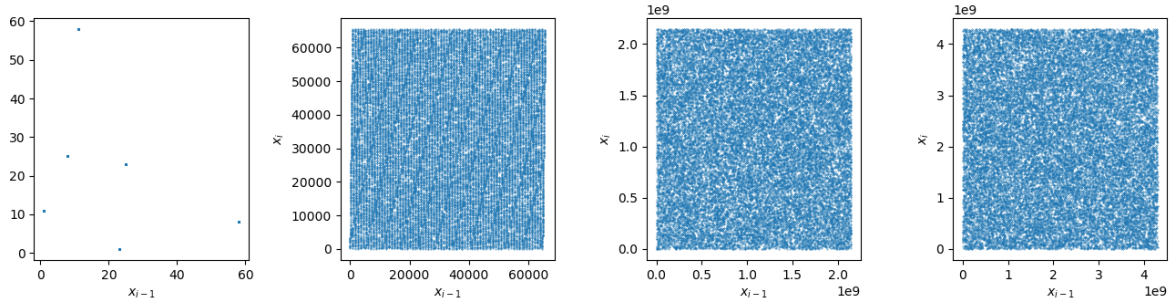


Abbildung 5.5: Fassen wir Paare (x_i, x_{i+1}) von aufeinanderfolgenden Pseudozufallszahlen als Koordinaten eines zweidimensionalen Pseudozufallsvektors auf, und betrachten die empirische Verteilung dieser Vektoren, so ergibt sich beim ZX81-Generator keine besonders gute Approximation einer zweidimensionalen Gleichverteilung.

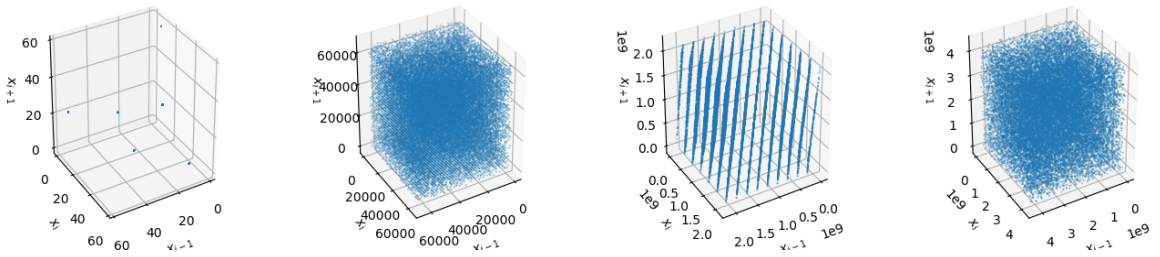


Abbildung 5.6: Fassen wir analog jeweils drei aufeinanderfolgende Pseudozufallszahlen als dreidimensionale Vektoren auf, dann konzentrieren sich diese beim RANDU-LCG auf 15 Hyperebenen.

Kombination von Zufallszahlengeneratoren

Generatoren von Pseudozufallszahlen lassen sich kombinieren, zum Beispiel indem man die von mehreren Zufallszahlengeneratoren erzeugten Folgen von Pseudozufallszahlen aus $\{0,1,\ldots,m-1\}$ modulo m addiert. Auf diese Weise erhält man sehr leistungsfähige Zufallszahlengeneratoren, zum Beispiel den Kiss-Generator von Marsaglia [[3]], der einen LCG und zwei Shift-Register-Generatoren kombiniert, Periode 2^{95} hat, und umfangreiche statistische Tests besteht.

Physikalische Zufallszahlengeneratoren

Alternativ werden Zufallszahlen auch mithilfe von physikalischen und insbesondere quantenmechanischen Vorgängen erzeugt, z.B. durch radioaktive Zerfälle, thermisches Rauschen, Atmosphärenrauschen etc. Ein Nachteil ist, dass auf diese Weise nur eine begrenzte Anzahl unabhängiger Stichproben pro Zeiteinheit erzeugt werden kann. Zudem sind die erhaltenen Ergebnisse nicht reproduzierbar. Auch physikalische Zufallsgeneratoren können mit algorithmischen Pseudozufallszahlengeneratoren kombiniert werden.

