Stochastik
für
Informatiker(innen)

00110000101010010

Peter Pfaffelhuber,
Thorsten Schmidt
Universität Freiburg
Mathematisches Institut
Abteilung für Stochastik
http://www.stochastik.uni-freiburg.de/~schmidt

Inhaltsverzeichnis

1. Grundlegendes

1.1 Motivation

In den letzten 20 Jahren wurden verschiedene Verfahren aus der Wahrscheinlichkeitstheorie in einigen Bereichen der Informatik immer wichtiger. Das Verständnis der grundlegenden Konzepte ist nicht nur für Simulationen wichtig, sondern hat sich als essentielles Hilfsmittel für verschiedenste Probleme entwickelt[1].

- Analyse von Daten (Suchmaschinen (Google); Aktien (Algorithmisches Handeln))
- Ein großer Teil der künstlichen Intelligenz verwendet stochastische Verfahren
- Spracherkennung, Robotersteuerung, Mensch-Computer-Interaktion werden von zufälligen und verrauschten Eingabedaten bestimmt.
- Randomisierte Algorithmen haben oft eine deutlich höhere Effizienz als deterministische Verfahren. Beispiele sind Genetische Algorithmen, Evolutionäre Algorithmen, Schwarm Intelligenz[2].
- Analyse von Routing und der Ausfall von Bauteilen ist eine wichtige Komponente für den Aufbau von robusten Systemen. Das *Cloud-Computing* unterstreicht die Wichtigkeit dieser Ansätze noch mehr.
- Noch ein paar aktuelle Stichworte: reinforcement learning (wie bei AlphaGoZero) oder PFNs- prior fitted networks basieren alle auf Techniken die in dieser Vorlesung vermittelt werden.

Das Ziel dieser Vorlesung ist zum einen die Vermittlung der wahrscheinlichkeitstheoretischen Grundlagen als auch die Einführung in die stochastische Denkweise. Neben der Befähigung, selbst Simulationen durchzuführen sollten Sie aber auch in der Lage sein, komplexere Fragestellungen etwa zum maschinellen Lernen (wie zum Beispiel das Erstellen eines statistisch validen Modells der beobachteten Daten für Vorhersagen und Analysen) selbständig zu beantworten.

Umgekehrt haben einige Felder der Wahrscheinlichkeitstheorie in den letzten Jahren mit immer größeren Datenmengen zu kämpfen, die nach effizienten Verfahren verlangen (Big Data Analysis, Aktienmärkte mit Millisekundenhandel, Analyse von Genomdaten ...) so

[1] Quelle: M. Sahami, A Course on Probability Theory for Computer Scientists

[2] Schöne Beispiele findet man etwa in: Applications of Evolutionary Computation, Springer (2010).

dass die Nachfrage nach Informatiker(innen) mit statistischem Know-How ungebrochen hoch ist.

Die Wahrscheinlichkeitstheorie oder Stochastik beschäftigt sich mit dem systematischen Studium zufälliger Ereignisse. Auf den ersten Blick erscheint es erstaunlich, dass man Regeln für den Zufall finden kann. Aber obwohl man vorab nicht genau weiß, *welches* Ergebnis ein Zufallsexperiment liefert, so kann man doch häufig angeben, *mit welcher Wahrscheinlichkeit* verschiedene Ergebnisse auftreten. Eine nicht zufällige Größe heißt *deterministisch*. Die (deterministischen) Wahrscheinlichkeiten für die möglichen Ereignisse werden *Wahrscheinlichkeitsverteilung* genannt.

Bevor wir diskutieren, wie man solche Wahrscheinlichkeiten festlegt (in der Modellierung von Zufallsvariablen) oder sie schätzt (in der Statistik) behandelt dieses Kapitel zunächst die grundlegenden Rechenregeln für Wahrscheinlichkeiten. Zentral hierfür ist der Begriff der *Zufallsvariable* und die *Kolmogorovsche* Axiomatik. In diesem ersten Kapitel werden außerdem ein paar klassische Modelle kennen lernen.

Zur Illustration starten wir mit einem interessanten Beispiel.

B 1.1 *Fahrerflucht*: Ein Zeuge[3] beobachtet nachts einen Taxifahrer, der ein parkendes Auto beschädigt und Fahrerflucht begeht. Bei der Polizei gibt er an, ein blaues Auto gesehen zu haben. In der Stadt gibt es zwei Taxiunternehmen und nur eines davon hat blaue Autos.
Um die Aussage zu untermauern, wird ein Test durchgeführt, schließlich war es dunkel. Am nächsten Abend wird unter ähnlichen Sichtverhältnissen ein Test mit dem Zeugen durchgeführt. Das Ergebnis: mit 80%iger Wahrscheinlidchkeit identifiziert er die richtige Wagenfarbe. Ist das ein ausreichender Hinweis ?

Überraschenderweise fehlt uns noch eine wichtige Angabe: Wieviele Taxis haben denn die Unternehmen? Es stellt sich heraus, dass es 25 grüne und nur 5 blaue Taxis gibt. Das heißt, von den 5 blauen Taxis werden (im Mittel) 4 als blau identifiziert, 1 jedoch falsch eingeordnet. Wir erhalten folgende Tabelle.

	Zeuge: ,,Blaues Taxi!"	Zeuge: ,,Grünes Taxi!"
Taxi ist blau	4	1
Taxi ist grün	5	20

Das heißt, wenn man alle 30 Autos vorfahren lässt, dann wird (im Mittel) der Zeuge 9 mal ein blaues Auto erkennen. Davon sind aber nur 4 in Wahrheit blau!

[3]Dies ist ein Beispiel aus dem wunderbaren Buch *Mathematikverführer* von Christoph Drösser, rororo Verlag.

Es drängt sich die Frage auf, wie man sich sicher sein kann, dass man in einer solchen Argumentation keinen Fehler macht? Hierauf findet man in der Stochastik eine tiefschürfende Antwort: Man benötigt einen präzisen, exakten Formalismus, der in der Lage ist, genaue Aussagen zu machen. Diesen werden wir in dieser Vorlesung kennenlernen.

1.2 Häufige Modelle: Münzwurf, Würfeln, Urnen

Bevor wir formal einführen, was Wahrscheinlichkeiten und Zufallsvariablen sind, behandeln wir häufige Beispiele.

B 1.2 *Münzwurf*: Wenn man eine Münze wirft, zeigt sie entweder *Kopf* oder *Zahl*. Die Wahrscheinlichkeit für das Auftreten des Ereignisses *Kopf* bezeichnen wir mit p und nennen das Experiment einen *p-Münzwurf*. Die Wahrscheinlichkeit für *Zahl* ist gerade $1-p$. Üblicherweise denkt man oft an den Fall $p = 1/2$, weil Kopf und Zahl mit derselben Wahrscheinlichkeit oben zu liegen kommen. Wird aber beispielsweise eine Reißzwecke geworfen, die entweder mit der spitzen Seite oben oder unten zu liegen kommt, wird $p \neq 1/2$ gelten.

Wie hoch ist die Wahrscheinlichkeit bei einem zweimaligen Münzwurf, dass gerade zweimal *Kopf* auftritt? Man erhält p^2 und p^n für den n-maligen Münzwurf.

B 1.3 *Würfelwurf*: Bei einem Würfelwurf gibt es statt 2 Möglichkeiten gerade 6 Möglichkeiten und die Wahrscheinlichkeit eine bestimmte Augenzahl zu werfen ist $1/6$. Die Wahrscheinlichkeit bei k Versuchen keine 6 zu werfen, ist gerade $(5/6)^k$.

B 1.4 *Urne*: In einer Urne befinden sich N farbige Kugeln wobei es n unterschiedliche Farben gibt. Dabei haben K_i Kugeln haben die Farbe i, $i \in \{1, \dots, n\}$; also ist $K_1 + \cdots + K_i = N$. Zieht man (mit verschlossenen Augen) eine Kugel aus der Urne, so hat sie die Farbe i mit Wahrscheinlichkeit K_i/N. Zieht man anschließend nochmal aus der Urne (ohne dass die erste Kugel zurückgelegt wurde), so ist die Wahrscheinlichkeit nochmal eine Kugel der Farbe i zu ziehen $\max\{(K_i-1)/(N-1), 0\}$.

Die Wahrscheinlichkeit, zwei Kugeln derselben Farbe zu ziehen, berechnet sich wie folgt:

$$\sum_{i=1}^{n} \frac{K_i}{N} \frac{K_i - 1}{N-1} \mathbb{1}_{\{K_i \geq 2\}}.$$

B 1.5 *Alphabet*: Betrachten wir ein Alphabet aus n verschiedenen Buchstaben. Nimmt man an, dass alle Buchstaben gleichberechtigt sind, so ist die Wahrscheinlichkeit, dass ein bestimmter Buchstabe auftaucht, gerade n^{-1}. Wir nehmen an, dass $n \geq 2$ und bezeichnen die ersten beiden Buchstaben mit A und B. Die Wahrscheinlichkeit, dass ein Wort der Länge 2 gerade aus den Buchstaben A und B besteht (also AB oder BA) ist

$$2 \cdot \frac{1}{n} \cdot \frac{1}{n}.$$

Wie im Urnenbeispiel, Beispiel 1.4, kann man auch unterschiedliche Wahrscheinlichkeiten für die Buchstaben betrachten: Bezeichnen wir mit p_i die Wahrscheinlichkeit, dass der i-te Buchstabe auftritt, so ist $0 \leq p_i \leq 1$ und $p_1 + \cdots + p_n = 1$. Die Wahrscheinlichkeit, dass ein Wort der Länge 2 gerade aus den Buchstaben A und B besteht ist dann

$$2 \cdot p_1 \cdot p_2.$$

Wie hoch ist die Wahrscheinlichkeit für ein Wort der Länge 3 aus den Buchstaben A, B und C?

B 1.6 *Geburtstagsproblem*: In einem Raum befinden sich 23 Personen. Wie groß ist die Wahrscheinlichkeit, dass es zwei Personen gibt die am selben Tag Geburtstag haben?

