Algorithmische Mathematik II

Andreas Eberle

24. Juni 2017

Inhaltsverzeichnis

Teil I

Diskrete Stochastik

„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:

$$x_{n+1} \;=\; (\,69069\, x_n + 1\,) \bmod m\,,$$

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

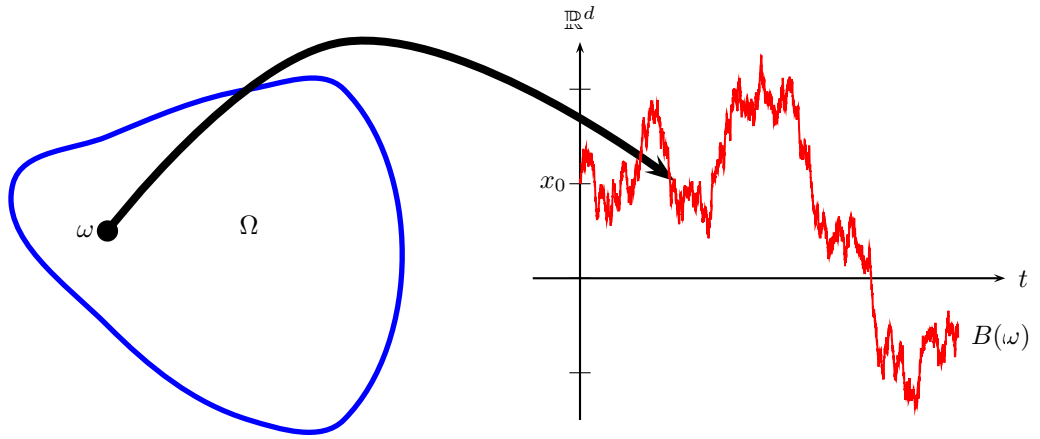
man die Bewegung der Hand des Spielers genau beschreiben könnte. Da diese Bewegung zu kompliziert ist, verwendet man ein elementares stochastisches Modell, das in der Regel die Folge der Augenzahlen sehr gut beschreibt.

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.



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.
- 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 innermathematisch 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

0 0 0 1 0 1 1 0 0 1 0 1 0 0 0 1 1 0 1 0 0 0 0 1 1 0 1 1 0 1
0 0 1 0 0 0 1 1 0 0 0 1 0 0 0 1 1 0 0 1 1 1 0 0 0 1 1 0 1 0
0 1 1 1 0 0 1 1 0 0 1 0 1 1 1 0 0 0 0 1 0 1 1 0 0 1 0 1 1 0
0 0 1 1 0 1 0 0 1 1 1 0 1 1 0 1 0 1 1 0 1 0 1 1 0 1 0 0 1 0

pa

0 1 1 0 0 1 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
0 1 0 1 0 1 0 1 1 1 1 1 1 1 1 1 1 0 1 0 1 0 1 0 1 0 1 0 1 1
1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 1 0 1 0
1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 0 1 0 0 1 0 0 1 0

pb

1 1 1 1 0 1 0 1 0 1 0 0 0 0 0 0 1 0 1 0 1 0 1 0 0 0 0 0 0 1
0 1 0 1 0 1 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 1 0 1 0 1 0 0 0 1
1 0 0 1 1 0 0 0 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 0 1 1 1 1 1 0
0 0 1 0 1 1 1 1 0 1 0 1 0 0 0 0 1 0 1 0 1 1 1 1 0 1 0 1 0 0

ta

0 0 0 0 0 1 1 0 0 0 1 0 1 1 0 1 1 1 0 1 0 0 1 1 1 0 1 1 1 0
0 1 1 1 1 1 1 0 0 1 0 0 0 0 1 0 0 1 0 1 1 0 1 1 0 0 1 1 0 1
0 0 0 1 0 0 1 0 0 1 0 1 0 0 1 1 1 1 1 0 1 1 1 1 1 1 0 0 1 0
0 0 0 1 0 0 0 0 0 0 1 0 1 1 0 0 1 1 0 0 0 0 1 0 1 1 0 1 0 1

fa

1 0 1 0 1 0 1 0 0 0 1 0 0 1 1 1 0 1 0 1 0 0 1 0 1 1 1 0 0 1
0 1 1 1 0 0 0 0 1 0 0 1 0 1 1 0 0 1 0 1 0 1 0 1 1 1 0 1 0 1
0 1 0 1 0 0 0 1 0 1 0 1 1 0 1 0 1 0 1 1 1 0 0 0 0 1 1 1 0 1
0 1 0 0 0 1 1 0 0 1 1 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0

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:

$$\begin{aligned} &\text{Die Werte } x_1, x_2, \ldots \text{ sind Realisierungen einer Folge } X_1, X_2, \ldots \\ &\text{von unabhängigen, auf } \{0,1\} \text{ gleichverteilten Zufallsvariablen.} \end{aligned} \tag{0.0.1}$$

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 (0.0.1) 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 (0.0.1) 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 (0.0.1) annehmen.

Zusammenfassend ist (0.0.1) 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 (0.0.1) 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 (0.0.1) 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.

Kapitel 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

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

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

$$\text{»}\bigcup_{i\in I} A_i\text{ tritt ein«} \qquad \Leftrightarrow \qquad \text{»mindestens eines der Ereignisse } A_i\text{ tritt ein«}.$$

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. *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 (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 1.1 (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

$$P[A \cup B] = P[A] + P[B] \qquad \textbf{\textit{»endliche Additivität«}}.$$

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

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

Insbesondere folgt

$$\begin{aligned} P[A] &\leq P[B], && \textbf{\textit{»Monotonie«,}} \\ P[A^C] &= 1 - P[A], && \textbf{\textit{»Gegenereignis«,}} \\ P[A] &\leq 1. \end{aligned}$$

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

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

Das Einschluss-/Ausschlussprinzip werden wir auf eine elegantere Weise am Ende dieses Kapitels beweisen (siehe Satz 1.8).

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 1.2. *(i) Sei $0 \le p(\omega) \le 1$, $\sum_{\omega \in \Omega} p(\omega) = 1$, 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. *Die Funktion $p\colon \Omega \to [0,1]$ 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

$$\sum_{\omega \in A} p(\omega) \;:=\; \sum_{i=1}^{\infty} p(\omega_i), \tag{1.1.1}$$

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 (1.1.1) wohldefiniert ist:

Lemma 1.3. *(i) Unabhängig von der gewählten Abzählung gilt*

$$\sum_{\omega \in A} p(\omega) \;=\; \sup_{\substack{F \subseteq A \\ |F| < \infty}} \sum_{\omega \in F} p(\omega). \tag{1.1.2}$$

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

$$\sum_{\omega \in A} p(\omega) \leq \sum_{\omega \in B} p(\omega). \tag{1.1.3}$$

(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 $p(\omega_i) \geq 0$ 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 »$\leq$«, und anschließend »$\geq$«:

»$\leq$«: 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 »$\leq$«.

»$\geq$«: 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 »$\geq$«.

(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 »$\leq$« und »$\geq$« separat:

»$\leq$«: 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 (1.1.3):

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

Also folgt nach (i) auch:

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

»$\geq$«: Seien $F_i \subseteq A_i$ endlich. Da die F_i wieder disjunkt sind, folgt mit σ-Additivität und Gleichung (1.1.3) 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 von Satz 1.2. (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 1.3.(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

$$P[A] \;=\; \sum_{\omega \in A} \frac{1}{|\Omega|} \;=\; \frac{|A|}{|\Omega|} \;=\; \frac{\text{Anzahl »günstiger« Fälle}}{\text{Anzahl aller Fälle}}. \tag{1.1.4}$$

Die Verteilung P heißt **Gleichverteilung** auf Ω und wird auch mit $\mathrm{Unif}(\Omega)$ bezeichnet. Laplace (1814) benutzte (1.1.4) 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 $\omega\colon \{1,2,\ldots,n\} \to \{1,2,\ldots,n\}$. Der 1 können n verschiedene Zahlen zugeordnet geordnet werden, der 2 die verbleibenden $n-1$, usw. Somit gibt es insgesamt $n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 1$ 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 1.8) berechnen:

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

Hierbei haben wir benutzt, dass es $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ k-elementige Teilmengen $\{i_1, \ldots, i_k\}$ von $\{1, \ldots, n\}$ 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

$N_k[A] := |\{i \in \{1, \ldots, k\} \mid x_i \in A\}|$ die Häufigkeit der Werte in A unter $x_1, \ldots, x_k$, und

$P_k[A] := N_k[A]/k$, die entsprechende relative Häufigkeit von Werten in A.

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

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

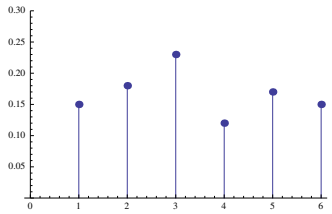
durch die relativen Häufigkeit der möglichen Merkmalsausprägungen unter $x_1, \ldots, x_k$ gegeben ist. Die Wahrscheinlichkeitsverteilung P_k heißt **empirische Verteilung** der Werte $x_1, \ldots, x_k$. In der beschreibenden Statistik analysiert man empirische Verteilungen mithilfe verschiedener Kenngrößen.

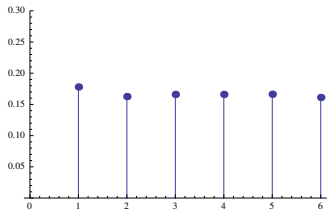
Beispiele. a) ABZÄHLUNG ALLER MÖGLICHEN FÄLLE:

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 $\{1, 2, 3, 4, 5, 6\}$:

```
x=RandomChoice[{1,2,3,4,5,6},n];
ListPlot[BinCounts[x[[1 ;; n], {1, 7, 1}]/n,
 Filling -> Axis, PlotRange -> {0, 0.3},
 PlotStyle -> PointSize[Large]], {n, 1, 100, 1}
```

$n = 100$: 

$n = 10000$: 

Das **empirische Gesetz der großen Zahlen** besagt, dass sich die empirischen Verteilungen für $k \to \infty$ der zugrundeliegenden Wahrscheinlichkeitsverteilung P (hier der Gleichverteilung auf $\{1, 2, \ldots, 6\}$) annähern:

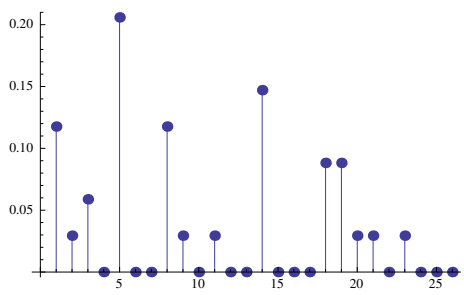
$$P_k[A] = \frac{|\{i \in \{1, \ldots, k\} \mid x_i \in A\}|}{k} \to P[A] \qquad \text{für } k \to \infty.$$

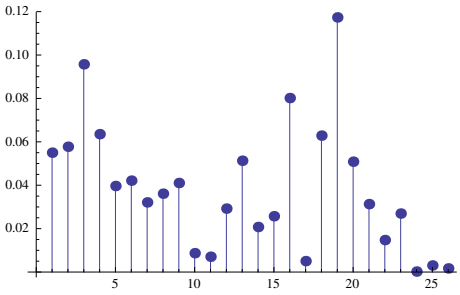
Diese Aussage wird auch als frequentistische „Definition" der Wahrscheinlichkeit von A in den empirischen Wissenschaften verwendet. Wir werden die Konvergenz der empirischen Verteilungen von unabhängigen, identisch verteilten Zufallsvariablen unten aus den Kolmogorovschen Axiomen herleiten.

c) EMPIRISCHE VERTEILUNG DER BUCHSTABEN »A« BIS »Z« IN DEM WORT »EISENBAHNSCHRANKENWAERTERHAEUSCHEN« UND IN EINEM ENGLISCHEN WÖRTERBUCH:

```
freq = StringCount["eisenbahnschrankenwaerterhaeuschen", #] & /@
   CharacterRange["a", "z"];
relfreq = freq/Total[freq];
ListPlot[relfreq, Filling -> Axis, PlotStyle -> PointSize[Large]]
```

```
freq = Length[DictionaryLookup[# ~~ ___]] & /@
   CharacterRange["a", "z"];
relfreq = freq/Total[freq]; ListPlot[relfreq, Filling -> Axis,
    PlotStyle -> PointSize[Large]]
```



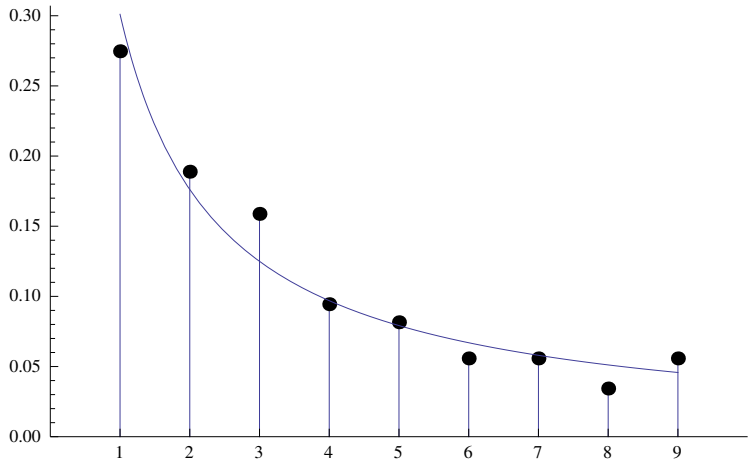


d) BENFORDSCHES GESETZ:

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. Das Benfordsche Gesetz lässt sich mithilfe des empirischen Gesetzes der großen Zahlen herleiten, wenn man annimmt, dass die Daten Realisierungen unabhängiger identisch verteilter Zufallsvariablen mit auf $[0,1)$ gleichverteilten logarithmierten Mantissen sind.



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

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

so dass für alle $a \in S$ gilt:

$$X^{-1}(a) \;=\; \{\omega \in \Omega \mid X(\omega) = a\} \;\in\; \mathcal{A}. \tag{1.2.1}$$

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 1.4. *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

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

□

Definition. *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 1.4 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 $k \leq n$ und $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ 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 1.5 (Poissonapproximation der Binomialverteilung)**.** *Sei $\lambda \in (0, \infty)$. Dann gilt:*

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

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

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

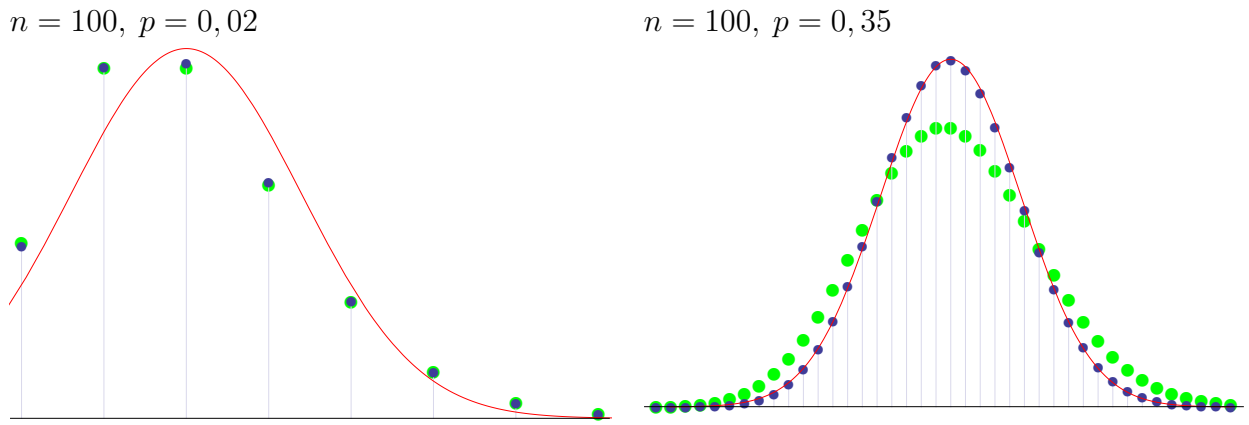
□

Definition. *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:



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 $\binom{m}{n}$ Ele-

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

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

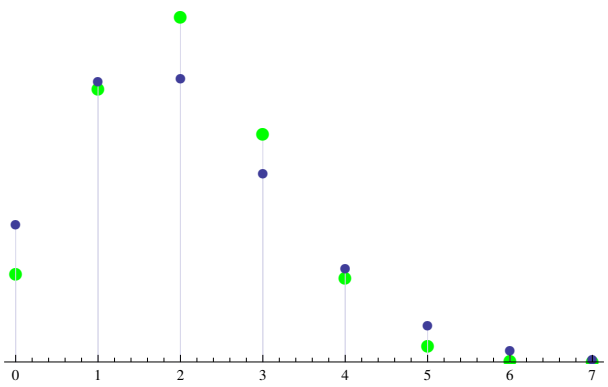
wird ***hypergeometrische Verteilung mit Parametern*** m***,*** r ***und*** n *genannt.*