Statistische Tests für Zufallszahlengeneratoren

Wie wir schon in den Abbildung 5.4, 5.5 und 5.6 gesehen haben, können Schwachstellen von Zufallszahlengeneratoren mithilfe statistischer Tests aufgezeigt werden. Wir wollen kurz auf die dabei zugrundeliegende Argumentation eingehen. Eine von einem Zufallszahlengenerator erzeugte Folge $x_1, x_2, x_3, \ldots$ soll eine Folge von Stichproben von *unabhängigen* Zufallsvariablen $X_1, X_2, X_3, \ldots$ simulieren, die auf der Menge $\{0, 1, \ldots, m-1\}$ *gleichverteilt* ist. Es stellt sich also die Frage, ob die erzeugte Zahlenfolge zu diesem mathematischen Modell passt. Um dies zu testen, leitet man aus den Modellannahmen Folgerungen her, und überprüft ob die erzeugte Zahlenfolge konsistent mit diesen Folgerungen ist.

> **Beispiel (Blocktest).** Sei d eine natürliche Zahl. Sind die Zufallsvariablen X_i unabhängig und gleichverteilt auf $\{0, 1, \ldots, m-1\}$, dann sind auch die Zufallsvektoren $(X_{(k-1)d+1}, X_{(k-1)d+2}, \ldots, X_{kd})$, $k \in \mathbb{N}$, wieder unabhängig und gleichverteilt auf dem Produktraum $\{0, 1, \ldots, m-1\}^d$. Genau dies haben wir in den Abbildungen 5.5 und 5.6 für $d = 2$ bzw. $d = 3$ graphisch getestet. In höheren Dimensionen versagt zwar der graphische Test, aber wir können weiterhin rechnerisch testen, ob sich zum Beispiel die relativen Häufigkeiten von Werten in einem bestimmten Bereich $A \subseteq \{0, 1, \ldots, m-1\}^d$ der simulierten Zufallsvektoren $(x_{(k-1)d+1}, x_{(k-1)d+2}, \ldots, x_{kd})$, $k = 1, \ldots, n$, für große n der Wahrscheinlichkeit von A unter der Gleichverteilung annähern.

Prinzipiell kann jede Folgerung aus den Modellannahmen zur Konzeption eines statistischen Tests verwendet werden. Beispielsweise haben wir in der Einleitung einen Test für 0-1-Zufallsfolgen betrachtet, der auf der Anzahl der Runs basiert. Da jeder Test nur einen bestimmten Aspekt berücksichtigen kann, ist auch für einen Zufallsgenerator, der viele der üblichen Tests besteht, noch nicht garantiert, dass er für eine konkrete Anwendung wirklich geeignet ist. Es kann daher sinnvoll sein, die Ergebnisse einer stochastischen Simulation mit verschiedenen Generatoren zu reproduzieren.

Direkte Simulationsverfahren

Aus den von einem Pseudozufallszahlengenerator zunächst erzeugten Pseudozufallszahlen mit Werten in der endlichen Menge $\{0, 1, \ldots, m-1\}$ kann man anschließend Pseudo-Stichproben von anderen Wahrscheinlichkeitsverteilungen erzeugen.

Zufällige Permutationen

Der folgende Algorithmus erzeugt eine (pseudo-)zufällige Permutation aus $\mathcal{S}_n$:

Algorithmus 3: Zufällige Permutation

Input : $n \in \mathbb{N}$
Output Zufällige Permutation der Länge n
:

```
1 for i ← 1 to n do
2 |  x_i ← i
3 for i ← 1 to n − 1 do
4 |  k ← i + ZufälligeGanzzahl({ 0,1,...,n-i});
5 |  Vertausche(x_i, x_k);
6 return (x_i)_{i=1}^n;
```

Zufallszahlen aus $[0, 1)$

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

$$u_n = \frac{x_n}{m}, \qquad x_n \in \{0, 1, \ldots, m-1\},$$

erzeugt, wobei m vorgegeben ist (zum Beispiel Darstellungsgenauigkeit des Computers), und x_n eine Folge ganzzahliger Pseudozufallszahlen aus {0, 1, . . . , m - 1} ist.