Um diese Wahrscheinlichkeit zu berechnen, ist es hilfreich, das Gegenteil *Alle Personen haben an unterschiedlichen Tagen Geburtstag* zu betrachten. Wir stellen die Personen (in Gedanken) in einer Reihe auf. Die Wahrscheinlichkeit dass die zweite Person an einem anderen Tag als die erste Person Geburtstag hat ist 364/365. (Von Schaltjahren und dem 29.2. sehen wir einmal ab.) Weiter ist die Wahrscheinlichkeit, dass die dritte Person an einem anderen Tag als die Personen eins und zwei Geburtstag hat, dann 363/365. Überlegen wir das weiter, so ist die Wahrscheinlichkeit, dass alle Personen an unterschiedlichen Tagen Geburtstag haben, gerade

$$\frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365 - 22}{365} \approx 0.493.$$

Damit ist die Wahrscheinlichkeit, dass es zwei Personen gibt, die am gleichen Tag Geburtstag haben, etwa $1 - 0.493 = 0.507$.

B 1.7 *Warteschlange*: Eine Webseite erhält sekündlich $\lambda/3600 \ll 1$ Anfragen, wobei die Anfragen jeweils unabhängig voneinander auflaufen. Herbei ist λ die erwartete Zahl von Anfragen pro Stunde. Die Wahrscheinlichkeit, dass in t Stunden kein Anfrage eingeht, ist gerade

$$\left(1 - \frac{\lambda}{3600}\right)^{3600t}.$$

Zerteilt man eine Stunde in noch kleinere Zeiteinheiten, behält jedoch die mittlere Anzahl von Anfragen pro Stunde bei, so beträgt die Wahrscheinlichkeit, dass in t Stunden keine Anfrage gefallen ist gerade

$$\left(1 - \frac{\lambda}{n}\right)^{nt} \xrightarrow{n\to\infty} e^{-\lambda t}. \tag{1.1}$$

1.3 Kombinatorik

Bereits in den Beispielen 1.3 und 1.4 wird klar, dass man oftmals etwas abzählen muss, um Wahrscheinlichkeiten zu berechnen. Diese Kunst wird auch als Kombinatorik bezeichnet.

Wir behandeln nun ein paar ausgewählte Fälle, eine Zusammenfassung findet man in Tabelle 1.1. Prinzipiell gibt es die Möglichkeiten Permutation, Variation und Kombination und hierfür jeweils mit und ohne Wiederholung.

Lemma 1.1 (Kombinatorik ohne Wiederholungen)**.** *Sei $\underline{x} = (x_1, ..., x_n)$ ein Vektor der Länge n, dessen Einträge alle verschieden sind.*

(i) Es gibt $n! := n \cdot (n-1) \cdots 1$ verschiedene Vektoren der Länge n mit denselben Einträgen wie $\underline{x}$.

(ii) Es gibt $n \cdot (n-1) \cdots (n-k+1)$ verschiedene Vektoren der Länge $k < n$, die aus den Einträgen von $\underline{x}$ gewonnen werden können (mit verschiedenen Einträgen).

(iii) Es gibt $\binom{n}{k} := \frac{n!}{k!(n-k)!}$ verschiedene, k-elementige Teilmengen von $\{x_1, ..., x_n\}$.

Beweis. (i) Stellt man einen Vektor zusammen, so hat der erste Eintrag genau n Möglichkeiten. Für jede dieser Möglichkeiten hat der zweite Eintrag dann $n-1$ Möglichkeiten usw. Daraus ergibt sich die Formel.
(ii) folgt genau wie (i) nur dass man nach dem k-ten Eintrag abbrechen muss.
(iii) Geht man von (ii) aus, so gibt es $\frac{n!}{(n-k)!}$ verschiedene Möglichkeiten, Vektoren der Länge k aus den Einträgen von $\underline{x}$ zu bekommen. Von diesen Vektoren enthalten jeweils $k!$ dieselben Elemente. Will man also die Menge der k-elementigen Teilmengen zählen, so liefert jede dieser Teilmengen genau $k!$ Vektoren, also folgt das Ergebnis. □

Lemma 1.2 (Kombinatorik mit Wiederholungen)**.** *Es gelten folgende Aussagen:*

(i) Von den Vektoren der Länge n, bestehend aus r verschiedenen Einträgen, wobei der i-te Eintrag jeweils k_i-mal vorkommt, gibt es genau $\frac{n!}{k_1! \cdots k_r!}$. Hierbei ist $k_1 + \cdots + k_r = n$ und $r \leq n$.

(ii) Zieht man aus den n Objekten genau k-mal mit Zurücklegen, so gibt es genau n^k Möglichkeiten, wenn man die Reihenfolge der gezogenen Objekte beachtet.

(iii) Zieht man aus den n Objekten genau k-mal mit Zurücklegen, so gibt es genau $\binom{n+k-1}{k}$ Möglichkeiten, wenn man die Reihenfolge der gezogenen Objekte nicht beachtet.

Beweis. (i) Könnte man alle Objekte unterscheiden, so gäbe es $n!$ Möglichkeiten. Da man jedoch von dieser Anzahl jeweils $k_1!, ..., k_r!$ für die möglichen Reihenfolgen der identischen Objekte zusammenfassen kann, folgt das Ergebnis.
(ii) ist klar.
(iii) Hier muss man etwas genauer nachdenken, wobei folgende Überlegung entscheidend ist: ist etwa $n = 5$, $k = 3$, so lässt sich jede gesuchte Möglichkeit durch einen Code der Art

$\bullet\bullet{*}{*}\bullet{*}{*}$ darstellen - die geben jeweils die Anzahl Objekte an und $*$ ist ein Trennzeichen. Der vorstehende Code bedeutet, dass vom ersten Objekt zwei Kopien gewählt wurden (deshalb $\bullet\bullet$), vom zweiten Objekt allerdings gar keines (deshalb ist zwischen den $*$'s kein $\bullet$), vom dritten eines, und vom vierten und fünften wiederum keines. Dieser Code ist also eine eindeutige Zuordnung der gesuchten Anordnungen auf die Menge der Vektoren der Länge $n+k-1$ mit k Einträgen $\bullet$ und $n-1$ Einträgen $*$. (Man beachten, dass $*$ ja hier ein Trennzeichen ist, und man $n-1$ Trennzeichen benötigt um n Felder zu trennen.) Die Anzahl dieser Vektoren ist gerade die gesuchte Anzahl, weil man nun die n-elementigen Teilmengen aller $n+k-1$ Einträge sucht. □

Bemerkung 1.3. Wir bezeichnen jede bijektive Abbildung $\mathcal{S}$ einer Menge in sich selbst als *Permutation* von $\mathcal{S}$. Die Permutationen der Zahlen $1,\ldots,n$ bezeichnen wir mit $\mathcal{S}_n$. Lemma 1.1.1 besagt, dass die Menge der Bijektionen von $\{1,...,n\}$ gerade $n!$ Elemente hat.

Als Beispiel betrachte man die Menge $\{1,2,3\}$, die man auf folgende Möglichkeiten bijektiv abbilden kann: (1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2) und (3,2,1).

	Permutation	Variation	Kombination
ohne Wiederholung	$n!$	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$
Beispiel:	Anordnung von n Zahlen, wobei jede Zahl nur einmal vorkommen darf	Ziehung von k Zahlen aus n Möglichen, wobei jede Zahl *nur einmal* vorkommen darf, *mit* Beachtung der Reihenfolge der Ziehung	Ziehung von k Zahlen aus n Möglichen, wobei jede Zahl *nur einmal* vorkommen darf, *ohne* Beachtung der Reihenfolge der Ziehung
mit Wiederholung	$\frac{n!}{k_1! \cdot \ldots \cdot k_r!}$	n^k	$\binom{n+k-1}{k}$
Beispiel:	Anordnung von n Zahlen, die nicht alle verschieden sind	Ziehung von k Zahlen aus n Möglichen, wobei jede Zahl *beliebig oft* vorkommen darf, *mit* Beachtung der Reihenfolge der Ziehung	Ziehung von k Zahlen aus n Möglichen, wobei jede Zahl *beliebig oft* vorkommen darf, *ohne* Beachtung der Reihenfolge der Ziehung

Tabelle 1.1: Kombinatorik beschäftigt sich mit dem Abzählen von Möglichkeiten.

1.4 Wahrscheinlichkeitsräume und Zufallsvariablen

In diesem Abschnitt betrachten wir Ergebnisse von Zufallsexperimenten, die reellwertig (oder mit Werten im $\mathbb{R}^n$) sind. Die zentrale Idee der zufälligen Modellierung lässt sich an folgendem Beispiel illustrieren.

B 1.8 *Würfelwurf und Augensumme*: Ein Würfel wird zweimal geworfen und die Augensumme berechnet. Wie hoch ist die Wahrscheinlichkeit für eine Augensumme von 3 ? Hierzu betrachten wir zunächst die elementaren Ereignisse $\Omega = \{(i,j) : 1 \leq i,j \leq 6\}$. Ein jedes Element dieser Menge hat die Wahrscheinlichkeit $1/36$. Die Augensumme ist gegeben durch die Abbildung $X : (i,j) \mapsto i+j$. Welche Elemente führen zur Augensumme von 3? Das sind gerade $(1,2)$ und $(2,1)$. Bemerkenswert ist, dass diese als *Urbild* der Abbildung X gedeutet werden können:

$$X^{-1}(3) := \{\omega \in \Omega : X(\omega) = 3\} = \{(1,2),(2,1)\},$$

und wir errechnen leicht die Wahrscheinlichkeit von $2/36$.

Im Folgenden werden wir diese Idee aufgreifen und Zufallsvariablen als Abbildungen einführen. Vorab definieren wir Ereignisse und fordern für Ereignisse, dass Schnitte, Vereinigungen und Komplemente wieder Ereignisse sind.

Wir starten mit einer beliebigen Menge Ω. Diese Menge bezeichnen wir als *Grundraum*. Ereignisse sind Teilmengen des Grundraumes, aber nicht alle Teilmengen sind von Interesse. Wir betrachten deswegen eine Teilmenge $\mathcal{A}$ der Potenzmenge $\mathcal{P}(\Omega)$ von Ω. Die Elemente von $\mathcal{A}$ nennen wir *Ereignisse*.