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 $m,r\to\infty$ mit $p=r/m$ fest gilt

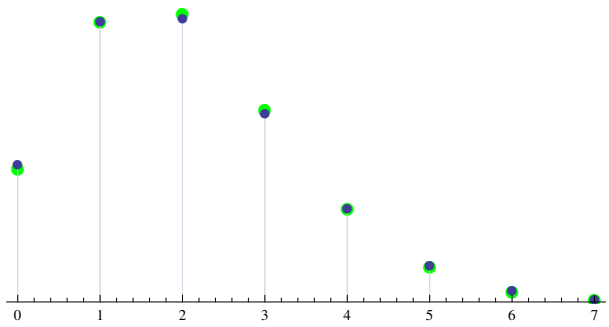
$$h_{m,r,n}(k) \;\longrightarrow\; \binom{n}{k}\,p^k\,(1-p)^k,$$

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. *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 $X : \Omega \to S$ 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 1.6 (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 $X : \Omega \to \mathbb{R}$ 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 1.7 (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:*

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

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

$$\begin{aligned} E[\lambda\, X + \mu\, Y] \;&=\; E[g(X,Y)] \;=\; \sum_{(a,b)\in S_X\times S_Y} g(a,b)\, P[(X,Y)=(a,b)] \qquad (1.3.1)\\ &= \sum_{a\in S_X}\sum_{b\in S_Y} (\lambda\, a + \mu\, b)\, P[X=a, Y=b] \\ &= \lambda \sum_{a\in S_X} a \sum_{b\in S_Y} P[X=a,Y=b] + \mu \sum_{b\in S_Y} b \sum_{a\in S_X} P[X=a,Y=b] \\ &= \lambda \sum_{a\in S_X} a\, P[X=a] + \mu \sum_{b\in S_Y} b\, P[Y=b] \\ &= \lambda\, E[X] + \mu\, E[Y]. \end{aligned}$$

Hierbei haben wir benutzt, dass die Reihe in (1.3.1) absolut konvergiert, da nach einer analogen Rechnung

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

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 (Monotonie des Erwartungswerts)**.** *Seien die Voraussetzungen von Satz 1.7 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 $\omega\colon \{1,\ldots,n\} \to \{1,\ldots,n\}$, 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

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

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 $\{\omega(1), \omega(2), \ldots, \omega(n)\}$ 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 1.8 (Einschluss-/Ausschlussprinzip)**.** *Für $n \in \mathbb{N}$ und Ereignisse $A_1, \ldots, A_n \in \mathcal{A}$ gilt:*

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

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

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

Damit folgt

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

□

Kapitel 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. *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 $P[B] \neq 0$ 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 $1/2$:

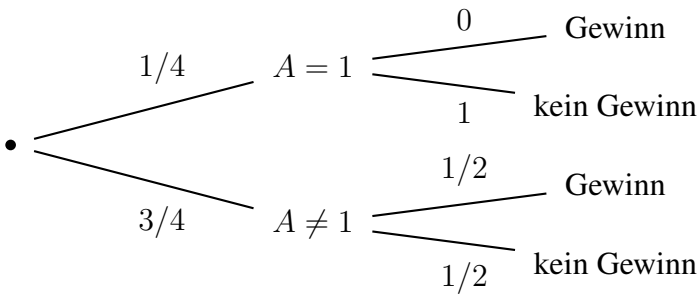
$$P[\text{»beide Mädchen«} \mid \text{»das erste ist Mädchen«}] \;=\; \frac{|\{MM\}|}{|\{MM, MJ\}|} \;=\; \frac{1}{2}.$$

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:

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

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

$$P[A] \;=\; \sum_{\substack{i \in I \\ P[H_i] \neq 0}} P[A|H_i] \cdot P[H_i] \tag{2.1.1}$$

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 2.1:

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

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

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

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 2.1 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:

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

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

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:

$$P[X_1 = x_1] = p_1(x_1) \qquad \text{für alle } x_1 \in S_1, \qquad \text{sowie} \tag{2.2.1}$$

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

$$P[X_k = x_k \mid X_1 = x_1, \ldots, X_{k-1} = x_{k-1}] = p_k(x_k \mid x_1, \ldots, x_{k-1}) \tag{2.2.2}$$

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

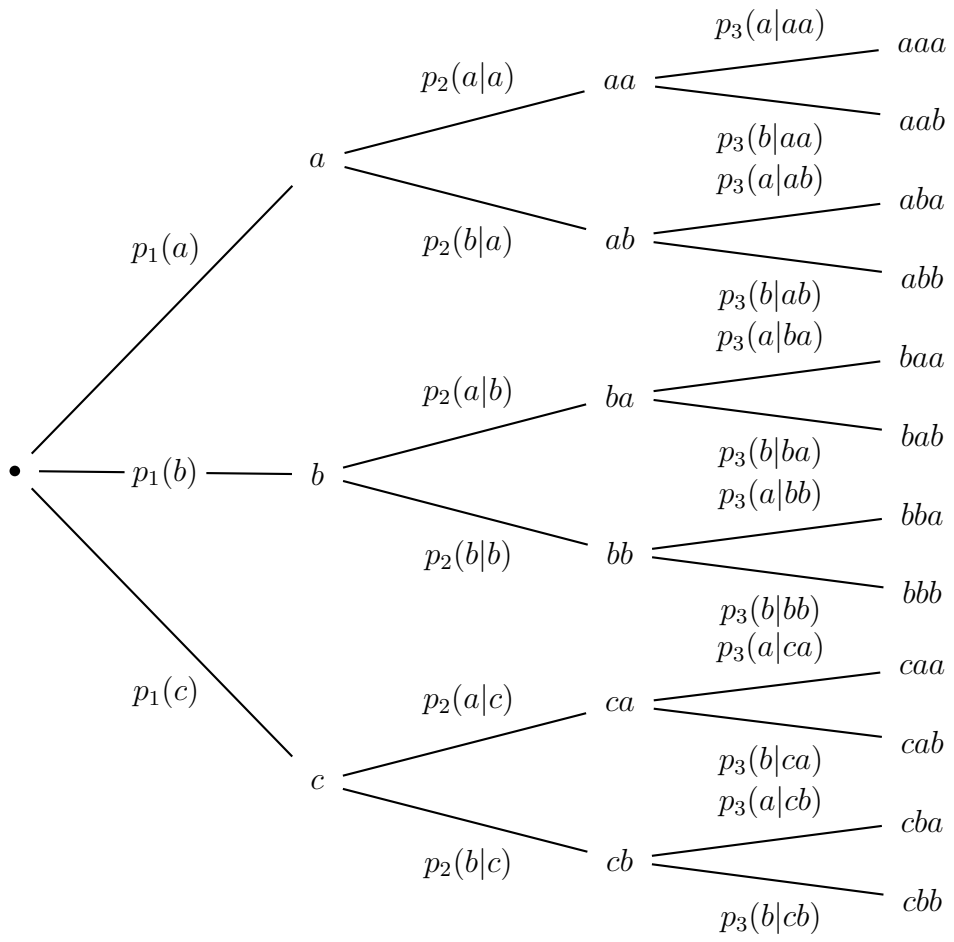


Abbildung 2.1: Dreistufiges Modell mit $S_1 = \{a, b, c\}$ und $S_2 = S_3 = \{a, b\}$.

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 (2.2.1) und (2.2.2) 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 2.2 (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* (2.2.1) *und* (2.2.2)*. 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 (2.2.1) und (2.2.2) und $k = 1, \ldots n$ gilt:

$$P[X_1 = x_1, \ldots, X_k = x_k] \;=\; p_1(x_1) \cdot p_2(x_2 \mid x_1) \cdots p_k(x_k \mid x_1, \ldots, x_{k-1}). \tag{2.2.3}$$

Nach (2.2.1) ist dies für $k = 1$ der Fall. Zudem folgt aus (2.2.3) für $k - 1$ nach (2.2.2):

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

also die Behauptung (2.2.3) für k, falls $P[X_1 = x_1, \ldots, X_{k-1} = x_{k-1}] \neq 0$. Andernfalls verschwinden beide Seiten in (2.2.3) 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 (2.2.1) und (2.2.2) 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

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

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 $x_1, \ldots, x_{i-1}$ ab, dann gilt

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

mit einer von $x_1, \ldots, x_{i-1}$ 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

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_i(x_i), \qquad x \in S_1 \times \cdot \times S_n. \tag{2.2.4}$$

Definition. *Seien P_i, $i = 1, \ldots n$, Wahrscheinlichkeitsverteilungen auf abzählbaren Mengen S_i mit Massenfunktionen p_i. Die durch die Massenfunktion* (2.2.4) *bestimmte Wahrscheinlichkeitsverteilung $P = P_1 \otimes \ldots \otimes P_n$ auf $\Omega = S_1 \times \ldots \times S_n$ heißt* ***Produkt*** *von $P_1, \ldots, P_n$.*

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

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

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

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

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

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

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

Satz 2.3. *Bezüglich des Produkts $P = P_1 \otimes \cdots \otimes P_n$ gilt für beliebige Ereignisse $A_i \subseteq S_i, i = 1, \ldots, n$:*

$$\begin{array}{rcl} P[X_1 \in A_1, \ldots, X_n \in A_n] & = & \displaystyle\prod_{i=1}^n P[X_i \in A_i] \\ \| & & \| \\ P[A_1 \times \ldots \times A_n] & & \displaystyle\prod_{i=1}^n P_i[A_i] \end{array} \tag{2.2.5}$$

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 2.3 besagt, dass die Koordinatenabbildungen $X_i(\omega) = \omega_i$ im Produktmodell *unabhängige Zufallsvariablen* sind, siehe Abschnitt 2.3.

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

$$p_k(x_k \mid x_0, \ldots, x_{k-1}) \;=\; p_k(x_{k-1}, x_k), \tag{2.2.6}$$

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. *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.*

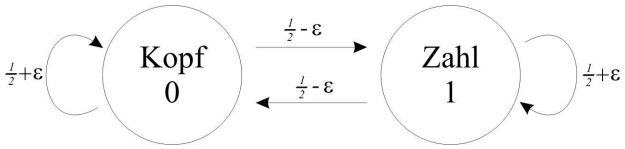
Sei $\nu : S \to [0,1]$ die Massenfunktion der Verteilung von X_0, also der **Startverteilung** der zufälligen Entwicklung. Als Massenfunktion des mehrstufigen Modells ergibt sich dann aus Gleichung (2.2.6):

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

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.

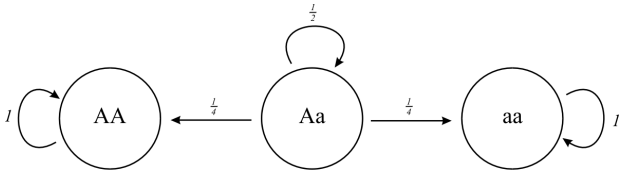
b) ABHÄNGIGE MÜNZWÜRFE: Ein einfaches Modell für abhängige Münzwürfe ist eine Markovkette mit Zustandsraum $S = \{0, 1\}$ und den folgenden Übergangswahrscheinlichkeiten:



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



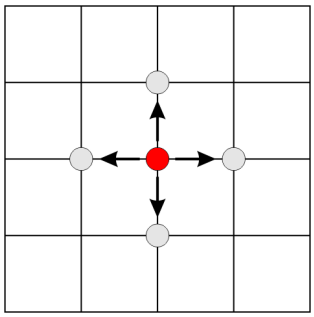
gegeben, und die Übergangsmatrix einer entsprechenden Markovkette ist

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

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

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

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 2.4 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 $x_0, x_1, \ldots, x_{k-1}$.

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 3.2 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

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

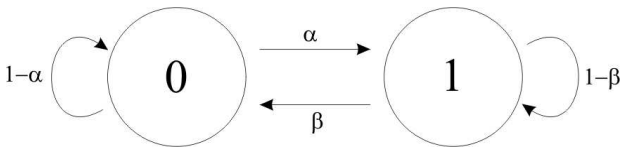
Für die unbedingten Wahrscheinlichkeiten ergibt sich

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

□

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

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

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

$$P[A \cap B] \;=\; P[A] \cdot P[B], \tag{2.3.1}$$

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. *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 2.5 (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 2.5 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. *Sei $p \in [0,1]$. Die Wahrscheinlichkeitsverteilung μ auf $\mathbb{N} \cup \{\infty\}$ mit Massenfunktion*

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

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

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

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

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 2.5 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.*

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. *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 $I = \{1, \dots, n\}$ mit $n \in \mathbb{N}$. Sind $X_1 : \Omega \to S_1, \dots, X_n : \Omega \to S_n$ diskrete Zufallsvariablen, dann ist auch $(X_1, \dots, X_n)$ eine diskrete Zufallsvariable mit Werten im Produktraum $S_1 \times \cdots \times S_n$.

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

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

$$p_{X_1,\dots,X_n}(a_1, \dots, a_n) \;=\; P[X_1 = a_1, \dots, X_n = a_n] \tag{2.3.2}$$

Sie enthält Informationen über den Zusammenhang zwischen den Zufallsgrößen X_i.

Satz 2.6. *Die folgenden Aussagen sind äquivalent:*

(i) $X_1, \dots, X_n$ *sind unabhängig.*

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

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

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

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

$(ii) \Rightarrow (iii)$ gilt nach (2.3.2).

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

$(iv) \Rightarrow (i)$: Seien $A_i \subseteq S_i$ $(i = 1, \ldots, n)$ und $1 \leq i_1 < i_2 < \ldots < i_k \leq n$. 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 2.3:

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

□

Als Konsequenz aus Satz 2.6 ergibt sich insbesondere:

Korollar. *Sind $X_i : \Omega \to S_i$, $1 \leq i \leq n$, diskrete Zufallsvariablen, und hat die gemeinsame Massenfunktion eine Darstellung in Produktform*

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

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

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

Beweis. Die Werte

$$\widetilde{g}_i(a) := \frac{g_i(a)}{\sum\limits_{b \in S_i} g_i(b)}, \qquad a \in S_i,$$

sind die Gewichte eine Wahrscheinlichkeitsverteilung μ_i auf S_i. Nach Voraussetzung gilt für $(a_1, \ldots, a_n) \in S_1 \times \ldots \times S_n$:

$$\mu_{X_1,\ldots,X_n}[\{a_1\} \times \ldots \times \{a_n\}] = p_{X_1,\ldots,X_n}(a_1, \ldots, a_n) = \widetilde{c} \cdot \prod_{i=1}^{n} \mu_i[\{a_i\}] \tag{2.3.3}$$

mit einer reellen Konstante $\widetilde{c}$. Da auf beiden Seiten von (2.3.3) bis auf den Faktor $\widetilde{c}$ die Massenfunktionen von Wahrscheinlichkeitsverteilungen stehen, gilt $\widetilde{c} = 1$, und damit

$$\mu_{X_1,\ldots,X_n} = \bigotimes_{i=1}^{n} \mu_i.$$

Also sind die X_i unabhängige Zufallsvariablen mit Verteilung μ_i, d.h. mit Massenfunktion $\widetilde{g}_i$. □

Beispiel (**Zwei 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 beispielsweise folgende Möglichkeiten:

(1). X, Y unabhängig.

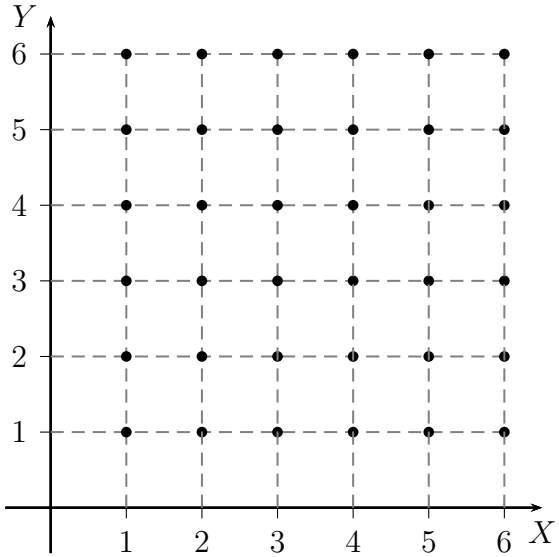


Abbildung 2.2: X, Y unabhängig; $\mu_{X,Y} = \mu_X \otimes \mu_Y$. Gewichte der Punkte sind jeweils $\frac{1}{36}$.