Stichproben von diskreten Zufallsvariablen

Wir nehmen nun an, dass wir eine Folge $u_1, u_2, \ldots$ von Stichproben von auf $(0, 1)$ gleichverteilten, unabhängigen Zufallsvariablen $U_1, U_2, \ldots$ gegeben haben. Die oben beschriebenen Probleme beim Generieren solcher Stichproben werden wir im folgenden ignorieren. Stattdessen wollen wir uns nun überlegen, wie wir aus der Folge (u_n) Stichproben von einer vorgegebenen Wahrscheinlichkeitsverteilung μ auf einer abzählbaren Menge S erzeugen können.

Dazu können wir das schon im Beispiel in Abschnitt 4.5 beschriebene Verfahren (siehe Abbildung 4.11) leicht an den allgemeineren Rahmen anpassen. Wir gehen davon aus, dass wir die Gewichte $\mu(a) = \mu[\{a\}]$ zumindest bis auf eine Normierungskonstante kennen bzw. berechnen können. Sei $a_1, a_2, \ldots$ eine Abzählung der Elemente von S. Wir betrachten die durch

$$s_k \;:=\; \sum_{i=1}^{k} \mu(a_i) \;=\; \mu[\{a_1, \ldots, a_k\}] \tag{5.21}$$

definierte Verteilungsfunktion. Wir gehen davon aus, dass wir die Werte $\mu(a_i)$ und damit auch s_i für jedes $i \in \mathbb{N}$ berechnen können. Für $n, i \in \mathbb{N}$ setzen wir

$$x_n \;:=\; a_i \qquad \text{falls } s_{i-1} < u_n \le s_i.$$

Dann ist x_n eine Stichprobe von der Zufallsvariable

$$X_n \;:=\; \sum_i a_i \, I_{s_{i-1} < U_n \le s_i}.$$

Lemma 5.20. *Sind U_n ($n \in \mathbb{N}$) unabhängige Zufallsvariablen mit Verteilung $U_n \sim \mathrm{Unif}(0,1)$, dann sind X_n ($n \in \mathbb{N}$) unabhängige Zufallsvariablen mit Verteilung $X_n \sim \mu$.*

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

$$P[X_n = a_i] = P[s_{i-1} < U_n \le s_i] = P[U_n \le s_i] - P[U_n \le s_{i-1}] = s_i - s_{i-1} = \mu(a_i).$$

Also hat X_n die Verteilung μ. Der Nachweis der Unabhängigkeit ist eine Übungsaufgabe. ■

Algorithmus 4: Direkte Simulation einer Stichprobe von einer diskreten Wahrscheinlichkeitsverteilung

Input : Gewichte $(\mu(a_i))_{i\in\mathbb{N}}$
Output Pseudozufallsstichprobe x von μ
:

```
1 i ← 1 ;
2 s ← μ(a_1) ;
3 u ← Stichprobe(Unif[0,1]);
4 while u > s do
5 |   i ← i + 1;
6 |   s ← s + μ(a_i)
7 return x := a_i;
```

Bemerkung (Mittlere Laufzeit). Die mittlere Anzahl von Schritten des Algorithmus ist gleich $\sum i\, \mu(a_i)$.

Nach der Bemerkung ist das direkte Verfahren im Allgemeinen nur dann praktikabel, wenn die Gewichte $\mu(a_i)$ für große i rasch abfallen. In einigen einfachen Spezialfällen kann man jedoch eine explizite Formel zur Berechnung von x_n aus u_n angeben, für deren Auswertung die Schleife in Algorithmus 4 nicht durchlaufen werden muss:

Übung (Simulation von Stichproben einer geometrischen Verteilung). Geben Sie ein direktes Verfahren an, dass in einem Schritt aus einer Stichprobe von der Gleichverteilung auf dem Intervall $(0,1)$ eine Stichprobe von der geometrischen Verteilung mit Parameter $p \in (0,1)$ erzeugt.