Definition 1.4. Eine Menge $\mathcal{A} \subset \mathcal{P}(\Omega)$ heißt *σ-Algebra*, falls

(i) $\Omega \in \mathcal{A}$,
(ii) für alle $A \in \mathcal{A}$ gilt, dass $\bar{A} = \Omega \backslash A \in \mathcal{A}$, *(Komplement)*
(iii) für $A_1, A_2, \cdots \in \mathcal{A}$ gilt, dass $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ *(abzählbare Vereinigung)*

Man spricht von einer *Algebra*, wenn statt (iii) nur gilt, dass: $A_1, A_2 \in \mathcal{A} \Rightarrow A_1 \cup A_2 \in \mathcal{A}$ (endliche Vereinigung). Wir nennen eine endliche oder abzählbare Menge von Mengen $A_1, A_2, \ldots$ paarweise disjunkt (p.d.), falls $A_i \cap A_j = \emptyset$ für alle $i \neq j$.

A 1.1 *Schnitte von Mengen*: Zeigen Sie, dass auch abzählbare Schnitte wieder in der σ-Algebra liegen.

Als nächstes führen wir Wahrscheinlichkeiten ein. Das ist eine Abbildung, die jedem Ereignis eine Zahl zwischen Null und Eins zuordnet. Darüber hinaus hat sie die wichtige Eigenschaft, dass sich die Wahrscheinlichkeiten von zwei (oder allgemeiner abzählbar vielen) Ereignissen, die niemals gleichzeitig auftreten addieren, wenn man sie mit *oder* verknüpft.

Definition 1.5. Sei $\mathcal{A}$ eine σ-Algebra und $P : \mathcal{A} \to \mathbb{R}$ eine Abbildung mit Werten zwischen Null und Eins. P heißt *Wahrscheinlichkeitsmaß,* falls

(i) $P(\Omega) = 1$,
(ii) $P(A) \geq 0$ für alle $A \in \mathcal{A}$,
(iii) für paarweise disjunkte Mengen $A_1, A_2, \cdots \in \mathcal{A}$ gilt, dass

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

A 1.2 *Rechenregeln*: Einige Rechenregeln folgen direkt aus der Definition: Seien $A, B \in \mathcal{A}$

(i) Gilt $A \cap B = \emptyset$, so ist $P(A \cup B) = P(A) + P(B)$,
(ii) immer gilt: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$,
(iii) Für $B \subset A$ gilt $P(A \backslash B) = P(A) - P(B)$.

Nun werden Zufallsvariablen als Abbildungen mit Werten in den reellen Zahlen eingeführt, analog lassen sich auch $\mathbb{R}^d$-wertige Zufallsvariablen einführen. Uns interessiert beispielsweise die Wahrscheinlichkeit, dass die Zufallsvariable kleiner 1 ist, was also ein Ereignis sein muss, damit man die Wahrscheinlichkeit bestimmen (messen) kann. Dies führt zu dem Begriff von *Messbarkeit*, was wir aber nicht deutlich vertiefen werden. Für eine detaillierte Ausführung sei etwa auf (**?**; **?**) verwiesen.

Definition 1.6. Eine (reellwertige) *Zufallsvariable* ist eine Abbildung von $X : \Omega \to E \subset \mathbb{R}$, für welche für alle $a \in \mathbb{R}$ gilt, dass

$$\{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{A}. \tag{1.2}$$

Eine Abbildung, welche die Eigenschaft (1.3) besitzt, heißt *meßbar.* Wir schreiben verkürzt $\{X \leq a\}$ für $\{\omega \in \Omega : X(\omega) \leq a\} = X^{-1}((-\infty, a])$ und $P[X \leq a]$ für $P(\{X \leq a\})$. Man sieht leicht, dass auch $\{X > a\}$, $\{X \in (a, b]\}$ und $\{X < a\}$ Ereignisse sind.

Ist E endlich oder abzählbar, so genügt für die Meßbarkeit bereits, dass $\{X = e\} \in \mathcal{A}$ für alle $e \in E$. Eine solche Zufallsvariable nennen wir *diskret.*

Definition 1.7. Sei X eine E-wertige Zufallsvariable.

(i) Angenommen, E ist höchstens abzählbar. Dann heißt X *diskret.* Die Funktion

$$x \mapsto P[X = x]$$

heißt *Verteilung* von X.

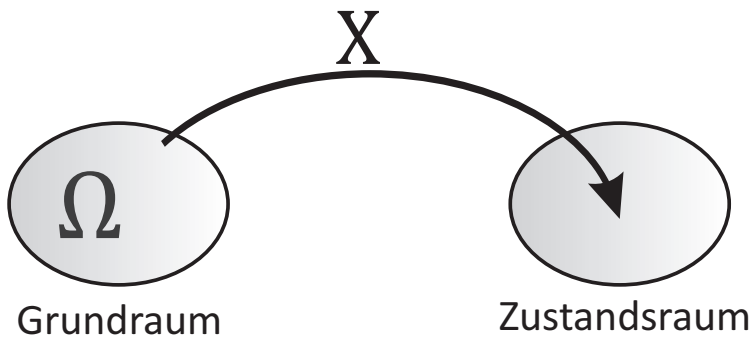


Abbildung 1.1: Eine Zufallsvariable ist eine Abbildung $X : \Omega \mapsto E$.

(ii) Angenommen, $E \subseteq \mathbb{R}$. Gibt es eine Funktion f, so dass für jedes Intervall $[a, b] \subseteq \mathbb{R}$ gilt, dass

$$P[X \in [a, b]] = \int_a^b f(x)dx,$$

so heißt f *Dichte* von X und die Zufallsvariable X heißt *stetig*. Wir schreiben dann auch $P[X \in dx] = f(x)dx$.

Etwas allgemeiner: Man kann Wahrscheinlichkeiten auch direkt auf der Bildmenge E betrachten. Hierzu wählt man ein Mengensystem $\mathcal{E} \subset \mathcal{P}(E)$, welches eine σ-Algebra ist und betrachtet

$$Q(B) = P[X \in B], \qquad B \in \mathcal{E}.$$

Dann ist Q wieder ein Wahrscheinlichkeitsmaß und wir nennen Q die *Wahrscheinlichkeitsverteilung* (oder einfach die Verteilung) von X. Bemerkenswert ist, dass es oft genügt, ein einfacheres Objekt zu studieren:

Definition 1.8. Sei X eine reellwertige Zufallsvariable. Dann heißt $F : \mathbb{R} \to [0, 1]$, definiert durch

$$F(x) = P[X \leq x], \quad x \in \mathbb{R}$$

die *Verteilungsfunktion* von X.

Ist X stetig, so gilt darüber hinaus die Beziehung (**??**), also

$$F(x) = \int_{-\infty}^x f(z)dz.$$

Ist F differenzierbar, dann ist sogar $F' = f$, die Dichte ist also die Ableitung der Verteilungsfunktion. Eigenschaften von Verteilungen und Verteilungsfunktionen studieren wir in Kapitel 2.

Bilder von Zufallsvariablen sind wieder Zufallsvariablen, solange die Meßbarkeitseigenschaft nicht zerstört wird. Das ist zum Beispiel für stetige Abbildungen der Fall. Ist X eine $\mathbb{R}$-wertige Zufallsvariable und h eine stetige Abbildung, so ist $Y := h(X)$ wieder eine Zufallsvariable und

$$\{Y = y\} = \{X \in h^{-1}(y)\}.$$

Demnach ist die Verteilung von Y leicht zu errechnen durch

$$P[Y \in B] = P[X \in h^{-1}(B)].$$

B 1.9 *Münzwurf*: Bei einem zweimaligen p-Münzwurf sei $X = (X_1, X_2)$ der zufällige Vektor der den Ausgang des Wurfes beschreibt. X hat Werte in $E = \{\text{Kopf,Zahl}\}^2$. Es gilt etwa

$$P[X = (\text{Kopf, Kopf})] = P[X_1 = \text{Kopf}, X_2 = \text{Kopf})] = p^2.$$

B 1.10 *Warteschlange*: Sei wie in Beispiel 1.7 X die Wartezeit bis zur nächsten Anfrage bei einer Webseite. Für das Intervall $A = [a, b]$ folgt aus (1.1), dass

$$P[X \in A] = P[X \geq a] - P[X \geq b] = e^{-\lambda a} - e^{-\lambda b} = \int_a^b \lambda e^{-\lambda x} dx.$$

Daraus folgt, dass

$$f(x) := \lambda e^{-\lambda x}$$

die Dichte von X ist. Diese werden wir *Exponentialverteilung* zum Parameter λ nennen. Nun können wir die Frage beantworten, wie hoch die Wahrscheinlichkeit ist, dass mehr als N Anfragen in einer Sekunde eintreffen (Wie?).

Wir lernen nun noch eine wichtige Formel zum Rechnen mit Wahrscheinlichkeiten kennen. Die Folgende Formel nennt man die Einschluss- Ausschlussformel oder die Siebformel von Poincaré und Sylvester.

Satz 1.9. *Sei X eine E-wertige Zufallsvariable und $A_1, ..., A_n \subseteq E$. Dann gilt*

$$P[X \in A_1 \cup \cdots \cup A_n] = \sum_{\emptyset \neq T \subset \{1,...,n\}} (-1)^{|T|-1} P(\cap_{i \in T} A_i).$$

Diese etwas komplizierte Formel hat folgende einfache Gestalt:

$$\begin{aligned} P[X \in A_1 \cup \cdots \cup A_n] = &\sum_{i=1}^{n} P[X \in A_i] \\ &- \sum_{1 \le i < j \le n} P[X \in A_i \cap A_j] \\ &+ \sum_{1 \le i < j < k \le n} P[X \in A_i \cap A_j \cap A_k] \\ &- \cdots \\ &+ (-1)^{n-1} P[X \in A_1 \cap \cdots \cap A_n]. \end{aligned}$$

Beweis. Zunächst sei $n = 2$. Direkt aus Definition 1.5 (siehe etwa Aufgabe 1.2) folgt, dass $P[X \in A_2 \setminus (A_1 \cap A_2)] = P[X \in A_2] - P[X \in A_1 \cap A_2]$, da $A_1 \cap A_2$ und $A_2 \setminus (A_1 \cap A_2)$ disjunkt sind. Damit erhalten wir, dass

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

Wir beweisen die Aussage mittels vollständiger Induktion. Es gilt wie im Fall $n = 2$

$$\begin{aligned} P[X \in A_1 \cup \cdots \cup A_{n+1}] = &P[X \in A_1 \cup \cdots \cup A_n] \\ &+ P[X \in A_{n+1}] - P[X \in (A_1 \cap A_{n+1}) \cup \cdots \cup (A_n \cap A_{n+1})]. \end{aligned}$$

Wir wenden auf den ersten und den letzten Teil die Induktionsvoraussetzung an und erhalten

$$\begin{aligned} &= \sum_{i=1}^{n+1} P[X \in A_i] - \sum_{1 \le i < j \le n} P[X \in A_i \cap A_j] - \sum_{i=1}^{n} P[X \in A_i \cap A_{n+1}] \\ &+ \cdots + (-1)^{n-1} P[X \in A_1 \cap \cdots \cap A_n] - \sum_{i=1}^{n} (-1)^{n-2} P[X \in A_1 \cap \cdots \cap A_{i-1} \cap A_{i+1} \cap \cdots \cap A_{n+1}] \\ &- (-1)^{n-1} P[X \in A_1 \cap \cdots \cap A_{n+1}] \end{aligned}$$

und der Satz ist bewiesen. □

Das folgende Beispiel beschäftigt Mathematiker bereits seit Anfangs des 18. Jahrhunderts[4]. Es kann auch wie folgt interpretiert werden: Beim Wichteln bringt jeder ein Geschenk mit, alle Geschenke werden in gleiche Kartons verpackt. Nun werden die Kartons gemischt und jeder bekommt einen. Wie hoch ist die Wahrscheinlichkeit, dass mindestens einer sein eigenes Geschenk bekommt ?