(2). X, Y deterministisch korreliert, z.B. $Y = (X + 1) \mod 6$.

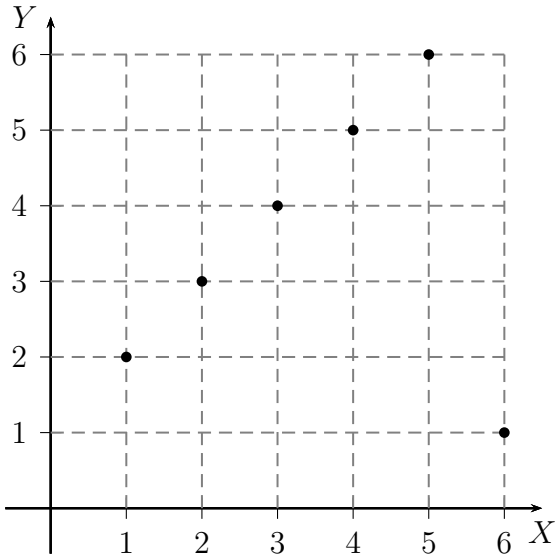


Abbildung 2.3: $Y = (X + 1) \mod 6$. Das Gewicht eines einzelnen Punktes ist $\frac{1}{6}$.

(3). $Y = (X + Z) \mod 6$, Z unabhängig von X, $Z = 0, \pm 1$ mit Wahrscheinlichkeit $\frac{1}{3}$.

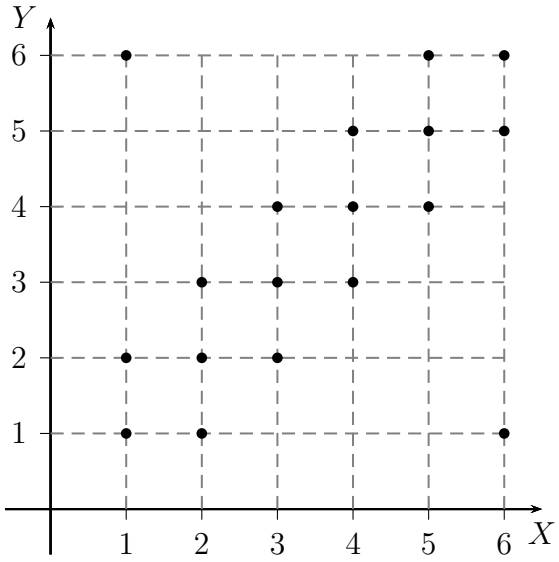


Abbildung 2.4: $Y = (X+Z) \mod 6$; $Z \sim \text{unif}\{-1, 0, 1\}$. Das Gewicht eines einzelnen Punktes ist $\frac{1}{18}$.

Random Walks 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

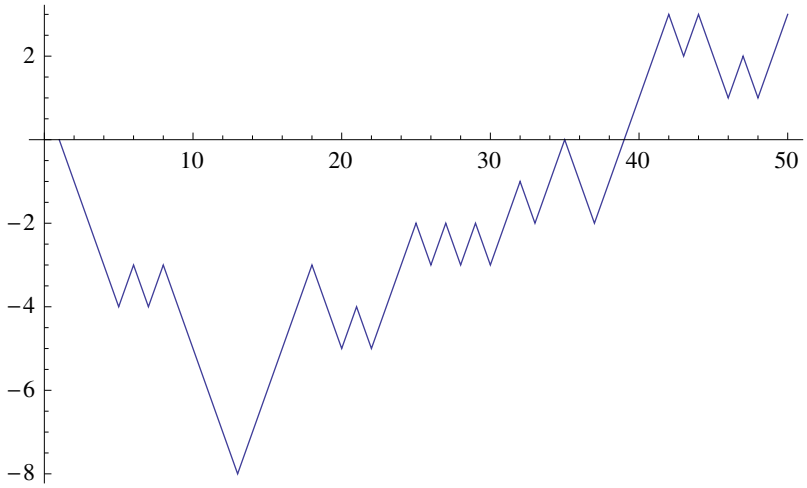
$$S_n = a + X_1 + X_2 + \cdots + X_n.$$

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.

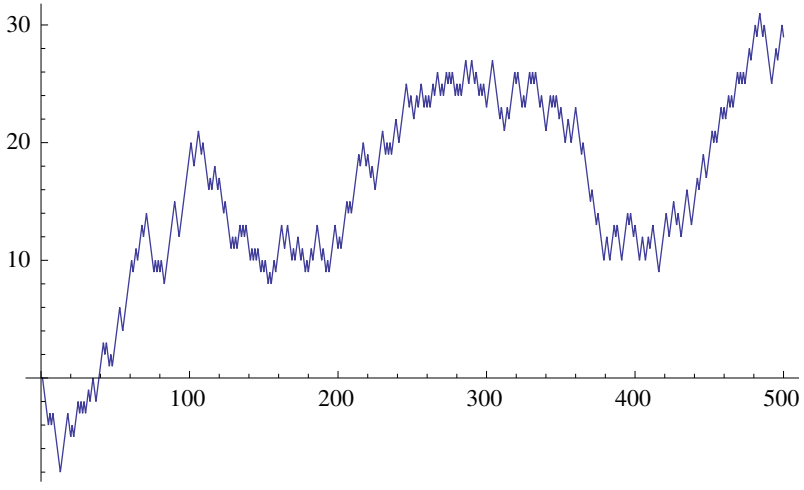
Beispiel (Symmetrischer Random Walk, $p = 1/2$)**.** Die folgende Mathematica-Routine simuliert 10.000 Schritte eines Random Walks für $p = 1/2$, und plottet den Verlauf der ersten $nmax$ Schritte.

```
zufall = RandomChoice[{-1, 1}, 10000];
randomwalk = FoldList[Plus, 0, zufall];
Manipulate[
 ListLinePlot[randomwalk[[1 ;; nmax]]], {nmax, 10, 10000, 10}]
```

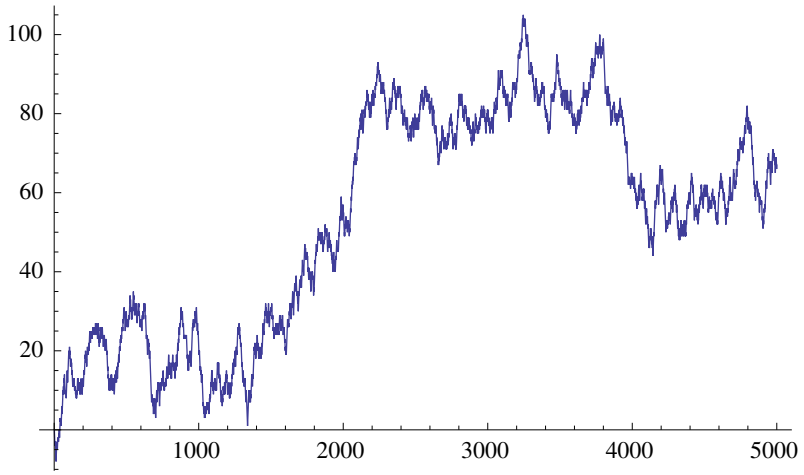
- $nmax = 50$:



- $nmax = 500$:



- $nmax = 5000$:



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

Lemma 2.7 (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 der Random Walk 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 des Random Walks 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 2.8. *Für* $(s_0, s_1, \ldots, s_n) \in \widehat{\Omega}_a^{(n)}$ *gilt*

$$\mu_a[\{(s_0, \ldots, s_n)\}] \;=\; p^{\frac{n+k}{2}} (1-p)^{\frac{n-k}{2}}, \qquad \textit{wobei } \; k = s_n - s_0. \tag{2.3.4}$$

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 (2.3.4), da für $s_n - s_0 = k$ genau k der Inkremente $s_1 - s_0, \ldots, s_n - s_{n-1}$ gleich $+1$ und die übrigen gleich -1 sind. □

Symmetrischer Random Walk und Reflektionsprinzip

Ab jetzt betrachten wir nur noch die symmetrische Irrfahrt mit $p = \frac{1}{2}$. Lemma 2.8 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 2.3 klar.

Satz 2.9 (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.*

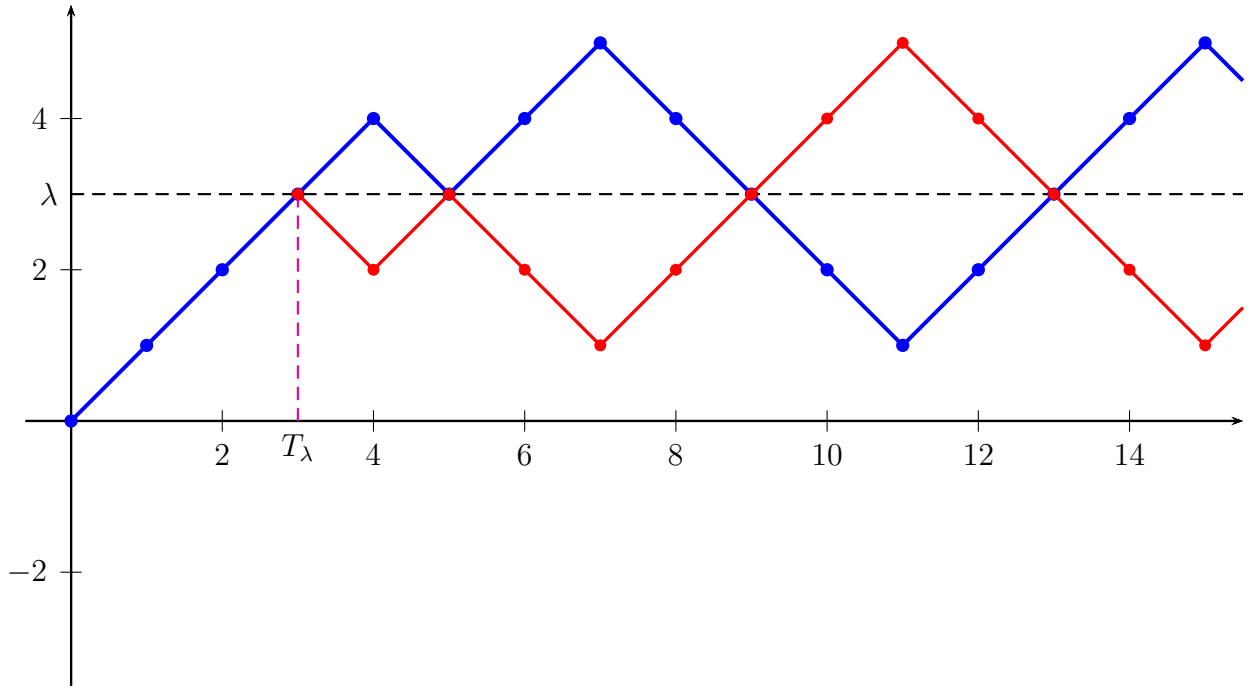


Abbildung 2.5: Reflektionsprinzip

Beweis. Es gilt

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

Die im Bild 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

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

□

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

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

(i)

$$P[T_\lambda \le n] = \begin{cases} P[S_n \ge \lambda] + P[S_n > \lambda] & \textit{falls } \lambda > a, \\ P[S_n \le \lambda] + P[S_n < \lambda] & \textit{falls } \lambda < a. \end{cases}$$

(ii)

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

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 2.9:

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

(ii) Aus (i) folgt

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

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:

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

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

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

□

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

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

des Random Walks bis zur Zeit n.

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

$$P[M_n \geq \lambda] \ = \ P[T_\lambda \leq n] \ = \ P[S_n \geq \lambda] + P[S_n > \lambda].$$

Kapitel 3

Konvergenzsätze für Zufallsvariablen und ihre Verteilungen

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 $A_1, \ldots, A_n$ 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 $A_1, A_2, \ldots$ 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 $A_1, \ldots, A_n$, die eintreten.

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

$$P\left[\frac{S_n}{n} \geq p + \varepsilon\right] \leq e^{-2\varepsilon^2 n}, \qquad \textit{und} \qquad P\left[\frac{S_n}{n} \leq p - \varepsilon\right] \leq e^{-2\varepsilon^2 n}.$$

Insbesondere ist

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

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

$$p\, e^{\lambda q} + q\, e^{-\lambda p} \;\leq\; e^{\lambda^2/8} \tag{3.1.1}$$

gilt. Damit erhalten wir dann

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

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 (3.1.1). Sei dazu

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

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

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

□

Zur Illustration des Satzes simulieren wir den Verlauf von S_k und S_k/k für $k \leq n$ und $p = 0.7$ mehrfach (m-mal), und plotten die Massenfunktionen von S_n.

VERLAUF VON S_k FÜR $k \leq n$

```
m = 30; nmax = 1000; p = 0.7;
(Wir erzeugen m × nmax Bernoulli-Stichproben mit Wahrscheinlichkeit p)
x = RandomChoice[{1 - p, p} -> {0, 1}, {nmax, m}]; s = Accumulate[x];
Das Feld s enthält m Verläufe von s_n = x_1 + ... + x_n, n = 1, ..., nmax
Manipulate[Show[
  ListLinePlot[Transpose[s[[1 ;; n]]]],
  ListLinePlot[p*Range[n], PlotStyle -> {Black, Thick}]]
  , {{n, 50}, 1, nmax, 1}]
(Vergleich der m Verläufe von s_n mit np)
```

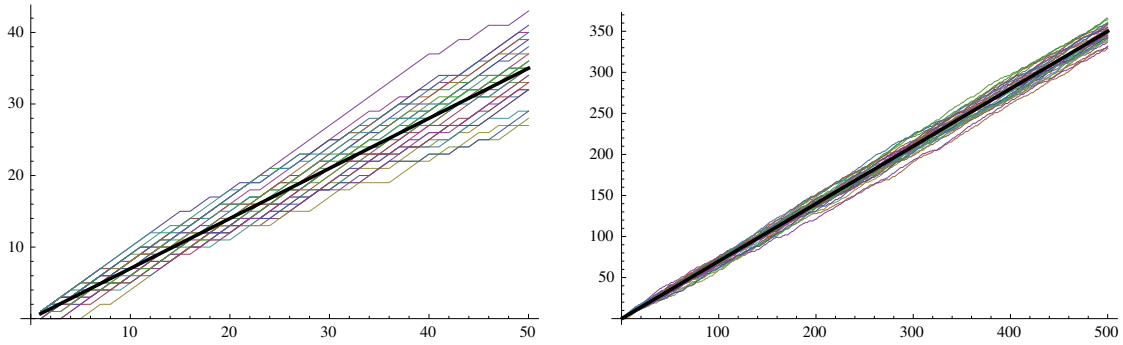


Abbildung 3.1: Verlauf von S_k für $k \leq 50$ und $k \leq 500$

VERLAUF VON S_k/k FÜR $k \leq n$

```
mean = s / Range[nmax];
(Das Feld mean enthält m Verläufe der Werte von s_n / n)
Manipulate[Show[
  ListLinePlot[Transpose[mean[[1 ;; n]]]],
  ListLinePlot[ConstantArray[p, n], PlotStyle -> {Black, Thick}] ], {{n,
     50}, 1, nmax, 1}]
```

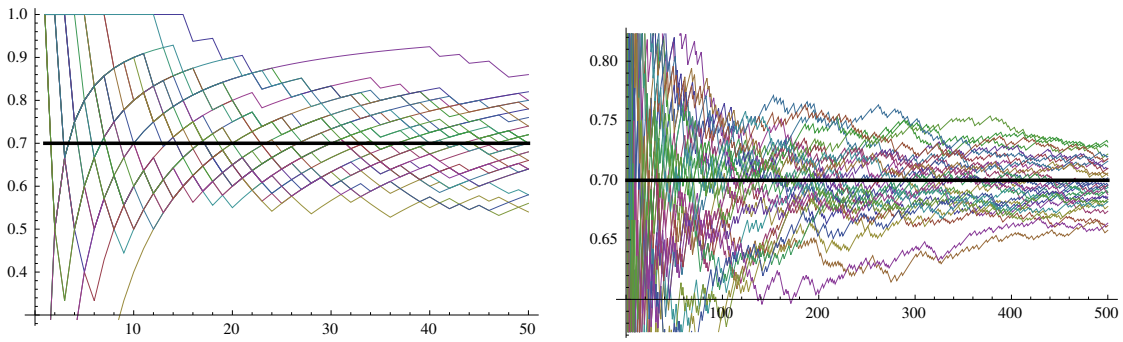


Abbildung 3.2: Verlauf von S_k/k für $k \leq 50$ und $k \leq 500$

VERTEILUNG VON S_n

```
Manipulate[
 ListPlot[Table[{k, PDF[BinomialDistribution[n, p], k]}, {k, 0, n}],
  PlotRange -> All, Filling -> Axis]
 , {{n, 50}, 1, nmax, 1}]
```