Stichproben von reellen Zufallsvariablen

Auch das direkte Simulationsverfahren zum Erzeugen von Stichproben von allgemeinen reellwertigen Zufallsvariablen haben wir in Abschnitt 4.5 bereits besprochen. Ist μ eine Wahrscheinlichkeitsverteilung auf $\mathbb{R}$, und sind $u_1, u_2, \ldots$ Stichproben von auf $(0,1)$ gleichverteilten, unabhängigen Zufallsvariablen $U_1, U_2, \ldots$, dann sind die Werte

$$x_n := \underline{G}(u_n)$$

Stichproben von den unabhängigen, nach μ verteilten, Zufallsvariablen $X_n = \underline{G}(U_n)$, siehe den Algorithmus und die Beispiele am Ende von Abschnitt 4.5 Die direkte Anwendung dieses Verfahrens setzt natürlich die explizite Berechenbarkeit der Quantilfunktion $\underline{G}$ voraus.

Stichproben von normalverteilten Zufallsvariablen

Die Verteilungsfunktion einer $N(0,1)$-verteilten Zufallsvariable ist

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt\ .$$

Das Integral ist nicht explizit lösbar und die Inverse Φ^{-1} ist dementsprechend nur approximativ berechenbar. Daher ist die Simulation einer Standardnormalverteilung durch Inversion der Verteilungsfunktion relativ aufwändig. Ein einfacheres Verfahren zur Simulation von Stichproben einer Normalverteilung ergibt sich, wenn wir eine zweidimensionale Standardnormalverteilung betrachten und auf Polarkoordinaten transformieren.

Algorithmus 5: Box-Muller-Verfahren

Input : $m \in \mathbb{R}, \sigma > 0$
Output Zwei unabhängige Stichproben z_1, z_2 von $N(0,1)$
:
1 Erzeuge unabhängige Zufallszahlen $u_1, u_2 \sim \mathrm{Unif}_{(0,1)}$;
2 $z_1 := \sqrt{-2\log u_1}\cos(2\pi u_2)$, $z_2 := \sqrt{-2\log u_1}\sin(2\pi u_2)$;
3 **return** z_1, z_2;

Der Beweis, dass z_1 und z_2 tatsächlich unabhängige Stichproben von der Standardnormalverteilung sind, basiert auf dem mehrdimensionalen Dichtetransformationssatz, siehe die Vorlesung Einführung in die Wahrscheinlichkeitstheorie. Aus diesem folgt, dass die zweidimensionale Zufallsvariable

$$Z \;=\; (Z_1, Z_2) \;=\; \left(\sqrt{-2\log U_1}\cos(2\pi U_2)\,,\, \sqrt{-2\log U_1}\sin(2\pi U_2)\right)$$

genau dann standardnormalverteilt ist, wenn $U = (U_1, U_2)$ auf $(0,1)\times(0,1)$ gleichverteilt ist.

Unabhängige Stichproben x_1, x_2 von der Normalverteilung mit Mittelwert m und Varianz σ^2 kann man dann über die Transformation

$$x_1 \;:= \sigma z_1 + m, \quad x_2 \;:= \sigma z_2 + m$$

erzeugen.

Acceptance-Rejection-Verfahren

Da die direkten Simulationsverfahren oft nicht praktikabel sind, benötigen wir Alternativen. Eine häufig verwendete Methode besteht darin, zunächst unabhängige Stichproben von einer „einfacheren" Wahrscheinlichkeitsverteilung ν auf demselben Zustandsraum S zu generieren, und daraus mit einem Verwerfungsverfahren Stichproben von der Zielverteilung μ zu erzeugen.