[4]Siehe Wikipedia: http://de.wikipedia.org/wiki/Fixpunktfreie_Permutation.

B 1.11 *Fixpunkte in Permutationen*: Als Anwendung von Satz 1.9 berechnen wir die Wahrscheinlichkeit, dass es in einer Permutation der Zahlen $1, ..., n$ keine Fixpunkte gibt. (Ein Fixpunkt einer Permutation σ ist ein i mit $\sigma(i) = i$.) Alle Permutationen haben die gleiche Wahrscheinlichkeit und die Zufallsvariable X bezeichnet die aufgetretene Permutation. Sei

$$A_i := \{\sigma \in \mathcal{S}_n : \sigma(i) = i\}$$

und

$$Y := \sum_{i=1}^{n} \mathbb{1}_{\{X \in A_i\}}$$

die Anzahl der Fixpunkte von X. Klar ist, dass

$$P[X \in A_i] = \frac{(n-1)!}{n!} = \frac{1}{n},$$

weil es gerade $(n-1)!$ Permutationen gibt, die i als Fixpunkt haben. Analog gilt für $1 \leq i_1 < \cdots < i_k \leq n$

$$P[X \in A_{i_1} \cap \cdots \cap A_{i_k}] = \frac{(n-k)!}{n!} = \frac{1}{n \cdots (n-k+1)}.$$

Weiter gilt

$$\begin{aligned} P[Y > 0] &= P[X \in A_1 \cup \cdots \cup A_n] \\ &= nP[X \in A_1] - \binom{n}{2} P[A_1 \cap A_2] + \binom{n}{3} P[X \in A_1 \cap A_2 \cap A_3] - \cdots \\ &\qquad \pm \binom{n}{n} P[X \in A_1 \cap \cdots \cap A_n] \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots \pm \frac{1}{n!}. \end{aligned}$$

Damit ist die Wahrscheinlichkeit, dass X keine Fixpunkte besitzt, gerade

$$P[Y = 0] = 1 - P[Y > 0] = \frac{1}{2!} - \frac{1}{3!} + \cdots \mp \frac{1}{n!}.$$

Wir erhalten, dass für wachsendes n der Anteil der fixpunktfreien Permutationen sehr schnell gegen den Kehrwert der Eulerschen Zahl konvergiert,

$$P[Y = 0] = \sum_{k=2}^{n} \frac{(-1)^k}{k!} = \sum_{k=0}^{n} \frac{(-1)^k}{k!} \xrightarrow{n \to \infty} e^{-1}.$$

1.5 Unabhängigkeit

In den Beispielen 1.2, 1.3, 1.6, 1.7 haben wir Wahrscheinlichkeiten miteinander multipliziert. Etwa hatten wir gesagt, dass die Wahrscheinlichkeit beim p-Münzwurf für das zweimalige Auftreten von *Kopf* gerade p^2 ist. Begründet haben wir dies mit der „Unabhängigkeit" der Münzwürfe. Um dieses Konzept einzuführen, definieren wir zunächst die *gemeinsame* Verteilung von Zufallsvariablen.

Definition 1.10. Eine n-dimensionale *Zufallsvariable* ist eine Abbildung von $X : \Omega \to \mathbb{R}^n$, für welche für alle $x \in \mathbb{R}^n$ gilt, dass

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{A}. \tag{1.3}$$

Hier nutzen wir die Konvention, dass $\{X \leq x\} = \{X_1 \leq x_1, \ldots, X_n \leq x_n\}$. Analog zu eindimensionalen Zufallvariablen definieren wir die Eigenschaften diskret und stetig sowie die mehrdimensionale Verteilungsfunktion $F(x) = P[X \leq x]$.

Definition 1.11. Ist $X = (X_1, ..., X_n)$ eine n-dimensionale Zufallsvariable, so heißt die *Verteilung* von X gemeinsame Verteilung von $X_1, ..., X_n$. Die Verteilung von X_i heißt i-te *Randverteilung*. Gilt für alle Intervalle $A_1, ..., A_n \subseteq \mathbb{R}$, das dass

$$P[X \in A] = \int_A f(x_1, ..., x_n)\, dx_1 \cdots dx_n,$$

mit $A = A_1 \times \cdots \times A_n$, so heißt f *Dichte* von X.

Die gemeinsame Verteilung wird beschrieben durch die gemeinsame Verteilungsfunktion,

$$x \mapsto P[X_1 \leq x_1, \ldots, X_n \leq x_n],$$

wobei die i-te Randverteilung natürlich durch die Verteilungsfunktion von X_i beschrieben wird.

Bemerkung 1.12 (Bilder von Zufallsvariablen)**.** Auch wenn die Definition abstrakt erscheint, sind Bilder von Zufallsvariablen häufig anzutreffen. Sei etwa $(X_1, ..., X_n)$ das Ergebnis eines n-fachen Würfelwurfs. Dann beschreibt die Zufallsvariable $Z_i := h(X_i)$ mit $h(x) := 1_{x=6}$, ob im i-ten Wurf eine 6 gefallen ist oder nicht. Weiter zählt die Zufallsvariable $\ell(Z_1, ..., Z_n)$ mit $\ell(z_1, ..., z_n) := z_1 + \cdots + z_n$ die Anzahl der 6er.

B 1.12 *p-Münzwurf*: Wir betrachten einen n-fachen p-Münzwurf und setzen $q := 1 - p$. Das Ergebins des Münzwurfes sei $X = (X_1, ..., X_n)$, eine $\{0, 1\}^n$-wertige Zufallsvariable mit

$$P[X = (x_1, ..., x_n)] = p^{\sum_{i=1}^n x_i}(1 - p)^{n - \sum_{i=1}^n x_i}. \tag{1.4}$$

Mit $h(x_1, ..., x_n) = x_1 + \cdots + x_n$ erhalten wir die Augensumme: Es gilt $|h^{-1}(k)| = \binom{n}{k}$, also

$$\begin{aligned} P[h(X) = k] &= P[X \in h^{-1}(k)] \\ &= \sum_{(x_1,...,x_n) \in h^{-1}(k)} p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned} \tag{1.5}$$

In Gleichung (1.4) wurden die Wahrscheinlichkeiten multipliziert. Dies motiviert folgende, präzise Definition des zentralen Konzepts der Unabhängigkeit.

Definition 1.13. Seien $X_1, ..., X_n$ rellwertige Zufallsvaribalen. Gilt

$$P[X_1 \in A_1, ..., X_n \in A_n] = P[X_1 \in A_1] \cdots P[X_n \in A_n] \tag{1.6}$$

für alle Intervalle $A_1, ..., A_n$, so heißen $X_1, ..., X_n$ *unabhängig*. Gilt lediglich

$$P[X_i \in A_i, X_j \in A_j] = P[X_i \in A_i] \cdot P[X_j \in A_j]$$

für beliebige Intervalle A_i, A_j und $i \neq j$, so heißen $X_1, ..., X_n$ *paarweise unabhängig*.

Für diskrete Zufallsvariablen erhält man ein einfaches Kriterium: Sind $X_1, ..., X_n$ diskret, so sind sie genau dann unabhängig, wenn

$$P[X_1 = x_1, ..., X_n = x_n] = P[X_1 = x_1] \cdots P[X_n = x_n]$$

für beliebige Wahl von $x_1, ..., x_n$.

Im stetigen Fall kann man die Unabhängigkeit darauf zurückführen, dass die gemeinsame Dichte in das Produkt der Rand-Dichten zerfällt.

Lemma 1.14. *Sind $X_1, ..., X_n$ stetig mit Dichte f, so sind sie genau dann unabhängig, wenn es Funktionen $f_1, ..., f_n$ gibt mit*

$$f(x_1, ..., x_n) = f_1(x_1) \cdots f_n(x_n).$$

Dabei gilt

$$f_i(x_i) = \int f(x_1, ..., x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n.$$

Der Beweis ist eine (maßtheoretische) Übungsaufgabe.

Beispiele für unabhängige Zufallsvariable haben wir bereits kennen gelernt, so führt die n-fache Wiederholung eines p-Münzwurfes zu unabhängigen Zufallsvariablen $X_1, ..., X_n$. Folgendes Beispiel verdeutlicht den Unterschied zwischen Unabhängigkeit und dem schwächeren Begriff der paarweisen Unabhängigkeit.