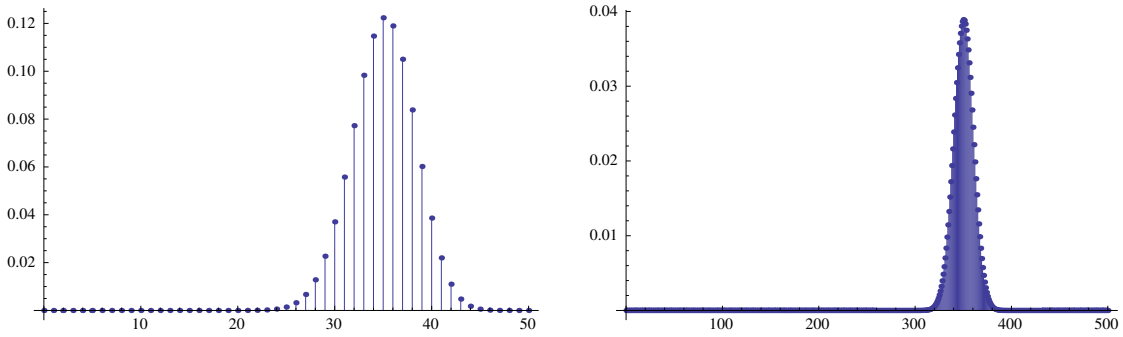


Abbildung 3.3: Verteilung von S_n für $n = 50$ und $n = 500$

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

$$\left\{\lim_{n\to\infty} \frac{S_n}{n} = p\right\} \ = \ \bigcap_{k=1}^{\infty} \bigcup_{n_0=1}^{\infty} \bigcap_{n=n_0}^{\infty} \left\{\left|\frac{S_n}{n} - p\right| \leq \frac{1}{k}\right\} \ \in \ \mathcal{A}. \tag{3.1.2}$$

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

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

Beweis. Wir zeigen mithilfe der Bernstein-Ungleichung, dass das Gegenereignis $\{S_n/n \not\to p\}$ Wahrscheinlichkeit Null hat. Nach (3.1.2) 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 $n_0 \in \mathbb{N}$:

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

Für $n_0 \to \infty$ konvergieren die Partialsummen auf der rechten Seite gegen Null. Also folgt $P[A_k] = 0$, und damit die Behauptung. □

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.

3.2 Konvergenz ins Gleichgewicht 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. *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 2.4 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. *i) Eine Wahrscheinlichkeitsverteilung* μ *auf* S *heißt* ***Gleichgewichtsverteilung*** *(oder* ***invariante Verteilung****) der Übergangsmatrix* π*, falls* $\mu\,\pi = \mu$ *gilt, d.h. falls*

$$\sum_{x\in S} \mu(x)\,\pi(x, y) = \mu(y) \qquad \textit{für alle } y \in S.$$

ii) μ *erfüllt die* ***Detailed Balance-Bedingung*** *bzgl. der Übergangsmatrix* π*, falls gilt:*

$$\mu(x)\,\pi(x, y) = \mu(y)\,\pi(y, x) \qquad \textit{für alle } x, y \in S \tag{3.2.1}$$

Satz 3.2. *Erfüllt* μ *die Detailed Balance-Bedingung* (3.2.1)*, dann ist* μ *eine Gleichgewichtsverteilung von* π.

Beweis. Aus der Detailed Balance-Bedingung folgt

$$\sum_{x\in S} \mu(x)\,\pi(x, y) \;=\; \sum_{x\in S} \mu(y)\,\pi(y, x) \;=\; \mu(y) \qquad \text{für alle } y \in S.$$

□

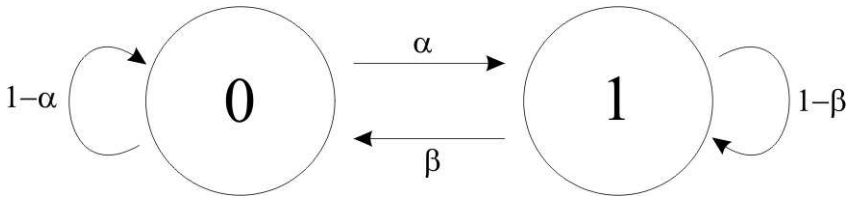
Bemerkung. Bei Startverteilung μ gilt:

$$\mu(x)\,\pi(x,y) = P[X_0 = x, X_1 = y].$$

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 (3.2.1). 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

$$\pi(k, k+1) = p, \qquad \pi(k, k-1) = 1-p.$$

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

$$\mu(x) \;=\; \frac{\deg(x)}{2|E|}.$$

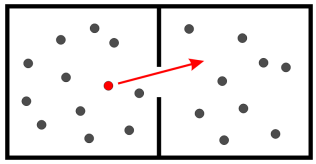
Alternativ können wir einen modifizierten Random Walk definieren, der die Gleichverteilung auf V als Gleichgewicht hat. Sei dazu $\triangle := \max_{x \in V} \deg(x)$ der maximale Grad, und

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

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.

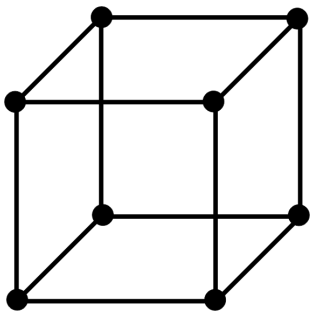
MIKROSKOPISCHE BESCHREIBUNG: Ein detailliertes Modell ergibt sich, wenn wir für jede einzelne Kugel notieren, ob sich diese in der ersten Urne befindet. Der Zustandsraum ist dann

$$S = \{0,1\}^n = \{(\sigma_1, \ldots, \sigma_n) \ : \ \sigma_i \in \{0,1\} \ \forall\, i\},$$

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

$$\mu(x-1)\,\pi(x-1,x) \;=\; \mu(x)\,\pi(x,x-1) \qquad \text{für } x = 1,\dots,n,$$

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 die 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. *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

$$\pi^r(x,y) \ge \delta\cdot\mu(y) \qquad \text{für alle } x,y\in S \text{ gilt.} \tag{3.2.2}$$

Satz 3.3 (Konvergenzsatz von W. Doeblin)**.** *Gilt die Minorisierungsbedingung* (3.2.2)*, 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 (3.2.2) 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 $\nu \in \mathrm{WV}(S)$:

$$\nu\,\pi^r = (1-\lambda)\,\mu + \lambda\,\nu\,q. \tag{3.2.3}$$

2. Wir zeigen mit vollständiger Induktion:

$$\nu\,\pi^{kr} = (1-\lambda^k)\,\mu + \lambda^k\,\nu\,q^k \qquad \text{für alle } k \geq 0, \quad \nu \in \mathrm{WV}(S). \tag{3.2.4}$$

Für $k = 0$ ist die Aussage offensichtlich wahr. Gilt (3.2.4) für ein $k \geq 0$, dann erhalten wir durch Anwenden von Gleichung (3.2.3) 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 $\nu \in \mathrm{WV}(S)$:

$$\begin{aligned}
\nu\,\pi^n &= \nu\,\pi^{kr}\,\pi^i = (1-\lambda^k)\,\underbrace{\mu\,\pi^i}_{=\mu} + \lambda^k\,\nu\,q^k\,\pi^i, \qquad \text{also} \\
\nu\,\pi^n - \mu &= \lambda^k\,(\nu\,q^k\,\pi^i - \mu), \qquad \text{und damit} \\
d_{TV}(\nu\,\pi^n, \mu) &= \frac{1}{2}\,\|\nu\,\pi^n - \mu\|_1 = \lambda^k\,d_{TV}(\nu\,q^k\,\pi^i, \mu) \leq \lambda^k.
\end{aligned}$$

□

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. *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

$$\text{Periode}(x) \;:=\; \text{ggT}\,\{n \in \mathbb{N} \mid \pi^n(x,x) > 0\}.$$

Man kann dann zeigen, dass x genau dann aperiodisch ist, wenn $\text{Periode}(x) = 1$ 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 (Konvergenzsatz für endliche Markov-Ketten)**.** *Ist der Zustandsraum S endlich, die Übergangsmatrix π irreduzibel, und existiert ein aperiodischer Zustand $a \in S$, dann gilt:*

$$\lim_{n\to\infty} d_{TV}(\nu\,\pi^n, \mu) = 0 \qquad \textit{für alle } \nu \in \text{WV}(S).$$

Beweis. Wir zeigen, dass zu jedem $x, y \in S$ eine natürliche Zahl $k(x,y)$ existiert, so dass

$$\pi^n(x,y) > 0 \qquad \text{für alle } n \geq k(x,y) \tag{3.2.5}$$

gilt. Da der Zustandsraum endlich ist, folgt hieraus, dass die Minorisierungsbedingung (3.2.2) 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.

3.3 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. *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 $E[X^2] = E[X] = p$, 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 $E\left[|X|^p\right] < \infty$. 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 bzgl. P. Gilt $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$, dann ist 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 3.4 (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

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

□

Kovarianz und Korrelation

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

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

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

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

(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.*

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

*Gilt $\mathrm{Cov}[X, Y] > 0$ bzw. < 0, dann heißen X und Y **positiv** bzw. **negativ korreliert**.*

Um elementare Eigenschaften der Kovarianz herzuleiten, bemerken wir, dass der Raum $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ mit einem positiv semidefiniten Skalarprodukt versehen ist:

Lemma 3.5 (L^2 Skalarprodukt und Cauchy-Schwarz-Ungleichung)**.**

(i) Der Raum $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ ist ein Vektorraum.

(ii) Durch

$$(X, Y)_{\mathcal{L}^2} := E[X \cdot Y], \qquad X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P),$$

ist eine positiv semidefinite symmetrische Bilinearform auf $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ definiert.

(iii) Für $X, Y \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ gilt $X \cdot Y \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ und

$$E[X \cdot Y]^2 \leq E\left[X^2\right] \cdot E\left[Y^2\right].$$

Insbesondere gilt also für eine Zufallsvariable $X \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ auch

$$E[|X|] \leq \sqrt{E\left[X^2\right]}\sqrt{E\left[1^2\right]} < \infty,$$

d.h. der Raum $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ ist in $\mathcal{L}^1(\Omega, \mathcal{A}, P)$ enthalten.

Beweis. (i) Seien $X, Y \in \mathcal{L}^2$ und $a \in \mathbb{R}$. Dann ist $aX + Y$ eine Zufallsvariable, für die wegen der Monotonie und Linearität des Erwartungswerts gilt:

$$E\left[(aX + Y)^2\right] = E\left[a^2X^2 + 2aX\,Y + Y^2\right] \leq 2a^2E\left[X^2\right] + 2E\left[Y^2\right] < \infty.$$

(ii) Für $X, Y \in \mathcal{L}^2$ gilt wegen der Monotonie des Erwartungswerts:

$$E\left[|X \cdot Y|\right] \leq E\left[(X^2 + Y^2)/2\right] \leq \frac{1}{2}E\left[X^2\right] + \frac{1}{2}E\left[Y^2\right] < \infty.$$

Also ist $(X, Y)_{\mathcal{L}^2} = E[X\,Y]$ wohldefiniert. Die Abbildung $(X, Y)_{\mathcal{L}^2}$ ist zudem symmetrisch, bilinear, da $E[\,\bullet\,]$ linear ist, und positiv semidefinit wegen $(X, X)_{\mathcal{L}^2} = E\left[X^2\right] \geq 0$ für alle $X \in \mathcal{L}^2$.

(iii) Da $(X, Y)_{\mathcal{L}^2}$ nach (ii) eine positiv semidefinite symmetrische Bilinearform ist, gilt die Cauchy-Schwarz-Ungleichung

$$(X, Y)_{\mathcal{L}^2}^2 \leq (X, X)_{\mathcal{L}^2}\,(Y, Y)_{\mathcal{L}^2}.$$

□

Korollar (Cauchy-Schwarz-Ungleichung für Kovarianz)**.**

(i) Die Kovarianz ist eine symmetrische Bilinearform auf $\mathcal{L}^2 \times \mathcal{L}^2$ mit

$$\mathrm{Cov}[X, X] \;=\; \mathrm{Var}[X] \;\geq\; 0 \qquad \textit{für alle } X \in \mathcal{L}^2.$$

(ii) Es gilt die Cauchy-Schwarz-Ungleichung

$$|\mathrm{Cov}[X, Y]| \;\leq\; \sqrt{\mathrm{Var}[X]} \cdot \sqrt{\mathrm{Var}[Y]} \;=\; \sigma[X] \cdot \sigma[Y].$$

Beweis. Das Korollar folgt durch Anwenden von Lemma 3.5 auf die zentrierten Zufallsvariablen $\widetilde{X} = X - E[X]$ und $\widetilde{Y} = Y - E[Y]$. □

Aus der Cauchy-Schwarz-Ungleichung für die Kovarianz folgt, dass der Korrelationskoeffizient $\varrho[X, Y]$ stets Werte zwischen -1 und 1 annimmt.

Beispiel (Empirischer Korrelationskoeffizent). 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

$$\begin{aligned} E[X] \;&=\; \frac{1}{n}\sum_{i=1}^{n} x_i \;=:\; \overline{x}_n, \qquad E[Y] \;=\; \overline{y}_n, \\ \mathrm{Cov}[X, Y] \;&=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n) \;=\; \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x}_n \overline{y}_n. \end{aligned}$$

Der entsprechende **empirische Korrelationskoeffizient** der Daten (x_i, y_i), $1 \leq i \leq n$, ist

$$\varrho[X, Y] \;=\; \frac{\mathrm{Cov}[X, Y]}{\sigma[X]\sigma[Y]} \;=\; \frac{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\displaystyle\left(\sum_{i=1}^{n}(x_i - \overline{x}_n)^2\right)^{1/2}\left(\sum_{i=1}^{n}(y_i - \overline{y}_n)^2\right)^{1/2}}$$

Grafik 3.3 zeigt Datensätze mit verschiedenen Korrelationskoeffizienten ϱ.