Der Einfachheit halber betrachten wir den diskreten Fall. Für absolutstetige Verteilungen kann man ähnlich verfahren, wenn man die Massenfunktionen durch die Dichten ersetzt. Wir nehmen an, dass wir den Quotienten $\mu(x)/\nu(x)$ der Massenfunktionen der Verteilungen μ bzw. ν bis auf eine Proportionalitätskonstante kennen, d.h. für $x \in S$ gilt

$$\mu(x) \;\propto f(x)\nu(x) \tag{5.22}$$

mit einer explizit bekannten Funktion $f : S \to \mathbb{R}$. Beispielsweise können wir $f(x) = \mu(x)/\nu(x)$ setzen, wenn dieses Verhältnis explizit bekannt ist. Wir setzen zudem voraus, dass wir eine obere Schranke c für die Funktion f kennen, d.h.

$$\text{es gibt ein } c \in [1,\infty), \text{ so dass} \qquad f(x) \le c \qquad \text{für alle } x \in S. \tag{5.23}$$

Angenommen, wir können Folgen von Stichproben x_n, u_n ($n \in \mathbb{N}$) von unabhängigen Zufallsvariablen X_n, U_n mit Verteilung ν bzw. $\mathrm{Unif}(0,1)$ erzeugen. Dann können wir daraus Stichproben von der Zielverteilung μ generieren, indem wir die x_n als Vorschlagswerte betrachten, die mit einer Wahrscheinlichkeit proportional zu $f(x_n)$ akzeptiert, und ansonsten verworfen werden. Aufgrund der Annahme (5.23) können die *Akzeptanzwahrscheinlichkeiten* dabei gleich $f(x)/c$ gewählt werden.

Algorithmus 6: Acceptance-Rejection-Verfahren (AR)

Input : $f : S \to [0,\infty)$, $c \in [1,\infty)$ mit (5.23)
Output Stichprobe x von Wahrscheinlichkeitsverteilung μ mit (5.22)
:

repeat
  $x \leftarrow$`Stichprobe(`ν`)` ;
  $u \leftarrow$`Stichprobe(`Unif$(0,1)$`)`;
until $u \le \frac{f(x)}{c}$;
return x;

Wir wollen den Algorithmus nun analysieren. Seien dazu $X_n \sim \nu$ und $U_n \sim \mathrm{Unif}(0,1)$ ($n \in \mathbb{N}$) unabhängige Zufallsvariablen, die auf einem gemeinsamen Wahrscheinlichkeitsraum definiert sind. Die diskrete Zufallsvariable

$$T(\omega) \;=\; \min\{n \in \mathbb{N} \,:\, U_n(\omega) \le f(X_n(\omega))/c\}$$

beschreibt dann die Anzahl der Durchläufe der Schleife bis erstmals ein Vorschlag X_n akzeptiert wird, und

$$X_T(\omega) \;=\; X_{T(\omega)}(\omega)$$

ist der akzeptierte Wert, der schließlich ausgegeben wird.

Satz 5.21 (Laufzeit und Output des AR-Verfahrens).

(i) T ist *geometrisch verteilt* mit Parameter $p = \sum_{a\in S} \frac{f(a)\nu(a)}{c}$. Insbesondere ist T fast sicher endlich.

(ii) Die Zufallsvariable X_T hat die Verteilung μ.

Der Satz zeigt, dass der Algorithmus tatsächlich eine Stichprobe von der Verteilung μ liefert. Die mittlere Anzahl von Schritten, bis ein Vorschlag akzeptiert wird, beträgt $E[T] = 1/p$. Ist $f = \mu/\nu$, dann ist $p = 1/c$, also die mittlere Laufzeit gleich c.

Beweis (von Satz 5.21). (i) Sei $A_n := \{U_n \le f(X_n)/c\}$ das Ereignis, dass der n-te Vorschlag akzeptiert wird. Aus der Unabhängigkeit der Zufallsvariablen $X_1, U_1, X_2, U_2, \ldots$ folgt, daß auch die Ereignisse $A_1, A_2, \ldots$ unabhängig sind. Zudem gilt wegen der Unabhängigkeit von X_n und U_n:

$$\begin{aligned} P[A_n] \;&=\; \sum_{a\in S} P\left[\{X_n = a\} \cap A_n\right] \;=\; \sum_{a\in S} P\left[X_n = a, U_n \le f(a)/c\right] \\ &=\; \sum_{a\in S} P\left[X_n = a\right] \cdot P\left[U_n \le f(a)/c\right] \;=\; \sum_{a\in S} \nu(a) f(a)/c = p. \end{aligned}$$

Also ist $T(\omega) = \min\{n \in \mathbb{N} : \omega \in A_n\}$ geometrisch verteilt mit Parameter p.