B 1.13 *Paarweise Unabhängigkeit impliziert nicht Unabhängigkeit*: Sei hierzu (X_1, X_2) die zweifache Wiederholung eines 1/2-Münzwurf. Wir betrachten dann $Z = (Z_1, Z_2, Z_3)$ mit $Z_1 = X_1, Z_2 = X_2, Z_3 = 1_{X_1=X_2}$. Die Zufallsgröße Z_3 gibt also gerade an, ob die beiden Ergebnisse der Münzwürfe identisch sind oder nicht. Es gilt $P[Z_3 = 1] = \mathbf{P}[X_1 = X_2 = 0 \text{ oder } X_1 = X_2 = 1] = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$. Weiter ist

$$P[Z_1 = 1, Z_3 = 1] = P[X_1 = X_2 = 1] = \tfrac{1}{4} = P[Z_1 = 1] \cdot P[Z_3 = 1]$$

und analog für das Paar (Z_2, Z_3). Damit sind Z_1, Z_2, Z_3 paarweise unabhängig. Jedoch gilt

$$P[Z_1 = 0, Z_2 = 1, Z_3 = 1] = 0 \neq \tfrac{1}{8} = P[Z_1 = 0] \cdot P[Z_2 = 1] \cdot P[Z_3 = 1],$$

und somit sind Z_1, Z_2, Z_3 *nicht* unabhängig.

2. Verteilungen

Einen besonderen Stellenwert haben reellwertige Zufallsvariablen und deren Verteilungen. Die Verteilungsfunktion einer Zufallsvariable ist $x \mapsto F_X(x) = P[X \leq x]$, siehe Definition 1.8. Sie ist monoton wachsend und deswegen in einem geeigneten Sinne invertierbar, vergleiche auch Definition 2.10.

Lemma 2.1. *Sei X eine (diskrete oder stetige), reellwertige Zufallsvariable. Dann gilt*

$$\lim_{x\to\infty} F_X(x) = 1, \qquad \lim_{x\to-\infty} F_X(x) = 0.$$

Ist X eine stetige Zufallsvariable mit Dichte $f(x)dx$, so gilt

$$F_X(x) = \int_{-\infty}^{x} f(z)dz.$$

Wie bereits erwähnt gilt also $F' = f$, falls F differenzierbar ist (Geben Sie ein Beispiel an, bei welchem F nicht differenzierbar ist !).

Beweis. Wir zeigen die ersten beiden Aussagen für diskrete Zufallsvariable. Sei $E \subseteq \mathbb{R}$ der diskrete Wertebereich von X. Dann gilt $P(\Omega) = \sum_{z\in E} P[X = z] = 1$. Es folgt, dass $P[X \leq x] = \sum_{z\in E, z\leq x} P[X = z] \xrightarrow{x\to\infty} 1$, da andernfalls die Gesamtsumme von $P[X = z]$ nicht konvergiert. Die zweite Gleichung folgt analog.

Zunächst erhält man mit Definition 1.5 (ii), dass

$$P[X \leq x] = P[X \in (-\infty, x]] = \lim_{n\to\infty} P[X \in (-n, x]]$$

und mit dem Satz der majorisierten Konvergenz folgt die Behauptung, da

$$\lim_{n\to\infty} \int_{-n}^{x} f(z)dz = \int_{-\infty}^{x} f(z)dz. \qquad \square$$

2.1 Laplace-Experimente

Man stelle sich einen Würfelwurf mit einem fairen Würfel vor. Hier sind alle möglichen Ausgänge gleichwahrscheinlich. In einem solchen Fall spricht man auch von einem Laplace-Experiment. Solche Experimente führen zu uniformen Verteilungen, die wir nun kennen lernen.

Definition 2.2. Sei X eine E-wertige Zufallsvariable und E endlich. Ist $x \mapsto P[X = x]$ konstant, so nennen wir X (diskret) *uniform* verteilt.

Wir schreiben $E = \{e_1, \dots, e_n\}$. Dann folgt, dass $1 = \sum_{i=1}^n P[X = e_i] = \sum_{i=1}^n P[X = e_1]$, also $P[X = e_i] = n^{-1}$. Für eine Teilmenge $A \subset E$ folgt, dass

$$P[X \in A] = \frac{|A|}{|E|}.$$

Definition 2.3. Sei X eine E-wertige Zufallsvariable mit $E = [a, b]$. Gilt für $a \le c \le d \le b$, dass

$$P[X \in (c, d)] = \frac{d - c}{b - a},$$

so heißt X *uniform* verteilt auf $[a, b]$ und wir schreiben auch $X \sim U([a, b])$.

Lemma 2.4. *Gilt $X \sim U([a, b])$, so hat X Dichte und Verteilungsfunktion*

$$f(x) = \frac{1}{b - a} \mathbb{1}_{\{[a,b]\}}(x)$$
$$F(x) = \mathbb{1}_{\{[a,b]\}}(x) \frac{x - a}{b - a} + \mathbb{1}_{\{x>b\}}.$$

Beweis. Zunächst berechnen wir, dass

$$P[X \le x] = P[X \in (a, x)] = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a} = \frac{1}{b-a} \int_a^x dz, & a \le x < b, \\ 1, & x \ge b. \end{cases}.$$

und die Behauptung über F folgt. Durch Ableiten erhält man die Dichte. □

2.2 Die Binomial-Verteilung

Denkt man an ein n-mal unabhängig wiederholt durchgeführtes Spiel, so ist es möglich, zwischen 0 und n-mal zu gewinnen. Die Anzahl der Gewinne ist eine Zufallsvariable, deren Verteilung wir nun als Binomialverteilung kennen lernen, vergleiche auch Beispiel 1.12.

Definition 2.5. Sei X eine E-wertige Zufallsvariable, $E = \{0, \dots, n\}$ und $p \in (0, 1)$. Gilt

$$P[X = k] = \binom{n}{k} p^k (1 - p)^{n-k}, \tag{2.1}$$

so heißt die Verteilung von X *Binomialverteilung* mit Parametern n und p und wir schreiben $X \sim B(n, p)$.

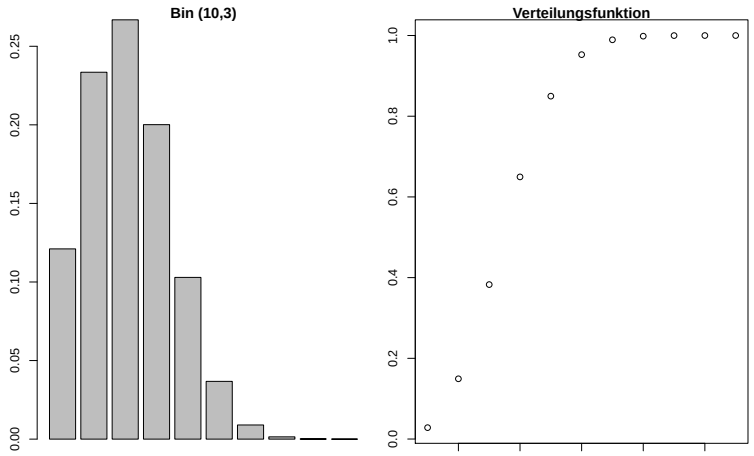


Abbildung 2.1: Verteilungsgewichte und Verteilungsfunktion von $B(10, 0.3)$.

Den Spezialfall $n = 1$ nennen wir auch *Bernoulliverteilung*.

Lemma 2.6. *Die Summe von n unabhängigen Bernoulli-Zufallsvariablen ist Binomialverteilt.*

Beweis. Folgt exakt wie in (1.5). □

2.3 Kenngrößen von Zufallsvariablen

Definition 2.7. Sei X eine E-wertige Zufallsvariable.

(i) Ist E abzählbar, so heißt

$$E[X] := \sum_{x \in E} x P[X = x]$$

Erwartungswert von X, falls die Summe konvergiert.

(ii) Ist X stetig mit Dichte f, so heißt

$$E[X] := \int x f(x) dx$$

Erwartungswert von X, falls das Integral existiert.

In den Fällen wo die Summe konvergiert bzw. das Integral existiert nennen wir die Zufallsvariable X *integrierbar*.

B 2.1 *Uniforme und Binomialverteilung*: Wir berechnen die Erwartungswerte wie folgt:

(i) Sei $X \sim B(n,p)$. Dann ist mit $q := 1-p$

$$\begin{aligned} E[X] &= \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k} = np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}q^{n-k} \\ &= np\sum_{k=0}^{n-1}\binom{n-1}{k}p^k q^{n-1-k} = np. \end{aligned} \tag{2.2}$$

(ii) Sei $X \sim U([a,b])$. Dann ist

$$E[X] = \frac{1}{b-a}\int_a^b x dx = \frac{1}{2}\frac{b^2-a^2}{b-a} = \frac{a+b}{2}.$$

Ist $h : E \to \mathbb{R}$ eine messbare Funktion, so ist $h(X)$ wieder eine Zufallsvariable. Der folgende *Transfomrationssatz* zeigt, wie man den Erwartungswert von $h(X)$ berechnen kann.

Lemma 2.8 (Transformationssatz)**.** *Sei X eine E-wertige, diskrete Zufallsvariable und $h : E \to \mathbb{R}$. Dann gilt*

$$E[h(X)] = \sum_{y\in h(E)} yP[h(X)=y] = \sum_{x\in E} h(x)P[X=x],$$

falls die Summe existiert.

Beweis. Zunächst ist $Y = h(X)$ eine diskrete Zufallsvariable mit Werten in $E' = h(E) = \{h(x) : x \in E\}$. Es gilt nach Definition des Erwartungswertes

$$\begin{aligned} E[Y] &= \sum_{y\in h(E)} yP[h(X)=y] = \sum_{y\in h(E)} y\sum_{x\in h^{-1}(y)} P[X=x] \\ &= \sum_{y\in h(E)}\sum_{x\in h^{-1}(y)} h(x)P[X=x] = \sum_{x\in E} h(x)P[X=x] \end{aligned}$$

und die Behauptung folgt. □

Ähnlich erhält man auch im stetigen Fall, dass

$$E[h(X)] = \int h(x)f(x)\,dx.$$

Der Ewartungswert ererbt von der Summe und von dem Integral die Eigenschaft der *Linearität*.

Satz 2.9. *Seien X, Y integrierbare Zufallsvariablen und $a, b \in \mathbb{R}$.*

(i) Dann gilt $E[aX + bY] = aE[X] + bE[Y]$ und

(ii) aus $X \leq Y$ folgt $E[X] \leq E[Y]$.