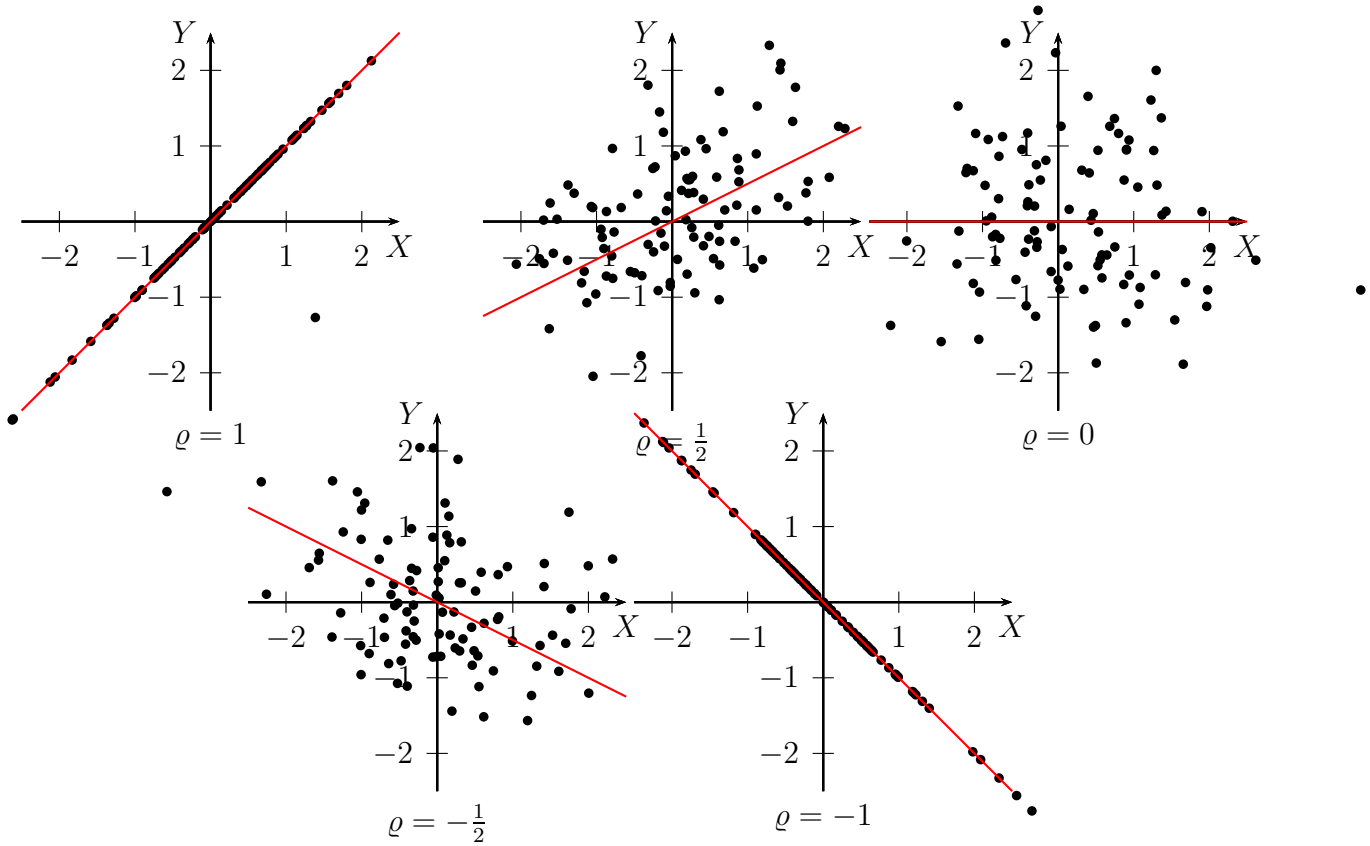


Abbildung 3.4: Empirische Korrelationskoeffizienten verschiedener Datensätze

Unabhängigkeit und Unkorreliertheit

Aus der Unabhängigkeit reellwertiger Zufallsvariablen in $\mathcal{L}^2$ folgt deren Unkorreliertheit. Allgemeiner gilt sogar:

Satz 3.6 (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 Definition der Unabhängigkeit ist Bedingung (i) äquivalent zu

$$P[X \in A, Y \in B] \;=\; P[X \in A]\, P[Y \in B] \qquad \text{für alle } A,\, B \in \mathcal{A}.$$

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

$$\begin{aligned} P[X=0, Y=0] &= 1/3 \;\neq\; 1/9 \;=\; P[X=0]\, P[Y=0], \\ E[X\,Y] &= 0 \;=\; E[X]\, E[Y]. \end{aligned}$$

3.4 GGZ 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.

$$\frac{S_n}{n} - \frac{E[S_n]}{n} \quad \stackrel{n\to\infty}{\longrightarrow} \quad 0. \tag{3.4.1}$$

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 $\dfrac{S_n}{n} = X_1$ 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 (3.4.1) 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 3.7. *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 3.7 für die Varianz der Binomialverteilung:

$$\mathrm{Var}[S_n] \;=\; \sum_{i=1}^n \mathrm{Var}[X_i] \;=\; n\,p\,(1-p).$$

Insbesondere ist die Standardabweichung einer 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

$$\sum_{n=0}^{\infty} c_n \;<\; \infty \qquad \text{und} \qquad \mathrm{Cov}[X_i, X_j] \;\le\; c_{|i-j|} \quad \text{für alle } i, j \in \mathbb{N}. \tag{3.4.2}$$

Die Annahme ist z.B. immer erfüllt, wenn die beiden folgenden Bedingungen erfüllt sind:

(i) Die Zufallsvariablen sind unkorreliert: $\mathrm{Cov}[X_i, X_j] \;=\; 0 \quad$ für alle $i \neq j$.

(ii) Die Varianzen sind beschränkt: $v \;:=\; \sup_{i\in\mathbb{N}} \mathrm{Var}[X_i] \;<\; \infty$.

In diesem Fall können wir in (3.4.2) $c_0 = v$ und $c_n = 0$ für $n \neq 0$ wählen. Insbesondere setzen wir keine Unabhängigkeit voraus, sondern nur Bedingungen an die Kovarianzen.

Satz 3.8 (Gesetz der großen Zahlen für schwach korrelierte Zufallsvariablen)**.** *Ist die Annahme erfüllt, dann gilt für alle $\varepsilon > 0$ und $n \in \mathbb{N}$:*

$$P\left[\left|\frac{S_n}{n} - \frac{E[S_n]}{n}\right| \ge \varepsilon\right] \;\le\; \frac{C}{\varepsilon^2\, n} \qquad \textit{mit} \qquad C \;:=\; c_0 + 2\sum_{n=1}^{\infty} c_n \;<\; \infty.$$

*Ist insbesondere $E[X_i] = m$ für alle $i \in \mathbb{N}$, dann **konvergieren** die Mittelwerte **stochastisch** gegen den Erwartungswert m, d.h.*

$$\lim_{n\to\infty} P\left[\left|\frac{S_n}{n} - m\right| \ge \varepsilon\right] = 0 \qquad \textit{für jedes } \varepsilon > 0.$$

Der Beweis des Gesetzes der großen Zahlen ergibt sich unmittelbar aus Lemma 3.7 und Satz 3.4:

Beweis von Satz 3.8. Nach der Annahme und Lemma 3.7 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 $n \in \mathbb{N}$. □

Bemerkung (Starkes Gesetz der großen Zahlen)**.** Unter den Voraussetzungen von Satz 3.8 gilt auch ein *starkes Gesetz der großen Zahlen*, d.h. $S_n(\omega)/n \longrightarrow m$ mit Wahrscheinlichkeit 1. Dies wird in der Vorlesung »Einführung in die Wahrscheinlichkeitstheorie« gezeigt.

Beispiel (IID Fall)**.** 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 $\operatorname{Var}[X_i] = v$ 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

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

für den Abstand des Mittelwerts der Zufallsvariablen vom Erwartungswert.

Anwendung auf 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

$$Y_n \sim \mu \qquad \text{für alle } n \ge 0. \tag{3.4.3}$$

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 (Gesetz der großen Zahlen für stationäre Markovketten)**.** *Ist die Minorisierungsbedingung* (3.2.2) *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 des Korollars. Seien $A \subseteq S$ und $i, n \in \mathbb{Z}_+$. Um die Annahme in Satz 3.8 zu verifizieren, schätzen wir die Kovarianzen der Zufallsvariablen $I_A(Y_i)$ und $I_A(Y_{i+n})$ ab. Nach (3.4.3) 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

$$\begin{aligned}
\operatorname{Cov}\left[(Y_i), I_A(Y_{i+n})\right] \;&=\; E\left[I_A(Y_i)\, I_A(Y_{i+n})\right] \;-\; E\left[I_A(Y_i)\right]\, E\left[I_A(Y_{i+n})\right] \\
&= \sum_{a\in A}\sum_{b\in A} P[Y_i = a, Y_{i+n} = b] \;-\; \sum_{a\in A} P[Y_i = a] \sum_{b\in A} P[Y_{i+n} = b] \\
&= \sum_{a\in A}\sum_{b\in A} \mu(a)\pi^n(a,b) \;-\; \sum_{a\in A}\mu(a) \sum_{b\in A}\mu(b) \\
&= \sum_{a\in A}\mu(a) \sum_{b\in A}(\pi^n(a,b) - \mu(b)) \\
&\leq\; 2\sum_{a\in A}\mu(a)\; d_{TV}(\pi^n(a,\cdot),\mu) \\
&\leq\; 2\sum_{a\in A}\mu(a)\,(1-\delta)^{\lfloor n/r\rfloor} \;\leq\; 2\,(1-\delta)^{\lfloor n/r\rfloor}.
\end{aligned}$$

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 3.3 aus der Minorisierungsbedingung (3.2.2).

Aus der Abschätzung sehen wir, dass die Zufallsvariablen $X_i := I_A(Y_{i-1})$ die Annahme in (3.4.2) 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 3.8 anwenden. Die Behauptung folgt dann wegen $S_n = \sum_{i=1}^n X_i$ und

$$E[X_i] \;=\; P[Y_{i-1} \in A] \;=\; \mu[A] \qquad \text{für alle } i \in \mathbb{N}.$$

□

Teil II

Numerische Verfahren

Kapitel 4

Stochastische Simulation und Monte-Carlo-Verfahren

DIESES KAPITEL WIRD NOCH ÜBERARBEITET

Simulationsverfahren für Stichproben von Wahrscheinlichkeitsverteilungen gehen in der Regel von der Existenz einer Folge von auf dem Intervall $[0,1]$ gleichverteilten, unabhängigen Zufallszahlen aus, die durch einen Zufallszahlengenerator erzeugt werden. In Wirklichkeit simulieren Zufallszahlengeneratoren natürlich nur auf $\{k\,m^{-1} : k = 0, 1, \ldots, m-1\}$ 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. In Abschnitt 4.1 gehen wir kurz auf Verfahren und Probleme bei der Erzeugung von Pseudozufallszahlen mithilfe eines Zufallszahlengenerators ein. In den Abschnitten 4.2 und 4.3 betrachten wir dann verschiedene grundlegenden Verfahren, um Stichproben von allgemeineren Wahrscheinlichkeitsverteilungen zu simulieren. Schließlich betrachten wir in Abschnitt 4.4 Monte-Carlo-Verfahren, die Gesetze der großen Zahlen verwenden, um Wahrscheinlichkeiten und Erwartungswerte ausgehend von Stichproben näherungsweise zu berechnen.

Um Simulationsverfahren zu analysieren, benötigen wir noch den Begriff einer auf $[0,1]$ gleichverteilten reellwertigen Zufallsvariablen. Die Existenz solcher Zufallsvariablen auf einem geeigneten Wahrscheinlichkeitsraum wird hier vorausgesetzt, und kann erst in der Vorlesung »Analysis III« bzw. in der »Einführung in die Wahrscheinlichkeitstheorie« gezeigt werden.

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

*(i) Eine **reellwertige Zufallsvariable** ist eine Abbildung $U : \Omega \to \mathbb{R}$, für die gilt:*

$$\{\omega \in \Omega : U(\omega) \leq y\} \in \mathcal{A} \qquad \text{für alle } y \in \mathbb{R}.$$

(ii) Eine reellwertige Zufallsvariable U heißt ***gleichverteilt auf dem Intervall*** $[0,1]$*, falls*

$$P[U \leq y] = y \qquad \textit{für alle } y \in [0,1] \textit{ gilt.}$$

(iii) Reellwertige Zufallsvariablen $U_i \colon \Omega \to \mathbb{R}$, $i \in I$, heißen ***unabhängig****, falls die Ereignisse $\{U_i \leq y_i\}$, $i \in I$, für alle $y_i \in \mathbb{R}$ unabhängig sind.*

4.1 Pseudozufallszahlen

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.

Zufallszahlengeneratoren

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. Wir betrachten einige Beispiele:

Lineare Kongruenzgeneratoren (LCG)

Bei linearen Kongruenzgeneratoren ist die Rekurrenzrelation vom Typ

$$x_{n+1} = (a x_n + c) \bmod m, \qquad n = 0, 1, 2, \ldots.$$

Hierbei sind a, c und m geeignet zu wählende positive ganze Zahlen, zum Beispiel:

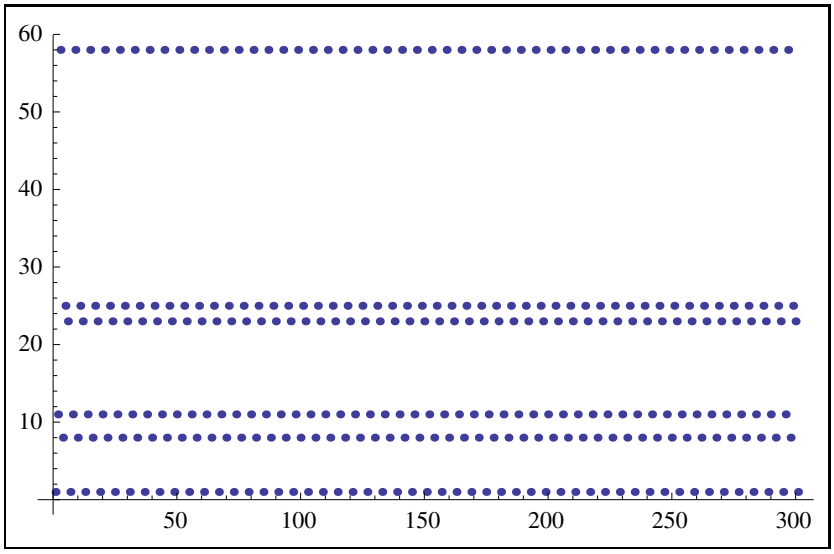
ZX81-Generator :	$m = 2^{16}+1,$	$a = 75,$	$c = 0.$
RANDU, IBM 360/370 :	$m = 2^{31},$	$a = 65539,$	$c = 0.$
Marsaglia-Generator :	$m = 2^{32},$	$a = 69069,$	$c = 1.$
Langlands-Generator :	$m = 2^{48},$	$a = 142412240584757,$	$c = 11.$

Um einen ersten Eindruck zu erhalten, wie die Qualität der erzeugten Pseudozufallszahlen von a, c und m abhängt, implementieren wir die Generatoren mit »Mathematica«:

```
f[x_] := Mod[a x + c, m]
```

Beispiel. Wir beginnen zur Demonstration mit dem Beispiel eines ganz schlechten LCG:

```
a = 11; c = 0; m = 63; pseudorandomdata = NestList[f, 1, 300];
  ListPlot[pseudorandomdata]
```



Die Folge von Zufallszahlen ist in diesem Fall periodisch mit einer Periode, die viel kleiner ist als die maximal mögliche (63). Dies rechnet man auch leicht nach.

Periodizität mit Periode kleiner als m kann man leicht ausschließen. Es gilt nämlich:

Satz (Knuth)**.** *Die Periode eines LCG ist gleich m genau dann, wenn*

i) *c und m teilerfremd sind,*

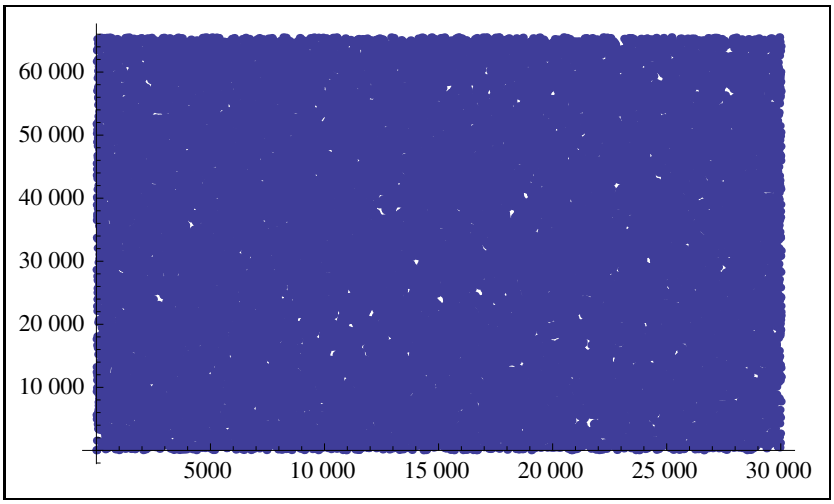
ii) *jeder Primfaktor von m ein Teiler von $a-1$ ist, und*

iii) *falls 4 ein Teiler von m ist, dann auch von $a-1$.*

Beweis. siehe `D. Knuth`: »The art of computer programming, Vol. 2.« □

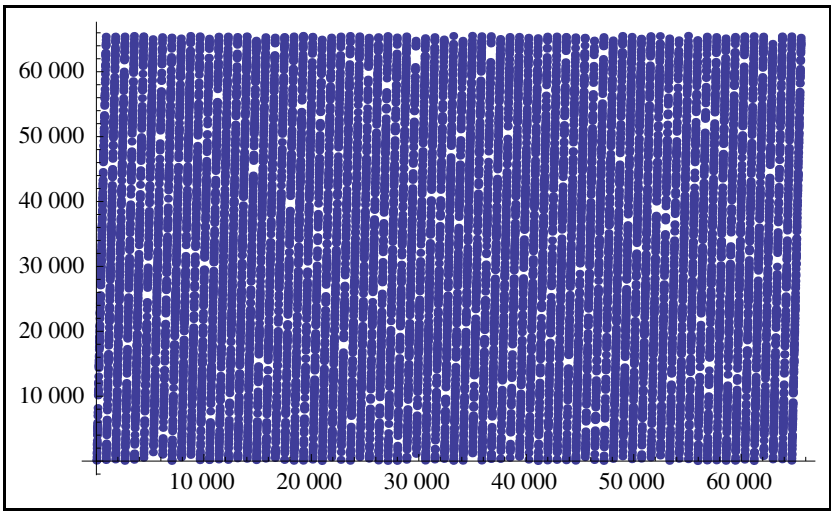
Beispiel (ZX 81-Generator)**.** Hier ergibt sich ein besseres Bild, solange wir nur die Verteilung der einzelnen Zufallszahlen betrachten:

```
a = 75; c = 0; m = 2^16 + 1; pseudorandomdata = NestList[f, 1, 30000];
    ListPlot[pseudorandomdata]
```