(ii) Für $a \in S$ gilt

$$\begin{aligned} P[X_T = a] &= \sum_{n=1}^{\infty} P[\{X_T = a\} \cap \{T = n\}] \\ &= \sum_{n=1}^{\infty} P[\{X_n = a\} \cap A_n \cap A_1^C \cap \ldots \cap A_{n-1}^C] \\ &= \sum_{n=1}^{\infty} P[\{X_n = a, U_n \le f(a)/c\} \cap A_1^C \cap \ldots \cap A_{n-1}^C] \\ &= \sum_{n=1}^{\infty} \nu(a)\,\frac{f(a)}{c}\,(1-p)^{n-1} \;=\; \frac{f(a)\nu(a)}{pc}. \end{aligned}$$

Hierbei haben wir im letzten Schritt benutzt, dass die Ereignisse $\{X_n = a\}$, $\{U_n \leq f(a)/c\}$, sowie $A_1^C, \ldots, A_{n-1}^C$ unabhängig sind. Da μ die einzige Wahrscheinlichkeitsverteilung ist, deren Massenfunktion proportional zu $f(a)\nu(a)$ ist, folgt $X_T \sim \mu$. ■

Beispiel (Simulation von bedingten Verteilungen). Das Acceptance-Rejection-Verfahren kann prinzipiell verwendet werden, um Stichproben von einer bedingten Verteilung $\mu[A] = \nu[A|B]$ zu simulieren, wobei $B \subseteq S$ ein Ereignis mit $\nu[B] > 0$ ist. In diesem Fall gilt $\mu(x) = f(x)\nu(x)$ mit

$$f(x) \;=\; I_B(x)/\nu[B] \;\leq\; 1/\nu[B] \qquad \text{für alle } x \in S,$$

so dass wir $c = 1/\nu[B]$ wählen können. Das AR-Verfahren erzeugt dann Stichproben von der Verteilung ν und akzeptiert diese mit Wahrscheinlichkeit $I_B(x)$, d.h., Stichproben in B werden stets akzeptiert. Da die mittlere Laufzeit gleich c ist, ist das Verfahren nur dann praktikabel, wenn die Wahrscheinlichkeit von B nicht zu klein ist.

Monte-Carlo-Verfahren

Sei μ eine Wahrscheinlichkeitsverteilung auf einer Menge S mit σ-Algebra $\mathcal{B}$. Angenommen, wir wollen die Wahrscheinlichkeit

$$p \;:=\; \mu[B]$$

eines Ereignisses $B \in \mathcal{B}$ beziehungsweise, allgemeiner, den Erwartungswert

$$\theta \;:=\; E_\mu[f]$$

einer reellwertigen Zufallsvariable $f\colon S \to \mathbb{R}$ mit $E_\mu[f^2] < \infty$ näherungsweise berechnen. In einem solchen Fall können wir auf ein Monte-Carlo-Verfahren zurückgreifen. Hierbei simuliert man eine große Anzahl Stichproben $X_1(\omega), \ldots, X_n(\omega)$ von unabhängigen Zufallsvariablen mit Verteilung μ (*klassisches Monte-Carlo-Verfahren*), beziehungsweise von einer konvergenten Markovkette mit Gleichgewicht μ (*Markov Chain Monte Carlo*). Nach dem Gesetz der großen Zahlen liefern dann die relativen Häufigkeiten

$$\widehat{p}_n(\omega) \;:=\; \frac{1}{n}\sum_{i=1}^{n} I_B(X_i(\omega)).$$

bzw. die empirischen Mittelwerte

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

Schätzwerte für p bzw. θ, die sich für $n \to \infty$ den gesuchten Werten annähern. Wenn die Zufallsvariablen X_i alle die Verteilung μ haben, dann gilt