Beweis. Wir zeigen beide Aussagen nur für diskrete Zufallsvariablen. Es gilt

$$\begin{aligned} E[aX + bY] &= \sum_{x,y \in E} (ax + by) P[X = x, Y = y] \\ &= a \sum_{x,y \in E} x P[X = x, Y = y] + b \sum_{x,y \in E} y P[X = x, Y = y] \\ &= a \sum_{x \in E} P[X = x] + b \sum_{y \in E} y P[Y = y] = aE[X] + bE[Y]. \end{aligned}$$

Für (ii) genügt es zu zeigen, dass $E[X] \geq 0$ für $X \geq 0$. (Dann ist nämlich $E[Y] - E[X] = E[Y - X] \geq 0$.) Gelte also $X \geq 0$. Es folgt, dass

$$E[X] = \sum_{x \in E} x P[X = x] = \sum_{x \in E, x \geq 0} x P[X = x] \geq \sum_{x \in E, x \geq 0} 0 \cdot P[X = x] = 0. \qquad \square$$

Die Umkehrfunktion der Verteilungsfunktion, die Quantilfunktion, ist von besonderem Interesse.

Definition 2.10. Sei $\alpha \in (0, 1)$. Jedes $q_\alpha \in \mathbb{R}$ mit

$$P[X < q_\alpha] \leq \alpha \leq P[X \leq q_\alpha]$$

heißt *α-Quantil* von (der Verteilung von) X. Jedes 0.5-Quantil heißt *Median* von (der Verteilung von) X.

In der Statistik beobachtet man oft Daten von unabhängigen und identischen Wiederholungen eines Zufallsexperiments. In diesem Sinne kommt jeder einzelnen Beobachtung das gleiche Gewicht zu. Das motiviert die folgende empirische Verteilung.

Definition 2.11. Sind die Punkte $E = \{x_1, \ldots, x_n\}$ gegeben, so heißt die Verteilung auf E mit

$$P(X = x_i) = \frac{1}{n}, \qquad x_i \in E$$

die *empirische Verteilung*.

B 2.2 *Median und Mittelwert*: Der Erwartungswert der empirischen Verteilung ist

$$E[X] = \sum_{i=1}^{n} x_i P[X = x_i] = \frac{1}{n} \sum_{i=1}^{n} x_i =: \bar{X},$$

der *arithmetische Mittelwert* der Daten. Die Funktion

$$\bar{F}(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}}, \qquad x \in \mathbb{R}$$

heißt *empirische Verteilungsfunktion*. Ist n ungerade und sind alle x_i verschieden, so ist

$$q_{0.5} = x_{(n+1)/2:n},$$

ein Quantil, der so genannte *Median*; Hier haben wir mit $x_{1:n} \leq x_{2:n} \leq ... \leq x_{n:n}$ die Ränge, also die geordneten Datenpunkte bezeichnet.

Lemma 2.12. *Sei X eine $\mathbb{Z}_+$-wertige Zufallsvariable. Dann gilt*

$$E[X] = \sum_{x=0}^{\infty} P[X > x],$$
$$E[X(X-1)] = 2 \cdot \sum_{x=0}^{\infty} x \cdot P[X > x]$$

falls jeweils die Summe existiert.

Beweis. Für die erste Behauptung schreiben wir

$$E[X] = \sum_{x=0}^{\infty} xP[X = x] = \sum_{x=1}^{\infty} \sum_{y=1}^{x} P[X = x] = \sum_{y=1}^{\infty} \sum_{x=y}^{\infty} P[X = x]$$
$$= \sum_{y=1}^{\infty} P[X \geq y] = \sum_{y=0}^{\infty} P[X > y]$$

wobei die Umordnung der Doppelsumme wegen der absoluten Konvergenz der Reihe möglich ist. Für die zweite Behauptung ist analog

$$E[X(X-1)] = 2 \sum_{x=1}^{\infty} \frac{x(x-1)}{2} \cdot P[X = x] = 2 \sum_{x=1}^{\infty} \sum_{y=0}^{x-1} y \cdot P[X = x]$$
$$= 2 \sum_{y=0}^{\infty} \sum_{x=y+1}^{\infty} y \cdot P[X = x] = 2 \sum_{y=0}^{\infty} y \cdot P[X > y]. \qquad \square$$

Analog zur stetigen uniformen Verteilung können wir auch für den diskreten Fall den Erwartungswert direkt ausrechnen.

B 2.3 *Diskrete uniforme Verteilung*: Sei X uniform auf $\{k, ..., \ell\}$ verteilt. Dann ist

$$\begin{aligned} E[X] &= \sum_{x=k}^{\ell} x \frac{1}{\ell - k + 1} = \frac{1}{\ell - k + 1} \Big(\sum_{x=0}^{\ell} x - \sum_{x=0}^{k-1} x \Big) = \frac{\binom{\ell+1}{2} - \binom{k}{2}}{\ell - k + 1} \\ &= \frac{(\ell - k + 1)(\ell + k)}{2(\ell - k + 1)} = \frac{\ell + k}{2}. \end{aligned}$$

Genauso können wir schreiben

$$\begin{aligned} E[X] &= \sum_{x=0}^{\ell} P[X > x] = \sum_{x=0}^{k-1} 1 + \sum_{x=k}^{\ell} \frac{\ell - x}{\ell - k + 1} = k + \frac{1}{\ell - k + 1} \sum_{x=0}^{\ell-k} x \\ &= k + \frac{\binom{\ell-k+1}{2}}{\ell - k + 1} = k + \frac{\ell - k}{2} = \frac{\ell + k}{2}. \end{aligned}$$

Mit Hilfe der Linearität lässt sich der Erwartungswert einer Binomialverteilten Zufallsvariable sehr leicht ausrechnen.

B 2.4 *p-Münzwurf*: Ist $X = (X_1, ..., X_n)$ ein p-Münzwurf, so wissen wir bereits, dass $Y := X_1 + \cdots + X_n \sim B(n, p)$. Also muss auch

$$E[Y] = E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n] = nE[X_1] = nP[X_1 = 1] = np$$

in Übereinstimmung mit (2.2).

Definition 2.13. Sei X eine integrierbare Zufallsvariable mit Erwartungswert $\mu := E[X]$. X heißt *quadratintegrierbar*, falls X^2 integrierbar ist. In diesem Fall nennen wir die Größe

$$\sigma^2 := \mathrm{Var}[X] = E[(X - \mu)^2]$$

Varianz von X und

$$\sigma := \sqrt{\mathrm{Var}[X]}$$

Standardabweichung von X.

Bemerkung 2.14. Eine quadratintegrierbare Zufallsvariable ist immer integrierbar, die Umkehrung gilt allerdings nicht. Für diskrete Zufallsvariablen ist also

$$\mathrm{Var}[X] = \sum_{x \in E} (x - \mu)^2 P[X = x],$$

für stetige mit Dichte $f(x)dx$ gerade

$$\mathrm{Var}[X] = \int (x-\mu)^2 f(x)dx$$

falls Summe und Integral existieren.

Satz 2.15 (Rechenregeln für die Varianz)**.** *Sei X eine quadratintegrierbare Zufallsvariable. Dann gilt*

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 = E[X(X-1)] + E[X] - E[X]^2.$$

Beweis. Die zweite Gleichheit ist klar wegen der Linearität des Erwartungswertes. Sei $\mu := E[X]$. Dann schreiben wir

$$\mathrm{Var}[X] = E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - \mu^2.$$

Damit ist alles gezeigt. □

B 2.5 *Uniforme und Binomialverteilung*: Wir berechnen die Varianzen in diesen beiden Fällen:

(i) Sei $X \sim B(n,p)$ und $q := 1-p$. Dann ist

$$\begin{aligned} E[X(X-1)] &= \sum_{k=0}^{n} k(k-1)\binom{n}{k} p^k q^{n-k} \\ &= n(n-1)p^2 \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2} q^{n-k} = n(n-1)p^2, \end{aligned}$$

also

$$\mathrm{Var}[X] = n(n-1)p^2 + np - (np)^2 = npq.$$

(ii) Sei $X \sim U([a,b])$. Dann ist

$$\begin{aligned} \mathrm{Var}[X] &= \frac{1}{b-a}\int_a^b x^2 dx - \frac{(a+b)^2}{4} = \frac{1}{3}\frac{b^3-a^3}{b-a} - \frac{(a+b)^2}{4} \\ &= \frac{4b^2+4ab+4a^2-3b^2-6ab-3b^2}{12} = \frac{(b-a)^2}{12}. \end{aligned}$$

Definition 2.16. Seien X, Y quadratintegrierbare Zufallsvariablen mit $\mu_X := E[X], \mu_Y := E[Y]$ sowie Varianzen $\sigma_X^2 := \mathrm{Var}[X], \sigma_Y^2 := \mathrm{Var}[Y]$. Dann existiert

$$\mathrm{Cov}[X,Y] := E[(X-\mu_X)(Y-\mu_Y)]$$

und heißt *Kovarianz* von X und Y. Gilt $\mathrm{Cov}[X,Y] = 0$, so nennen wir X und Y *unkorreliert*. Der *Korrelationskoeffizient* von X und Y ist

$$\mathrm{Kor}[X,Y] := \frac{\mathrm{Cov}[X,Y]}{\sigma_X \cdot \sigma_Y}.$$

Lemma 2.17 (Rechenregeln für Covarianzen)**.** *Seien X, Y, Z quadratintegrierbare Zufallsvariablen und $a, b \in \mathbb{R}$. Dann gilt*

$$\begin{aligned}
\operatorname{Cov}[X, Y] &= \operatorname{Cov}[Y, X], \\
\operatorname{Var}[X] &= \operatorname{Cov}[X, X], \\
\operatorname{Cov}[X, Y] &= E[XY] - E[X]E[Y], \\
\operatorname{Cov}[X, aY + bZ] &= a \operatorname{Cov}[X, Y] + b \operatorname{Cov}[X, Z] \\
\operatorname{Cov}[X, a] &= 0.
\end{aligned}$$

Beweis. Wir setzen $\mu_X := E[X], \mu_Y := E[Y]$. Die ersten beiden Gleichungen sind klar. Für die dritte Gleichung schreiben wir

$$E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X \mu_Y = E[XY] - \mu_X \mu_Y.$$