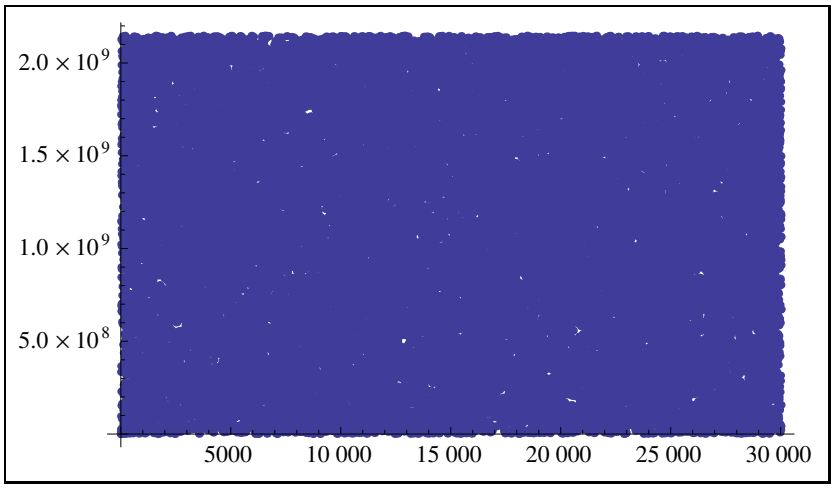
Fassen wir jedoch Paare (x_i, x_{i+1}) von aufeinanderfolgenden Pseudozufallszahlen als Koordinaten eines zweidimensionalen Pseudozufallsvektors auf, und betrachten die empirische Verteilung dieser Vektoren, so ergibt sich keine besonders gute Approximation einer zweidimensionalen Gleichverteilung:

```
blocks = Partition[pseudorandomdata, 2]; ListPlot[blocks]
```

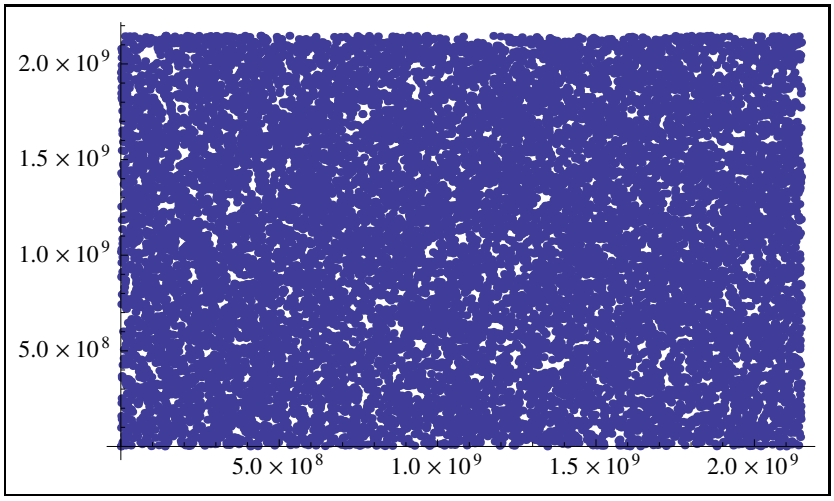


Beispiel (RANDU)**.** Hier scheinen sowohl die einzelnen Pseudozufallszahlen x_i als auch die Vektoren (x_i, x_{i+1}) näherungsweise gleichverteilt zu sein:

```
a = 65539; c = 0; m = 2^31; pseudorandomdata = NestList[f, 1, 30000];
    ListPlot[pseudorandomdata]
```

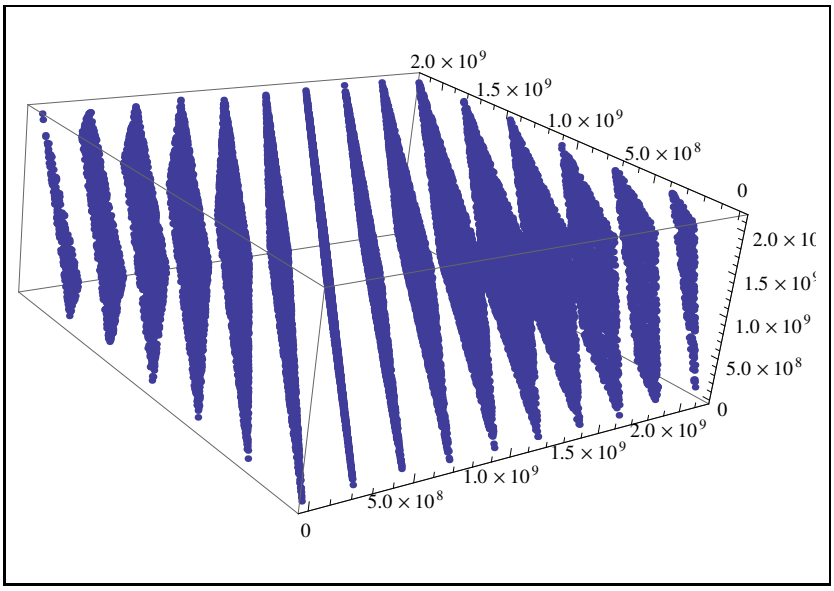



```
blocks = Partition[pseudorandomdata, 2]; ListPlot[blocks]
```



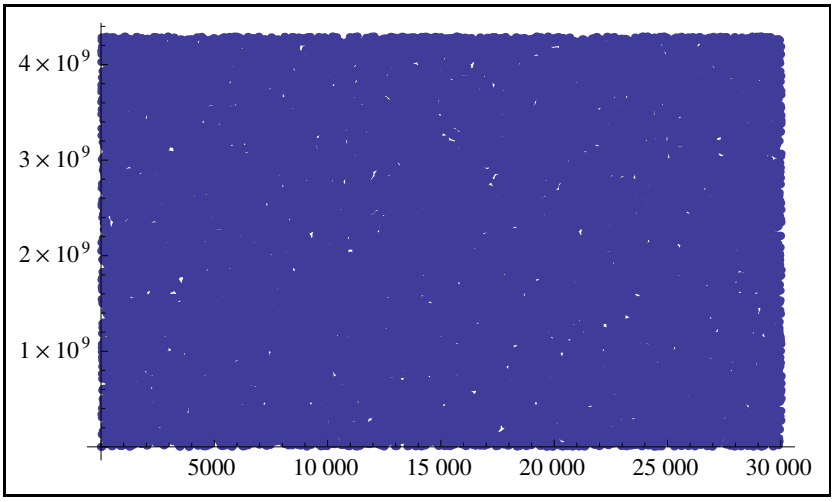
Fassen wir aber jeweils drei aufeinanderfolgende Pseudozufallszahlen als Koordinaten eines Vektors (x_i, x_{i+1}, x_{i+2}) im $\mathbb{Z}^3$ auf, dann ist die empirische Verteilung dieser Pseudozufallsvektoren keine Gleichverteilung mehr, sondern konzentriert sich auf nur 15 zweidimensionalen Hyperebenen:

```
blocks3 = Partition[pseudorandomdata, 3]; ListPointPlot3D[blocks3]
```

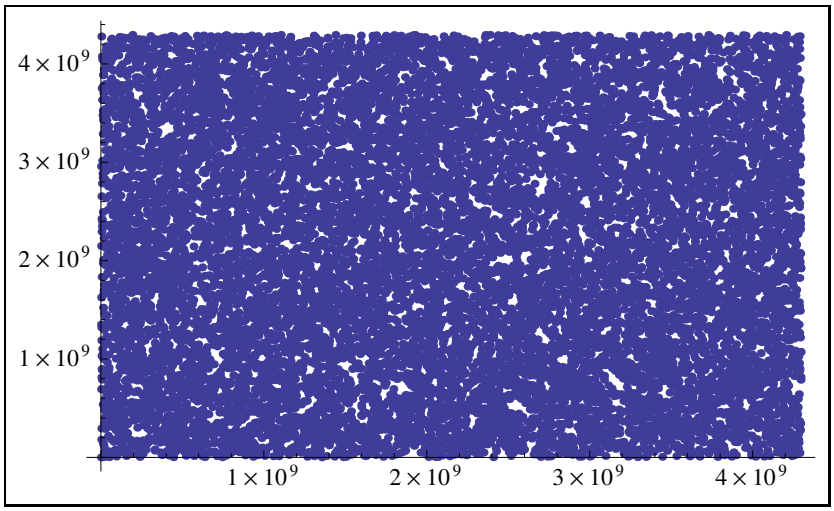


Beispiel (Marsaglia-Generator)**.** Der von Marsaglia 1972 vorgeschlagene LCG besteht dagegen alle obigen Tests (und einige weitere):

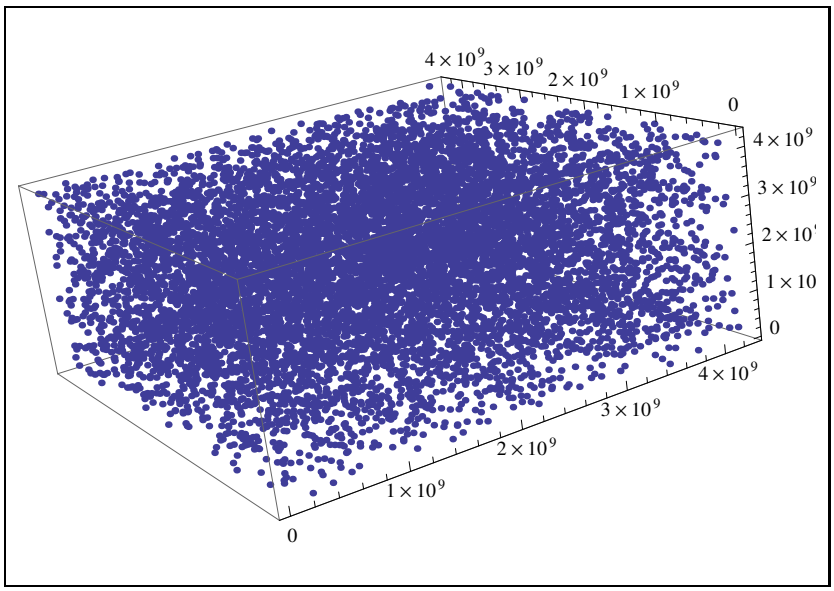
```
a = 60069; c = 1; m = 2^32; pseudorandomdata = NestList[f, 1, 30000];
    ListPlot[pseudorandomdata]
```




```
blocks = Partition[pseudorandomdata, 2]; ListPlot[blocks]
```




```
blocks3 = Partition[pseudorandomdata, 3]; ListPointPlot3D[blocks3]
```



Dies bedeutet natürlich nicht, daß die vom Marsaglia-Generator erzeugte Folge eine für *alle* Zwecke akzeptable Approximation einer Folge von unabhängigen Stichproben von der Gleichverteilung ist. Da die Folge in Wirklichkeit deterministisch ist, kann man einen Test konstruieren, der sie von einer echten Zufallsfolge unterscheidet.

Shift-Register-Generatoren

Bei Shift-Register-Generatoren 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:

$$x_{n+1} = Tx_n, \qquad n = 0,1,2,\ldots.$$

Kombination von Zufallszahlengeneratoren

Zufallszahlengeneratoren 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, der einen LCG und zwei Shift-Register-Generatoren kombiniert, Periode 2^{95} hat, und umfangreiche statistische Tests besteht.

Simulation von Gleichverteilungen

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 Dezimalzahlen 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 $\{0,\ 1,\ \ldots,\ \mathrm{m}\ -\ 1\}$ ist. In »Mathematica« kann man mit

```
RandomReal[spec, WorkingPrecision → k]
```

pseudozufällige Dezimalzahlen mit einer beliebigen vorgegebenen Anzahl k von Nachkommastellen erzeugen.

Zufallspermutationen

Der folgende Algorithmus erzeugt eine (pseudo-)zufällige Permutation aus S_n :

Algorithmus 4.1 (RPERM)**.**

```
rperm[n_] :=
 Module[{x = Range[n], k, a},   Beginn mit Liste 1,2,...,n
  Do[
   k = RandomInteger[{i, n}];
   a = x[[i]]; x[[i]] = x[[k]]; x[[k]] = a;     (Vertausche x[[i]] und x[[k]])
   , {i, n − 1}];   (Schleife, i läuft von 1 bis n − 1)
  x        (Ausgabe von x) ]
```

```
rperm[17] {12, 5, 13, 8, 17, 9, 10, 6, 1, 7, 16, 15, 14, 4, 2, 3, 11}
```

ÜBUNG:
Sei $\Omega_n = \{1, 2, \ldots, n\} \times \{2, 3, \ldots, n\} \times \cdots \times \{n-1, n\}$.

a) Zeigen Sie, daß die Abbildung $X(\omega) = \tau_{n-1,\omega_{n-1}} \circ \cdots \circ \tau_{2,\omega_2} \circ \tau_{1,\omega_1}$ eine Bijektion von Ω_n nach S_n ist ($\tau_{i,j}$ bezeichnet die Transposition von i und j).

b) Folgern Sie, daß der Algorithmus oben tatsächlich eine Stichprobe einer gleichverteilten Zufallspermutation aus S_n simuliert.

4.2 Simulationsverfahren

Ein Zufallszahlengenerator simuliert Stichproben $u_1 = U_1(\omega)$, $u_2 = U_2(\omega)$, ... von auf $[0, 1]$ gleichverteilten, unabhängigen Zufallsvariablen. Wie erzeugt man daraus Stichproben von diskreten Verteilungen ?

Das direkte Verfahren

Sei $S = \{a_1, a_2, \ldots\}$ endlich oder abzählbar unendlich, und μ eine Wahrscheinlichkeitsverteilung auf S mit Gewichten $p_i = p(a_i)$. Wir setzen

$$s_n := \sum_{i=1}^{n} p_i, \qquad n \in \mathbb{N}, \qquad \text{»kumulative Verteilungsfunktion«}.$$

Sei $U \colon \Omega \to [0, 1)$ eine gleichverteilte Zufallsvariable. Wir setzen

$$X(\omega) := a_i, \qquad \text{falls } s_{i-1} < U(\omega) \leq s_i, \qquad i \in \mathbb{N}.$$

Lemma 4.2. *Falls $U \sim \mathrm{Unif}[0, 1)$, gilt $X \sim \mu$.*

Beweis. Für alle $i \in \mathbb{N}$ gilt:

$$P[X = a_i] = P[s_{i-1} < U \leq s_i] = P[U \leq s_i] - P[U \leq s_{i-1}] = s_i - s_{i-1} = p_i.$$

□

Algorithmus 4.3 (Direkte Simulation einer diskreten Verteilung)**.**
INPUT: Gewichte $p_1, p_2, \ldots,$
OUTPUT: Pseudozufallsstichprobe x von μ.

$n := 1$
$s := p_1$
erzeuge Zufallszahl $u \sim \mathrm{Unif}[0,1)$
while $u > s$ **do**
  $n := n + 1$
  $s := s + p_n$
end while
return $x := a_n$

Bemerkung. a) Die mittlere Anzahl von Schritten des Algorithmus ist

$$\sum_{n=1}^{\infty} n\, p_n = \text{ Erwartungswert von Wahrscheinlichkeitsverteilung } (p_n) \text{ auf } \mathbb{N}.$$

b) Für große Zustandsräume S ist das direkte Verfahren oft nicht praktikabel, siehe Übung.

Das Acceptance-Rejection-Verfahren

Sei S eine endliche oder abzählbare Menge, μ eine Wahrscheinlichkeitsverteilung auf S mit Massenfunktion $p(x)$, und ν eine Wahrscheinlichkeitsverteilung auf S mit Massenfunktion $q(x)$. Angenommen, wir können unabhängige Stichproben von ν erzeugen. Wie können wir daraus Stichproben von μ erhalten?

IDEE: Erzeuge Stichprobe x von ν, akzeptiere diese mit Wahrscheinlichkeit proportional zu $\frac{p(x)}{q(x)}$, sonst verwerfe die Stichprobe und wiederhole.

ANNAHME:

$$\text{es gibt ein } c \in [1,\infty): \qquad p(x) \le c\, q(x) \qquad \text{für alle } x \in S.$$

Aus der Annahme folgt:

$$\frac{p(x)}{c\, q(x)} \le 1 \qquad \text{für alle } x \in S,$$

d.h. wir können $\frac{p(x)}{c\, q(x)}$ als **Akzeptanzwahrscheinlichkeit** wählen.