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

d.h. $\widehat{\theta}_n$ ist ein *erwartungstreuer Schätzer* für θ. Sind die Zufallsvariablen X_i unabhängig mit Verteilung μ, dann sind die Zufallsvariablen $f(X_i)$ unkorreliert, und es ergibt sich

$$\mathrm{MSE}\left[\widehat{\theta}_n\right] \;:=\; E\left[\left|\widehat{\theta}_n - \theta\right|^2\right] \;=\; \mathrm{Var}\left[\widehat{\theta}_n\right] \;=\; \frac{1}{n}\,\mathrm{Var}_\mu[f],$$

d.h. der *mittlere quadratische Fehler* des Schätzers ist von der Ordnung $O(1/n)$. Der mittlere quadratische Fehler fällt also relativ langsam in n ab. Ein großer Vorteil ist jedoch, dass die Abschätzung völlig *problemunabhängig* ist. Aus diesem Grund sind Monte-Carlo-Verfahren sehr universell einsetzbar. In komplizierten Modellen sind sie oft die einzige praktikable Option um Erwartungswerte näherungsweise zu berechnen. Abschätzungen für den Schätzfehler $\widehat{\theta}_n - \theta$ und Konfidenzintervalle kann man mit den in Abschnitt 5.2 beschriebenen Methoden herleiten, zum Beispiel über die Čebyšev-Ungleichung, mit exponentiellen Abschätzungen, oder über eine Normalapproximation.

Beispiel (Monte-Carlo-Schätzung von hochdimensionalen Integralen). Auch die Werte von mehrdimensionalen Integralen können mit Monte-Carlo-Verfahren näherungsweise berechnet werden. Dies ist besonders in hohen Dimensionen von Interesse, wo klassische numerische Verfahren in der Regel versagen. Soll beispielsweise der Wert des Integrals

$$\theta \coloneqq \int_{[0,1]^d} f(x)\,\mathrm{d}x \coloneqq \int_0^1 \dots \int_0^1 f(x_1,\dots,x_d)\,\mathrm{d}x_1 \dots \mathrm{d}x_d.$$

näherungsweise berechnet werden, dann können wir dazu Stichproben $u_1, u_2, \dots, u_{dn}$ von unabhängigen Zufallsvariablen $U_i \sim \mathrm{Unif}(0,1)$ simulieren. Die d-dimensionalen Zufallsvektoren $X^{(i)} \coloneqq (U_{di+1}, \dots, U_{d(i+1)})$, $i = 1, \dots, n$, sind dann unabhängig und gleichverteilt auf dem Produktraum $(0,1)^d$. Daher können wir den Wert θ des Integrals durch den Monte-Carlo-Schätzer

$$\hat{\theta}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)}) = \frac{1}{n}\sum_{i=1}^{n} f(u_1, u_2, \dots, u_{dn})$$

approximieren. Ist die Funktion f quadratintegrierbar, dann ergibt sich eine *dimensionsunabhängige* Abschätzung des mittleren quadratischen Fehlers, die nur von der Varianz von f bzgl. der Gleichverteilung auf dem Einheitswürfel $(0,1)^d$ abhängt. Da zum Erzeugen eines Stichprobenvektors $x^{(i)}$ d Zufallszahlen aus $(0,1)$ benötigt werden, beträgt der Aufwand $O(d)$, wenn ein vorgegebener mittlerer quadratischen Fehler für Funktionen mit Varianz kleiner gleich 1 unterschritten werden soll. Klassische numerische Integrationsverfahren haben dagegen in der Regel einen Aufwand, der exponentiell in der Dimension wächst.

Index

Literatur

[1] Achim Klenke. *Probability theory*. German. Universitext. A comprehensive course. Springer, London, 2014, S. xii+638. ISBN: 978-1-4471-5360-3; 978-1-4471-5361-0. URL: https://doi.org/10.1007/978-1-4471-5361-0.

[2] Donald E. Knuth. *The Art of Computer Programming, Volume 2 (3rd Ed.): Seminumerical Algorithms*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1997. ISBN: 0-201-89684-2.

[3] George Marsaglia und Arif Zaman. *The KISS generator*. Techn. Ber. Tech. rep., Department of Statistics, University of Florida, 1993.

[4] David Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991, S. xvi+251. ISBN: 0-521-40455-X; 0-521-40605-6. URL: https://doi.org/10.1017/CBO9780511813658.