Die vierte folgt aus der Definition der Covarianz und der Linearität des Erwartungswertes. Für die letzte Gleichung beobachte man, dass für $Y = a$ der Erwartungswert $\mu_Y = a$ ist, also folgt, dass

$$E[(X - \mu_X)(a - \mu_a)] = 0. \qquad \square$$

Lemma 2.18 (Cauchy–Schwartz Ungleichung)**.** *Es gilt*

$$(E[XY])^2 \leq (E[|XY|])^2 \leq E[X^2]E[Y^2].$$

Insbesondere ist

$$-1 \leq \operatorname{Kor}[X, Y] \leq 1. \tag{2.3}$$

Beweis. Die erste Ungleichung folgt aus Satz 2.9, da $-|XY| \leq XY \leq |XY|$ gilt und somit

$$-E[|XY|] \leq E[XY] \leq E[|XY|].$$

Daraus folgt direkt $(E[XY])^2 \leq (E[|XY|])^2$. Für die zweite Gleichung genügt es den Fall $E[X^2] > 0$ zu betrachten. Andernfalls muss nämlich $P[X = 0] = 1$ gelten und dann ist nichts zu zeigen. Sei also $E[X^2] > 0$. Für jedes $c \in \mathbb{R}$ gilt $0 \leq (-c|X| + |Y|)^2 = c^2X^2 - 2c|XY| + Y^2$. Insbesondere gilt für $c := E[|XY|]/E[X^2]$

$$\begin{aligned}
0 &\leq \frac{E[|XY|]^2}{E[X^2]} - 2\frac{E[|XY|]^2}{E[X^2]} + E[Y^2] \\
&= E[Y^2] - \frac{E[|XY|]^2}{E[X^2]},
\end{aligned}$$

woraus die erste Behauptung direkt folgt. Für Gleichung (2.3) betrachten wir $X' = X - E[X]$ und $Y' = Y - E[Y]$ so dass die Gleichung direkt aus der Definition 2.16 folgt. $\square$

Bemerkung 2.19. Meistens verwendet man die Cauchy-Schwartz-Ungleichung zur Abschätzung durch die zweiten Momente in folgender Version:

$$E[|XY|] \leq \big(E[X^2]E[Y^2]\big)^{1/2} \tag{2.4}$$

Eine wichtige Beobachtung ist, dass aus Unabhängigkeit Unkorrelliertheit folgt, wie der folgende Satz zeigt. Die Umkehrung allerdings ist im Allgemeinen nicht richtig, gilt aber unter der zusätzlichen Annahme, dass eine *gemeinsame* Normalverteilung vorliegt.

Satz 2.20. *Seien X, Y quadratintegrierbare und unabhängige Zufallsvariablen. Dann sind X, Y auch unkorreliert, d.h.* $\mathrm{Cov}[X, Y] = 0$.

Beweis. Wir zeigen die Behauptung für diskrete Zufallsvariablen, der stetige Fall ist eine gute Übungsaufgabe. Es gilt

$$\begin{aligned} E[XY] &= \sum_{x,y} xyP[X = x, Y = y] = \sum_{x} xP[X = x] \sum_{y} yP[Y = y] \\ &= E[X]E[Y]. \end{aligned}$$

□

Der folgende Satz beinhaltet die wichtige Regel von Bienaymé für die Varianz einer Summe.

Satz 2.21. *Seien $X_1, ..., X_n$ quadratintegrierbare Zufallsvariablen. Dann gilt*

$$\mathrm{Var}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2 \sum_{1 \leq i < j \leq n} \mathrm{Cov}[X_i, X_j].$$

Sind $(X_1, ..., X_n)$ paarweise unkorreliert, so gilt

$$\mathrm{Var}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathrm{Var}[X_i].$$

Beweis. Wir verwenden Lemma 2.17 und erhalten

$$\begin{aligned} \mathrm{Var}\Big[\sum_{i=1}^{n} X_i\Big] &= \mathrm{Cov}\Big[\sum_{i=1}^{n} X_i, \sum_{i=j}^{n} X_j\Big] = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}[X_i, X_j] \\ &= \sum_{i=1}^{n} \mathrm{Cov}[X_i, X_i] + \sum_{i=1}^{n}\sum_{j=1}^{i-1} \mathrm{Cov}[X_i, X_j] + \sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{Cov}[X_i, X_j] \\ &= \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2 \sum_{1 \leq i < j \leq n} \mathrm{Cov}[X_i, X_j]. \end{aligned}$$

□

Es gibt eine ziemlich zuverlässige Möglichkeit von Menschen aufgeschriebene "zufällige" Nullen und Einsen von einer echten Zufallsfolge zu unterscheiden: Beim Aufschreiben tendiert man dazu, eine Regelmäßigkeit einzubauen und vermeidet längere Runs.

B 2.6 *Runs in einem p-Münzwurf*: Sei $X = (X_1, ..., X_n)$ ein p-Münzwurf. Ein *Run* ist ein maximaler Block aus 0ern oder 1ern, also enthält 1100010 genau vier Runs. Es gibt immer mindestens einen Run. Sei Y die Anzahl der Runs in X und Z die Anzahl der Runs in $(X_2, \ldots, X_n)$. Dann gilt

$$Y = 1 + Z = 1 + \sum_{i=2}^{n} \mathbb{1}_{\{X_i \neq X_{i-1}\}}.$$

Wir setzen $Z_i := \mathbb{1}_{\{X_i \neq X_{i-1}\}}$ und beobachten, dass die (Z_i) nicht paarweise unabhängig sind. Wie können wir Erwartungswert und Varianz von Y ausrechnen? Zunächst ist mit $q = 1 - p$

$$E[Z_i] = P[X_i \neq X_{i-1}] = 2pq$$

und es folgt $E[Y] = 1 + 2pq(n-1)$. Weiterhin ist

$$\mathrm{Var}[Z_i] = E[Z_i] - E[Z_i]^2 = 2pq - (2pq)^2 = 2pq(1-2pq).$$

Für $2 \leq i \leq n-1$ ist außerdem

$$E[Z_i Z_{i+1}] = P[X_{i+1} \neq X_i \neq X_{i-1}] = pqp + qpq = pq.$$

Offenbar sind $\mathbb{1}_{E_i}, \mathbb{1}_{E_j}$ für $|i-j| > 1$ unabhängig. Damit gilt

$$\begin{aligned}\mathrm{Var}[Y] &= \sum_{i=2}^{n} \mathrm{Var}[Z_i] + 2\sum_{i=2}^{n-1} \mathrm{Cov}[Z_i, Z_{i+1}] \\ &= 2pq(1-2pq)(n-1) + pq(1-4pq)(n-2).\end{aligned}$$

Wir analysieren etwas genauer den Fall $p = q$: Wie hoch ist die Wahrscheinlichkeit für k Runs? Es gibt mindestens einen Run und wir betrachten die Zufallsvariable $Z \in \{0, \ldots, n-1\}$. Es gibt 2^n Möglichkeiten Nullen und Einsen anzuordnen und alle sind gleich wahrscheinlich, so dass es sich um ein Laplace-Experiment handelt. Wir identifizieren jeden Run mit seiner Endstelle und es gibt

$$\binom{n-1}{k}$$

Möglichkeiten, k Runs auf die $n-1$ Stellen zu verteilen. Wir müssen noch mit zwei Multiplizieren, um die Möglichkeiten $X_1 = 0$ und $X_1 = 1$ zu berücksichtigen. Wir erhalten

$$P[Z = k] = 2 \cdot \frac{1}{2^n} \binom{n-1}{k},$$

also ist Z Binomial$(n-1, p)$-verteilt. Wir erhalten $E[Y] = 1 + E[Z] = 1 + (n-1)\frac{1}{2}$, was sich mit den obigen Berechnungen deckt.

Berechnen Sie als Übungsaufgabe mit $n = 50$ die Wahrscheinlichkeit, dass Sie mehr als 32 Runs erhalten.

3. Wichtige Verteilungen

In diesem Kapitel stellen wir einige der wichtigsten Verteilungsklassen vor. Dies ist eine *kleine* Auswahl aus dem Zoo der Verteilungen, man schaue nur einmal in die Bände (**?**; **?**).

3.1 Die hypergeometrische Verteilung

Gegeben sei folgende Situation: In einer Urne befinden sich insgesamt N Kugeln, wovon w weiß sind. Wir ziehen n Kugeln blind und ohne Zurücklegen heraus und setzen

$$X := \text{ Anzahl der weißen Kugeln in der Stichprobe.}$$

Was ist die Verteilung von X? Um die Wahrscheinlichkeit $P[X = k]$ zu berechnen, stellen wir uns vor, dass die weißen Kugeln von 1 bis w und die anderen Kugeln von $w + 1$ bis N nummeriert sind. Insgesamt ist jede mögliche Ziehung (wenn man die Kugeln nur nach Nummern unterscheidet) gleichwahrscheinlich und es gibt genau $\binom{N}{n}$ mögliche Ziehungen nach Lemma 1.1 (wir unterscheiden ja die Reihenfolge nicht). Wie viele dieser Möglichkeiten führen nun gerade dazu, dass man k weiße Kugeln zieht? Hierzu muss man gerade k der w weißen und $n-k$ der $N-w$ andersfarbigen Kugeln ziehen. Deshalb gibt es hierfür $\binom{w}{k}\binom{N-w}{n-k}$ Möglichkeiten. Weil alle Möglichkeiten gleich wahrscheinlich sind, gilt also

$$P[X = k] = \frac{\binom{w}{k}\binom{N-w}{n-k}}{\binom{N}{n}}.$$

Definition 3.1. Sei $n, N \in \mathbb{N}$ und $w \in \mathbb{Z}_+$ mit $n, w \leq N$. Eine $\{0, ..., n\}$-wertige Zufallsvariable X heißt *hypergeometrisch verteilt* mit n, N, w, wenn

$$P[X = k] = \frac{\binom{w}{k}\binom{N-w}{n-k}}{\binom{N}{n}}, \qquad k = 0, 1, ..., n. \tag{3.1}$$

Wir schreiben dann auch $X \sim \text{Hyp}(n, N, w)$.

Satz 3.2. *Sei $X \sim Hyp(n, N, w)$. Dann gilt mit $p := \frac{w}{N}$, dass*

$$E[X] = np, \qquad \text{Var}[X] = np(1-p)\Big(1 - \frac{n-1}{N-1}\Big);$$

in der letzten Gleichung nutzen wir die Konvention $0/0 = 0$.