Algorithmus 4.4 (Acceptance-Rejection-Verfahren)**.**
INPUT: Gewichte $p(y)$, $q(y)$, c $(y \in S)$,
OUTPUT: Stichprobe x von μ.
repeat
  erzeuge Stichprobe $x \sim \nu$
  erzeuge Stichprobe $u \sim \mathrm{Unif}[0,1]$

until $\frac{p(x)}{c\,q(x)} \geq u$ {akzeptiere mit Wahrscheinlichkeit $\frac{p(x)}{c\,q(x)}$}
return x

ANALYSE DES ALGORITHMUS

Für die verwendeten Zufallsvariablen gilt:

$$\begin{aligned} &X_1, X_2, \ldots \sim \nu, \qquad \text{(Vorschläge)}, \\ &U_1, U_2, \ldots \sim \mathrm{Unif}[0,1]. \end{aligned}$$

Es gilt Unabhängigkeit, d.h.

$$P[X_1 = a_1, \ldots, X_n = a_n, U_1 \leq y_1, \ldots, U_n \leq q_n] = \prod_{i=1}^{n} P[X_i = a_i] \cdot \prod_{i=1}^{n} P[U_i \leq y_i]$$

für alle $n \in \mathbb{N}$, $a_i \in S$ und $y_i \in \mathbb{R}$.

Seien

$$\begin{aligned} T &= \min\left\{ n \in \mathbb{N} \;\middle|\; \tfrac{p(X_n)}{c\,q(X_n)} \geq U_n \right\} && \text{die »Akzeptanzzeit«, und} \\ X_T(\omega) &= X_{T(\omega)}(\omega) && \text{die ausgegebene Stichprobe.} \end{aligned}$$

des Acceptance-Rejection-Verfahrens. Wir erhalten:

Satz 4.5. *(i) T ist **geometrisch verteilt** mit Parameter $1/c$,*

(ii) $X_T \sim \mu$.

Bemerkung. Insbesondere ist die mittlere Anzahl von Schritten bis Akzeptanz:

$$E[T] = c.$$

Beweis. i) Sei

$$A_n := \left\{ \frac{p(X_n)}{c\,q(X_n)} \geq U_n \right\}.$$

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. Dies wird in der Vorlesung »Einführung in die Wahrscheinlichkeitstheorie« bewiesen. Zudem gilt wegen der Unabhängigkeit von X_n und U_n:

$$\begin{aligned} P[A_n] &= \sum_{a \in S} P\left[\left\{ U_n \leq \frac{p(a)}{c\,q(a)} \right\} \cap \{X_n = a\} \right] \\ &= \sum_{a \in S} P\left[\left\{ U_n \leq \frac{p(a)}{c\,q(a)} \right\} \right] \cdot P[X_n = a] \\ &= \sum_{a \in S} \frac{p(a)}{c\,q(a)} \cdot q(a) = \frac{1}{c}. \end{aligned}$$

Also ist

$$T(\omega) = \min\{n \in \mathbb{N} \mid \omega \in A_n\}$$

geometrisch verteilt mit Parameter $1/c$.

ii)

$$\begin{aligned}
P[X_T = a] &= \sum_{n=1}^{\infty} P[\{X_T = a\} \cap \{T = n\}] \\
&= \sum_{n=1}^{\infty} P[\{X_n = a\} \cap A_n \cap A_1^C \cap \dots A_{n-1}^C] \\
&= \sum_{n=1}^{\infty} P[\{X_n = a\} \cap \left\{ \frac{p(a)}{c\, q(a)} \geq U_n \right\} \cap A_1^C \cap \dots A_{n-1}^C] \\
&= \sum_{n=1}^{\infty} q(a)\, \frac{p(a)}{c\, q(a)} \left(1 - \frac{1}{c}\right)^{n-1} \\
&= \frac{p(a)}{c} \sum_{n=1}^{\infty} \left(1 - \frac{1}{c}\right)^{n-1} \\
&= \frac{p(a)}{c}\, \frac{1}{1 - (1 - \frac{1}{c})} = p(a).
\end{aligned}$$

□

4.3 Metropolis-Algorithmus und Gibbs-Sampler

Häufig sind direkte oder Acceptance-Rejection-Verfahren zur Simulation von Stichproben einer Wahrscheinlichkeitsverteilung μ nicht praktikabel. Eine Alternative ist die Simulation einer Markovkette (X_n) mit Gleichgewicht μ. Konvergiert die Markovkette ins Gleichgewicht, dann ist die Verteilung von X_n für hinreichend große n ungefähr gleich μ. Eine Stichprobe x_n von X_n ist daher auch eine Näherung einer Stichprobe von μ. Um eine Markovkette mit Gleichgewicht μ zu finden, benutzt man in der Regel die hinreichende Detailed-Balance-Bedingung. Die zwei wichtigsten Verfahren, die sich auf diese Weise ergeben, sind der Metropolis-Hastings-Algorithmus und der Gibbs Sampler.

Metropolis-Hastings-Algorithmus

Sei $q(x, y)$ eine symmetrische stochastische Matrix, d.h. $q(x, y) = q(y, x)$ für alle $x, y \in S$. Dann erfüllt die Gleichverteilung die Detailed-Balance-Bedingung (3.2.1). Sei nun μ eine be-

liebige Wahrscheinlichkeitsverteilung auf S mit $\mu(x) > 0$ für alle $x \in S$. Wie können wir die Übergangsmatrix q so modifizieren, dass die Detailed-Balance-Bedingung bzgl. μ erfüllt ist?

Algorithmus 4.6 (Metropolis-Algorithmus (Update $x \to y$))**.** schlage Übergang $x \to y$ mit Wahrscheinlichkeit $q(x, y)$ vor
akzeptiere Übergang mit Wahrscheinlichkeit $\alpha(x, y) \in [0, 1]$
sonst verwerfe Vorschlag und bleibe bei x

ÜBERGANGSMATRIX:

$$p(x,y) := \begin{cases} \alpha(x,y)\, q(x,y) & \text{für } y \neq x, \\ 1 - \sum_{y \neq x} \alpha(x,y)\, q(x,y) & \text{für } y = x. \end{cases}$$

Die Detailed Balance-Bedingung lautet:

$$\mu(x)\, \alpha(x,y) = \mu(y)\, \alpha(y,x) \qquad \text{für alle } x, y \in S.$$

Sie ist äquivalent dazu, dass

$$b(x,y) := \mu(x)\, \alpha(x,y)$$

symmetrisch in x und y ist. Was ist die größtmögliche Wahl von $b(x, y)$?
Aus $\alpha(x, y) \leq 1$ folgen

$$\begin{aligned} b(x,y) &\leq \mu(x), \\ b(x,y) = b(y,x) &\leq \mu(y), \end{aligned}$$

und somit

$$b(x,y) \leq \min(\mu(x), \mu(y)).$$

Der größtmögliche Wert $b(x, y) = \min(\mu(x), \mu(y))$ entspricht gerade

$$\alpha(x,y) = \min\left(1, \frac{\mu(y)}{\mu(x)}\right) = \begin{cases} 1 & \text{falls } \mu(y) \geq \mu(x), \\ \frac{\mu(y)}{\mu(x)} & \text{falls } \mu(x) \geq \mu(y). \end{cases}$$

Definition. *Die Markov-Kette mit Übergangsmatrix*

$$p(x,y) = \min\left(1, \frac{\mu(y)}{\mu(x)}\right) \cdot q(x,y) \qquad \textit{für } y \neq x$$

heißt ***Metropolis-Kette*** *mit Vorschlagsverteilung $q(x, y)$ und Gleichgewicht μ.*

Satz 4.7. *μ erfüllt die Detailed Balance-Bedingung bzgl. p.*

Beweis. siehe oben. □

Die Konvergenz ins Gleichgewicht der Metropolis-Kette folgt auf endlichen Zustandsräumen unter schwachen Voraussetzungen aus dem Konvergenzsatz für Markovketten. Ist S endlich, $\mu(x) > 0$ für alle $x \in S$ und nicht konstant, und $q(x, y)$ irreduzibel, dann ist $p(x, y)$ irreduzibel und aperiodisch. Somit erhalten wir Konvergenz ins Gleichgewicht nach Satz 3.2. Diese asymptotische Aussage löst aber noch nicht die praktischen Probleme, denn die Konvergenz ins Gleichgewicht kann sehr langsam erfolgen ! Wichtig sind daher Abschätzungen der Konvergenzgeschwindigkeit und explizite Fehlerschranken. Diese sind in der Regel stark problemabhängig, und in anwendungsrelevanten Fällen meist nicht leicht herzuleiten.

Gibbs-Sampler

Sei $S = S_1 \times \cdots \times S_d$ ein endlicher Produktraum, $\mu(x_1, \ldots, x_d)$ eine Wahrscheinlichkeitsverteilung auf S und

$$\mu_i(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) := \frac{\mu(x_1, \ldots, x_d)}{\sum_{z \in S_i} \mu(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_d)}$$

die bedingte Verteilung der i-ten Komponente gegeben die übrigen Komponenten.

Algorithmus 4.8 (Gibbs-Sampler (Update $x \to y$))**.** $y := x$
for $i := 1, \ldots d$ **do**
update $y_i \sim \mu_i(\, \bullet \mid y_1, \ldots y_{i-1}, y_{i+1}, \ldots, y_d)$
end for
return y

ÜBERGANGSMATRIX:

$$p = p_d \, p_{d-1} \cdots p_1,$$

wobei

$$p_i(x, y) = \begin{cases} \mu_i(y_i \mid y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_d) & \text{falls } y_k = x_k \text{ für alle } k \neq i, \\ 0 & \text{sonst.} \end{cases}$$

Satz 4.9. *i) μ erfüllt die Detailed Balance-Bedingung bzgl. p_i für alle $i = 1, \ldots, d$.*

ii) μ ist ein Gleichgewicht von p.

Beweis. i) Der Beweis der ersten Aussage ist eine Übungsaufgabe.

ii) Nach der ersten Aussage ist μ ein Gleichgewicht von p_i für alle i. Also gilt auch

$$\mu\, p = \mu\, p_d\, p_{d-1} \cdots p_1 = \mu.$$

□

Bemerkung. Zur Simulation von Y_n, $n \geq 0$, genügt es, die Massenfunktion $\mu(x)$ bis auf eine multiplikative Konstante zu kennen:
aus $\mu(x) = C\, f(X)$ folgt

$$\alpha(x, y) = \min\big(1, \frac{f(y)}{f(x)}\big) \qquad \text{unabhängig von } C.$$

Simulated Annealing

Beispiel (Rucksackproblem)**.** Gegeben:

$$\begin{aligned} \omega_1, \ldots, \omega_d \in \mathbb{R}, &\quad \text{»Gewichte«}, \\ v_1, \ldots, v_d \in \mathbb{R}, &\quad \text{»Werte«}. \end{aligned}$$

Rucksack mit maximalem Gewicht $b > 0$, packe soviel Wert wie möglich ein.

$$\begin{aligned} S &= \{0,1\}^d, && \text{alle Konfigurationen}, \\ S_b &= \{(z_1, \ldots, z_d) \in S : \textstyle\sum_{i=1}^d z_i\, w_i \leq b\}, && \text{zulässige Konfigurationen}, \\ & z_i = 1 : i\text{-ter Gegenstand im Rucksack}. \end{aligned}$$

RUCKSACKPROBLEM:

$$\text{maximiere } V(z) = \textstyle\sum_{i=1}^d z_i\, v_i \text{ unter Nebenbedingung } z \in S_b.$$

Das Rucksackproblem ist **NP-vollständig**, insbesondere ist keine Lösung in $O(d^k)$ Schritten für ein $k \in \mathbb{N}$ bekannt.

STOCHASTISCHER ZUGANG: SIMULATED ANNEALING
Für $\beta > 0$ betrachten wir die Wahrscheinlichkeitsverteilung

$$\mu_\beta(z) = \begin{cases} \frac{1}{Z_\beta}\, e^{\beta\, V(z)} & \text{für } z \in S_b, \\ 0 & \text{für } z \in S \backslash S_b, \end{cases}$$

auf S, wobei $Z_\beta = \sum_{z \in S_b} e^{\beta V(z)}$ eine Konstante ist, die μ zu einer Wahrscheinlichkeitsverteilung normiert. Für $\beta = 0$ ist μ_β die Gleichverteilung auf S_b. Für $\beta \to \infty$ konvergiert μ_β gegen die Gleichverteilung auf der Menge der globalen Maxima von V, denn:

$$\mu_\beta(z) = \frac{e^{\beta V(z)}}{Z_\beta} = \frac{1}{\sum_{y \in S_b} e^{\beta (V(y) - V(z))}} \longrightarrow \begin{cases} 0 & \text{falls } V(z) \neq \max V, \\ \frac{1}{|\{y \mid V(y) = \max V\}|} & \text{falls } V(z) = \max V. \end{cases}$$

IDEE: Simuliere Stichprobe z von μ_β für β groß ($\beta \to \infty$). Dann ist $V(z)$ **wahrscheinlich** nahe dem Maximalwert.

METROPOLIS-ALGORITHMUS: Sei $x^+ := \max(x, 0)$ der Positivteil von x. Wir wählen als Vorschlagsmatrix die Übergangsmatrix

$$q(z, w) := \begin{cases} \frac{1}{d} & \text{falls } z_i \neq w_i \text{ für genau ein } i \in \{1, \ldots, d\}, \\ 0 & \text{sonst,} \end{cases}$$

des Random Walks auf $\{0, 1\}^d$. Für die Akzeptanzwahrscheinlichkeit ergibt sich

$$\alpha_\beta(z, w) = \min\left(1, \frac{\mu_\beta(w)}{\mu_\beta(z)}\right) = \begin{cases} e^{-\beta (V(z) - V(w))} & \text{für } z, w \in S_b, \\ 0 & \text{für } z \in S_b, w \notin S_b. \end{cases}$$

Der Vorschlag w wir also mit Wahrscheinlichkeit 1 akzeptiert, wenn $V(w) \geq V(z)$ gilt – andernfalls wird der Vorschlag nur mit Wahrscheinlichkeit $\exp -\beta\,(V(z) - V(w))$ akzeptiert.

Algorithmus 4.10 (Simulation einer Markov-Kette mit Gleichgewicht μ_β)**.** initialisiere $z^{(0)} \in S_b$
for $n = 1, 2, \ldots$ **do**
  $z^{(n)} := w := z^{(n-1)}$
  erzeuge $i \sim \mathrm{Unif}\{1, \ldots, d\}$
  $w_i := 1 - w_i$
  if $w \in S_b$ **then**
    erzeuge $u \sim \mathrm{Unif}(0, 1)$
    if $u \leq \alpha_\beta(z, w)$ **then**
      $z^{(n)} := w$
    end if
  end if
end for

Algorithmus 4.11 (Simulated Annealing)**.** Wie Algorithmus 4.10 aber mit $\beta = \beta(n) \to \infty$ für $n \to \infty$.

Bemerkung. a) PHYSIKALISCHE INTERPRETATIONEN:

μ_β ist die Verteilung im thermodynamischen Gleichgewicht für die Energiefunktion $H(z) = -V(z)$ bei der Temperatur $T = 1/\beta$. Der Grenzwert $\beta \to \infty$ entspricht $T \to 0$ (»simuliertes Abkühlen«).

b) Die beim Simulated Annealing-Verfahren simulierte zeitlich inhomogene Markov-Kette findet im allgemeinen nicht das globale Maximum von V, sondern kann in lokalen Maxima »steckenbleiben«. Man kann zeigen, dass die Verteilung der Markov-Kette zur Zeit n gegen die Gleichverteilung auf den Maximalstellen konvergiert, falls $\beta(n)$ nur sehr langsam (logarithmisch) gegen $+\infty$ geht. In praktischen Anwendungen wird der Algorithmus aber in der Regel mit einem schnelleren »Cooling schedule« $\beta(n)$ verwendet. Das Auffinden eines globalen Maximums ist dann nicht garantiert – trotzdem erhält man ein oft nützliches **heuristisches** Verfahren.

4.4 Monte-Carlo-Verfahren

Sei μ eine Wahrscheinlichkeitsverteilung auf einer abzählbaren Menge S. Wir bezeichnen im folgenden die Massenfunktion ebenfalls mit μ, d.h. $\mu(x) := \mu[\{x\}]$. Sei $f : S \to \mathbb{R}$ eine reellwertige Zufallsvariable mit $E_\mu[f^2] = \sum_{x \in S} f(x)^2 \, \mu(x) < \infty$. Angenommen, wir wollen den Erwartungswert

$$\theta \;:=\; E_\mu[f] \;=\; \sum_{x \in S} f(x) \, \mu(x)$$

berechnen, aber die Menge S ist zu groß, um die Summe direkt auszuführen. In einem *Monte-Carlo-Verfahren* simuliert man eine große Anzahl unabhängiger Stichproben $X_1(\omega), \ldots, X_n(\omega)$ von μ, und approximiert den Erwartungswert θ durch den **Monte-Carlo-Schätzer**[1]

$$\widehat{\theta}_n(\omega) \;:=\; \frac{1}{n} \sum_{i=1}^{n} f(X_i(\omega)).$$

Wir wollen nun verschiedene Abschätzungen für den Approximationsfehler $|\widehat{\theta}_n - \theta|$ vergleichen. Nach dem Transformationssatz (Satz 1.6) und der Linearität des Erwartungswerts gilt:

$$E[\widehat{\theta}_n] \;=\; \frac{1}{n} \sum_{i=1}^{n} E[f(X_i)] \;=\; \frac{1}{n} \sum_{i=1}^{n} E_\mu[f] \;=\; E_\mu[f] \;=\; \theta,$$

[1] Als **Schätzer** bezeichnet man in der Statistik eine Funktion der gegebenen Daten (hier Stichproben von $X_1, \ldots, X_n$), die zum Schätzen eines unbekannten Parameters verwendet wird.

d.h. $\widehat{\theta}_n$ ist ein **erwartungstreuer Schätzer** für θ. Der **mittlere quadratische Fehler** (»MSE« = Mean Squared Error) des Schätzers ist daher durch die Varianz der Zufallsvariable $\widehat{\theta}_n$ gegeben:

$$\text{MSE}\left[\widehat{\theta}_n\right] \;=\; E\left[\left|\widehat{\theta}_n - \theta\right|^2\right] \;=\; \text{Var}\left[\widehat{\theta}_n\right].$$

Fehlerschranken für Monte-Carlo-Schätzer

Explizite Abschätzungen für den Approximationsfehler erhalten wir mit denselben Methoden wie beim Beweis von Gesetzen der großen Zahlen. Seien $X_1, X_2, \ldots$ auf $(\Omega, \mathcal{A}, P)$ unabhängige Zufallsvariablen mit Verteilung μ, und sei $\widehat{\theta}_n := \frac{1}{n}\sum_{i=1}^n f(X_i)$ für $n \in \mathbb{N}$. Eine einfache Fehlerschranke ergibt sich aus der die Čebyšev-Ungleichung:

Satz 4.12 (Čebyšev-Schranke)**.** *Für $\varepsilon > 0$ und $n \in \mathbb{N}$ gilt*

$$P\left[\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right] \;\leq\; \frac{1}{\varepsilon^2} E\left[\left|\widehat{\theta}_n - \theta\right|^2\right] \;=\; \frac{1}{n\,\varepsilon^2}\,\text{Var}_\mu[f].$$

*Insbesondere ist $\widehat{\theta}_n$ eine **konsistente Schätzfolge** für θ, d.h. für jedes $\varepsilon > 0$ gilt*

$$P\left[\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right] \longrightarrow 0 \quad \textit{für } n \to \infty.$$

Beweis. Da die Zufallsvariablen X_i unabhängig sind, sind $f(X_i)$, $i \in \mathbb{N}$, unkorreliert. Zudem gilt

$$E[f(X_i)] = \sum_{x \in S} f(x)\,\mu(x) = E_\mu[f] = \theta, \qquad \text{und}$$
$$\text{Var}[f(X_i)] = \sum_{x \in S} (f(x) - \theta)^2\,\mu(x) = \text{Var}_\mu[f] < \infty$$

nach Voraussetzung. Die Behauptung folgt nun aus Satz 3.8. □

Bemerkung. Insbesondere gilt

$$\|\widehat{\theta}_n - \theta\|_{\mathcal{L}^2} \;=\; \sqrt{E[|\widehat{\theta}_n - \theta|^2]} \;=\; O(1/\sqrt{n}).$$

Beispiel (Monte Carlo-Schätzung von Wahrscheinlichkeiten)**.** Angenommen, wir wollen die Wahrscheinlichkeit

$$p \;=\; \mu[B] \;=\; E_\mu[I_B]$$

eines Ereignisses $B \subseteq S$ näherungsweise berechnen. Ein Monte Carlo-Schätzer für p ist

$$\widehat{p}_n \;=\; \frac{1}{n}\sum_{i=1}^n I_B(X_i), \qquad X_i \text{ unabhängig mit Verteilung } \mu.$$

FEHLERKONTROLLE:

- Mithilfe der Čebyšev-Ungleichung ergibt sich:

$$P[|\widehat{p}_n - p| \geq \varepsilon] \leq \frac{1}{\varepsilon^2} \operatorname{Var}(\widehat{p}_n) = \frac{1}{n\varepsilon^2} \operatorname{Var}_\mu(I_B) = \frac{p\,(1-p)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

  Gilt beispielsweise $n \geq \frac{5}{\varepsilon^2}$, dann erhalten wir:

$$P[p \notin (\widehat{p}_n - \varepsilon, \widehat{p}_n + \varepsilon)] \leq 5\%, \qquad \text{unabhängig von } p,$$

  d.h. das zufällige Intervall $(\widehat{p}_n - \varepsilon, \widehat{p}_n + \varepsilon)$ ist ein **95%-Konfidenzintervall** für den gesuchten Wert p.

- Mithilfe der Bernstein-Ungleichung (Chernoff-Abschätzung) erhalten wir für $\delta > 0$ und $S_n := \sum_{i=1}^n I_B(X_i)$:

$$P[p \notin (\widehat{p}_n - \varepsilon, \widehat{p}_n + \varepsilon)] = P\big[\big|\frac{1}{n}S_n - p\big| \geq \varepsilon\big] \leq 2e^{-2n\varepsilon^2} \leq \delta, \qquad \text{falls } n \geq \frac{\log(2/\delta)}{2\varepsilon^2}.$$

  Für kleine δ ist die erhaltene Bedingung an n wesentlich schwächer als eine entsprechende Bedingung, die man durch Anwenden der Čebyšev-Ungleichung erhält. Für den **relativen Schätzfehler** $(\widehat{p}_n - p)/p$ ergibt sich:

$$P[|\widehat{p}_n - p| \geq \varepsilon p] \leq 2e^{-2n\varepsilon^2 p^2} \leq \delta, \qquad \text{falls } n \geq \frac{\log(2/\delta)}{2\varepsilon^2 p^2}.$$

  Die benötigte Anzahl von Stichproben für eine (ε, δ)-Approximation von p ist also polynomiell in ε, $\log(1/\delta)$ und $1/p$. Mit einer etwas modifizierten Abschätzung kann man statt der Ordnung $O(\frac{1}{p^2})$ sogar $O(\frac{1}{p})$ erhalten, siehe `Mitzenmacher und Upfal`: »Probability and Computing«.

Beispiel. Wie viele Stichproben sind nötig, damit der **relative Fehler** mit 95% Wahrscheinlichkeit unterhalb von 10% liegt? Mithilfe der Čebyšev-Ungleichung ergibt sich:

$$P[|\widehat{p}_n - p| \geq 0,1\,p] \leq \frac{p\,(1-p)}{n\,(0,1\,p)^2} = \frac{100\,(1-p)}{n\,p} \leq 0,05, \quad \text{falls } n \geq \frac{2000\,(1-p)}{p}.$$

So sind zum Beispiel für $p = 10^{-5}$ ungefähr $n \approx 2 \cdot 10^8$ Stichproben ausreichend. Dies ist nur eine obere Schranke, aber man kann zeigen, dass die tatsächlich benötigte Stichprobenzahl immer noch sehr groß ist. Für solch kleine Wahrscheinlichkeiten ist das einfache Monte Carlo-Verfahren ineffektiv, da die meisten Summanden von $\widehat{\theta}_n$ dann gleich 0 sind. Wir brauchen daher ein alternatives Schätzverfahren mit geringerer Varianz.

Beispiel (Monte-Carlo-Schätzung von $\theta = \int_{[0,1]^d} f(x)\,dx$)**.**
Das mehrdimensionale Integral ist folgendermaßen definiert:

$$\int_{[0,1]^d} f(x)\,dx := \int_0^1 \dots \int_0^1 f(x_1,\dots,x_d)\,dx_1 \dots dx_d.$$

Der Wert von θ kann mit dem folgenden Algorithmus geschätzt werden.

erzeuge Pseudozufallszahlen $u_1, u_2, \dots, u_{nd} \in (0,1)$
$x^{(1)} := (u_1, \dots, u_d)$
$x^{(2)} := (u_{d+1}, \dots, u_{2d})$
$\dots$
$x^{(n)} := (u_{(n-1)d+1}, \dots, u_{nd})$
$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n f(x^{(i)})$ ist Schätzwert für θ.

Varianzreduktion durch Importance Sampling

Sei ν eine weitere Wahrscheinlichkeitsverteilung auf S mit Massenfunktion $\nu(x) = \nu(\{x\})$. Es gelte $\nu(x) > 0$ für alle $x \in S$. Dann können wir den gesuchten Wert θ auch als Erwartungswert bzgl. ν ausdrücken:

$$\theta = E_\mu[f] = \sum_{x\in S} f(x)\,\mu(x) = \sum_{x\in S} f(x)\,\frac{\mu(x)}{\nu(x)}\nu(x) = E_\nu[f\,\varrho],$$

wobei

$$\varrho(x) = \frac{\mu(x)}{\nu(x)}$$

der Quotient der beiden Massenfunktionen ist. Ein alternativer Monte Carlo-Schätzer für θ ist daher

$$\widetilde{\theta}_n = \frac{1}{n}\sum_{i=1}^n f(Y_i)\,\varrho(Y_i),$$

wobei die Y_i unabhängige Zufallsvariablen mit Verteilung ν sind. Auch $\widetilde{\theta}_n$ ist erwartungstreu:

$$E_\nu[\widetilde{\theta}_n] = E_\nu[f\,\varrho] = \theta.$$

Für die Varianz erhalten wir:

$$\mathrm{Var}_\nu(\widetilde{\theta}_n) = \frac{1}{n}\,\mathrm{Var}_\nu(f\,\varrho) = \frac{1}{n}\big(\sum_{x\in S} f(x)^2\varrho(x)^2\nu(x) - \theta^2\big).$$

Bei geeigneter Wahl von ν kann die Varianz von $\widetilde{\theta}_n$ deutlich kleiner sein als die des Schätzers $\widehat{\theta}_n$.
Faustregel für eine gute Wahl von ν : $\nu(x)$ sollte groß sein, wenn $|f(x)|$ groß ist.
»Importance Sampling«: Mehr Gewicht für die wichtigen x!

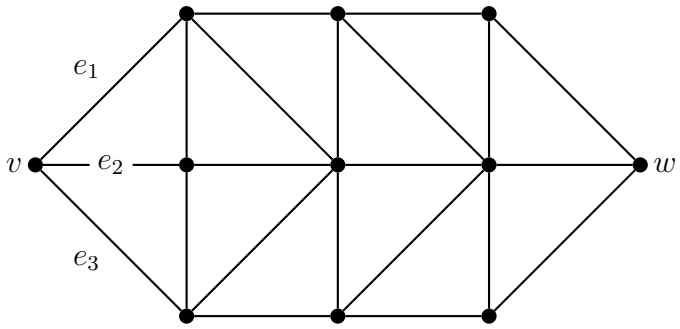


Abbildung 4.1: kleiner Beispielgraph für Perkolation

Beispiel (Zuverlässigkeit von Netzwerken; Perkolation)**.** Gegeben sei ein endlicher Graph (V, E), wobei V die Menge der Knoten und E die Menge der Kanten bezeichnet. Wir nehmen an, dass die Kanten unabhängig voneinander mit Wahrscheinlichkeit $\varepsilon \ll 1$ ausfallen. Seien $v, w \in E$ vorgegebene Knoten. Wir wollen die Wahrscheinlichkeit

$$p = P[\text{»}v \text{ nicht verbunden mit } w \text{ durch intakte Kanten«}]$$

approximativ berechnen. Sei

$$S = \{0,1\}^E = \{(x_e)_{e\in E} \mid x_e \in \{0,1\}\}$$

die Menge der Konfigurationen von intakten $(x_l = 0)$ bzw. defekten $(x_l = 1)$ Kanten und μ die Wahrscheinlichkeitsverteilung auf S mit Massenfunktion

$$\mu(x) = \varepsilon^{k(x)}(1-\varepsilon)^{|E|-k(x)}, \qquad k(x) = \sum_{e\in E} x_e.$$

Sei

$$A = \{x \in S \mid v, w \text{ nicht verbunden durch Kanten } e \text{ mit } x_e = 0\}.$$

Dann ist

$$p = \mu(A) = E_\mu[I_A].$$

Der »klassische Monte Carlo-Schätzer« für p ist

$$\widehat{p}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} I_A(X_i), \quad X_i \text{ unabhängig mit Verteilung } \mu.$$

Fordern wir nun zum Beispiel

$$\sigma(\widehat{p}_n) = \sqrt{\frac{p(1-p)}{n}} \overset{!}{\leq} \frac{p}{10},$$

dann benötigen wir eine Stichprobenanzahl

$$n \geq \frac{100\,(1-p)}{p},$$

um diese Bedingung zu erfüllen. Die Größenordnung von p für das in der obigen Graphik dargestellte Netzwerk mit $\varepsilon = 1\%$ lässt sich wie folgt abschätzen:

$$\begin{aligned} 10^{-6} = \mu(\text{»}e_1, e_2, e_3 \text{ versagen«}) \leq p \leq \mu(\text{»mindestens 3 Kanten versagen«}) \\ = \binom{22}{3} \cdot 10^{-6} \approx 1,5 \cdot 10^{-3}. \end{aligned}$$

Es sind also eventuell mehrere Millionen Stichproben nötig!
Um die benötigte Stichprobenanzahl zu reduzieren, wenden wir ein Importance Sampling-Verfahren an. Sei

$$\nu(x) = t^{-k(x)}\,(1-t)^{|E|-k(x)}, \qquad k(x) = \sum_{e \in E} x_e,$$

die Verteilung bei Ausfallwahrscheinlichkeit $t := \frac{3}{22}$. Da unter ν im Schnitt 3 Kanten defekt sind, ist der Ausfall der Verbindung bzgl. ν nicht mehr selten. Für den Schätzer

$$\widetilde{p}_n = \frac{1}{n} \sum_{i=1}^{n} I_A(Y_i) \frac{\mu(Y_i)}{\nu(Y_i)}, \qquad Y_i \text{ unabhängig mit Verteilung } \nu,$$

erhalten wir im Beispiel von oben:

$$\begin{aligned} \operatorname{Var}(\widetilde{p}_n) &= \frac{1}{n}\Big(\sum_{x \in S} I_A(x)^2 \frac{\mu(x)^2}{\nu(x)} - p^2\Big) \\ &\leq \frac{1}{n} \sum_{k=3}^{22} \binom{|E|}{k} \left(\frac{\varepsilon^2}{t}\right)^k \left(\frac{(1-\varepsilon)^2}{1-t}\right)^{|E|-k} \leq 0,0053\,\frac{p}{n}. \end{aligned}$$

Diese Abschätzung ist etwa um den Faktor 200 besser als die für den einfachen Monte Carlo-Schätzer erhaltene Abschätzung der Varianz.

Markov Chain Monte Carlo

Sei $\mu \in \mathrm{WV}(S)$, $f\colon S \to \mathbb{R}$.
GESUCHT:

$$\theta = E_\mu[f],$$

MARKOV-CHAIN-MONTE CARLO-SCHÄTZER:

$$\widehat{\theta}_{n,b} = \frac{1}{n} \sum_{k=b+1}^{b+n} f(X_k),$$

wobei $b \in \mathbb{N}$ eine feste Konstante (»burn-in-Zeit«) und $(X_k)_{k\in\mathbb{N}}$ irreduzible Markov-Ketten mit Gleichgewicht μ sind.

Satz (Ergodensatz / Gesetz der großen Zahlen für Markov-Ketten)**.** *: Für alle $b \in \mathbb{N}$ gilt:*

$$\lim_{n\to\infty} \widehat{\theta}_{n,b} = \theta \quad \textit{mit Wahrscheinlichkeit } 1,$$

Beweis. siehe Vorlesung »Stochastische Prozesse«. □

Die Analyse des Schätzfehlers ist im Allgemeinen diffizil!