Beweis. Sei

$$Z_i := \mathbb{1}_{\{i\text{-te gezogene Kugel ist weiß}\}},$$

also

$$X = \sum_{i=1}^{n} Z_i.$$

Klar ist, dass $P[Z_i = 1] = p$, $i = 1, ..., n$ und damit

$$E[X] = \sum_{i=1}^{n} E[Z_i] = np.$$

Weiter gilt für $i \neq j$ und $q = 1 - p$, dass

$$\mathrm{Cov}[Z_i, Z_j] = P[Z_i = Z_j = 1] - p^2 = p\Big(\frac{w-1}{N-1} - \frac{w}{N}\Big) = -p\frac{N-w}{N(N-1)} = -pq\frac{1}{N-1},$$

also nach Proposition 2.21

$$\begin{aligned}\mathrm{Var}[X] &= \sum_{i=1}^{n} \mathrm{Var}[Z_i] + \sum_{1\le i\neq j\le n} \mathrm{Cov}[Z_i, Z_j] \\ &= npq - n(n-1)pq\frac{1}{N-1} = npq\Big(1 - \frac{n-1}{N-1}\Big).\end{aligned}$$ □

Bemerkung 3.3 (Alternative Berechnung)**.** Wir können Erwartungswert und Varianz von X alternativ auch direkt berechnen. Hierzu bemerken wir zunächst, dass

$$\sum_{k=1}^{n} \binom{w}{k}\binom{N-w}{n-k} = \binom{N}{w},$$

andernfalls würden sich die Wahrscheinlichkeiten in (3.1) nicht zu 1 addieren. Nun schreiben wir

$$\begin{aligned}E[X] &= \sum_{k=0}^{n} k\frac{\binom{w}{k}\binom{N-w}{n-k}}{\binom{N}{n}} = w\sum_{k=1}^{n} \frac{\binom{w-1}{k-1}\binom{N-w}{n-k}}{\binom{N}{n}} = w\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = n\frac{w}{N} = np, \\ E[X(X-1)] &= \sum_{k=0}^{n} k(k-1)\frac{\binom{w}{k}\binom{N-w}{n-k}}{\binom{N}{n}} \\ &= w(w-1)\sum_{k=2}^{n} \frac{\binom{w-2}{k-2}\binom{N-w}{n-k}}{\binom{N}{n}} = w(w-1)\frac{\binom{N-2}{n-2}}{\binom{N}{n}} = np\tfrac{(n-1)(w-1)}{N-1},\end{aligned}$$

also

$$\begin{aligned}\mathrm{Var}[X] &= E[X(X-1)] + E[X] - E[X]^2 = np\tfrac{(n-1)(w-1)}{N-1} + np - (np)^2 \\ &= np\Big(1 - \frac{nw}{N} + \frac{(n-1)(w-1)}{N-1}\Big) \\ &= np\frac{N(N-1) + nw - (n+w-1)N}{N(N-1)} = npq\frac{N-n}{N-1}.\end{aligned}$$

Bemerkung 3.4 (Ziehen mit und ohne Zurücklegen)**.** Eine Alternative zum Ziehen *ohne* Zurücklegen ist natürlich das Ziehen *mit* Zurücklegen. Führt man dies n-mal durch, und fragt sich bei N Kugeln in der Urne, von denen genau w weiß sind, nach der Anzahl Y der weißen Kugeln in der Stichprobe, so erhält man $Y \sim B(n,p)$ mit $p = \frac{w}{N}$. Damit ist also

$$E[X] = E[Y] = np,$$

sowie

$$\mathrm{Var}[X] = npq\Big(1 - \frac{n-1}{N-1}\Big), \qquad \mathrm{Var}[Y] = npq.$$

Dies kann man etwa so interpretieren: Ist N groß, enthält die Urne also viele Kugeln, von denen etwa ein Anteil p weiß ist, und zieht man $n \ll N$ Kugeln ohne oder mit Zurücklegen heraus, so sind die Varianzen (und auch die Verteilungen) recht ähnlich.

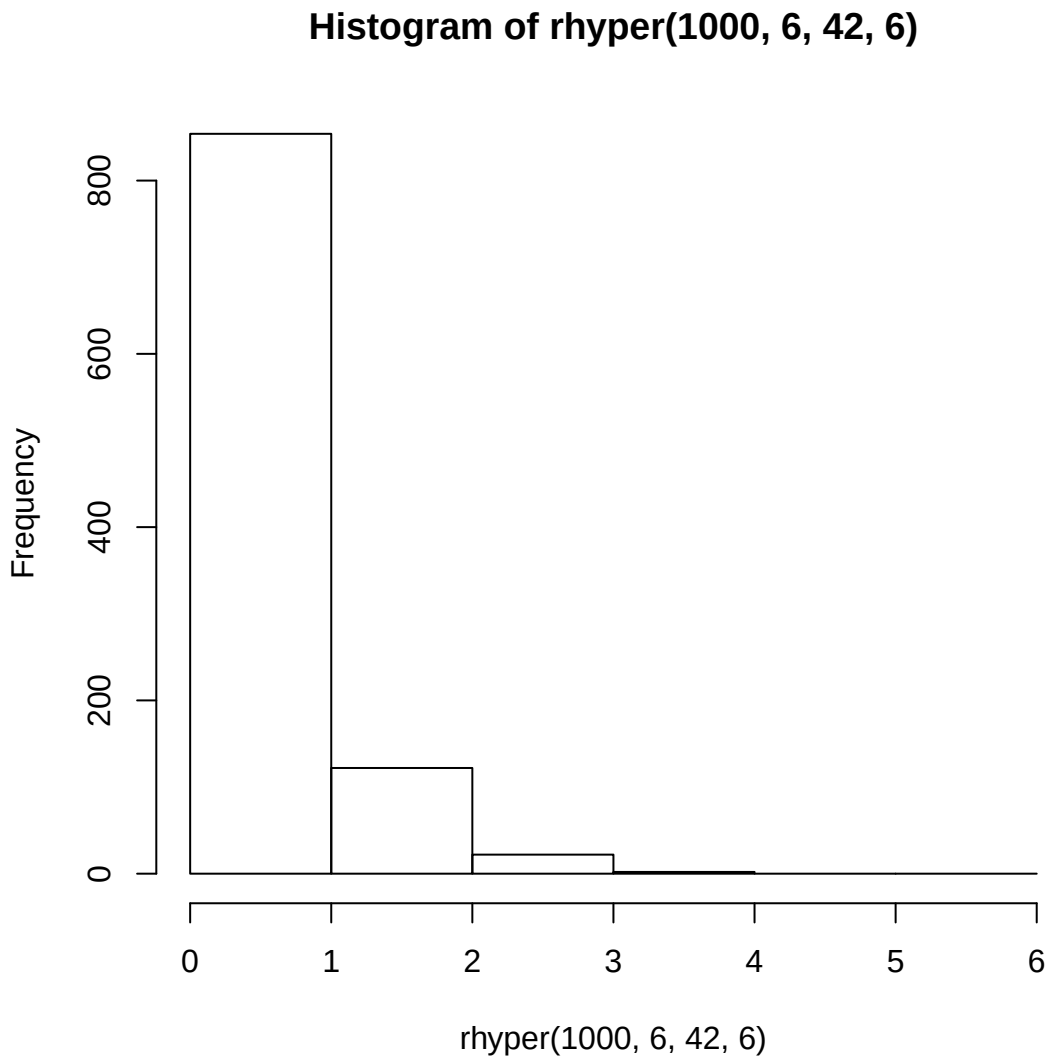
B 3.1 *Lotto*: Die Gewinnwahrscheinlichkeiten im Lotto kann man durch die hypergeometrische Verteilung beschreiben. Hier gibt es $g = 49$ Kugeln, von denen $w = 6$ die Gewinnzahlen sind. Füllt man einen Spielzetten aus, so kreuzt man $n = 6$ Zahlen an. Deshalb ist z.B. die Wahrscheinlichkeit, dass gerade $k = 4$ Zahlen richtig sind gerade

$$\frac{\binom{6}{4}\binom{43}{2}}{\binom{49}{6}} \approx 9.69 \cdot 10^{-4}.$$

B 3.2 *Simulation in R*: Es ist sehr einfach, solche Beispiele in R `www.r-project.org` zu implementieren. Der Code ist lediglich

```
> rhyper(100,6,43,6)     # Zieht 100 mal aus 49=6 (weiss)+43 jeweils 6 Kugeln
                         # und zählt weiß

> hist(rhyper (1000,6,42,6),breaks=c(0,1,2,3,4,5,6))
                         # Ergibt ein einfaches Histogramm:
```



3.2 Die Poisson-Verteilung und das Gesetz der kleinen Zahlen

Wir stellen uns (wie bei der Einführung der Binomialverteilung) vor, dass ein Zufallsexperiment unabhängig n-mal hintereinander ausgeführt wird, wobei in jedem Versuch mit Wahrscheinlichkeit p ein Erfolg zu verzeichnen ist. Wir betrachten nun den Fall großer n und kleiner p. Man denke hierbei etwa daran, dass viele Leute ein Glücksspiel (z.B. Lotto)

spielen, das für jede Person nur eine sehr kleine Gewinnwahrscheinlichkeit hat. Nachwievor ist natürlich die Anzahl der Gewinner $X \sim B(n,p)$, also in Erwartung $E[X] = np$. Es stellt sich nun heraus, dass man die Verteilung von X approximieren kann.

Satz 3.5 (Gesetz der kleinen Zahlen)**.** *Sei X_n verteilt nach $B(n,p_n)$ für $n = 1,2,...,$ so dass*

$$E[X_n] = n \cdot p_n \xrightarrow{n\to\infty} \lambda > 0.$$

Dann gilt

$$\lim_{n\to\infty} P[X_n = k] = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Beweis. Wir schreiben direkt

$$\begin{aligned} P[X_n = k] &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{1}{k!}(np_n)^k \Big(1 - \frac{np_n}{n}\Big)^n (1-p_n)^{-k}. \end{aligned}$$

Für jedes k gilt nun $\frac{n(n-1)\cdots(n-k+1)}{n^k} \xrightarrow{n\to\infty} 1$, $(1-np_n/n)^n \xrightarrow{n\to\infty} e^{-\lambda}$ und $(1-p_n)^{-k} \xrightarrow{n\to\infty} 1$. Insgesamt ergibt sich damit die Behauptung. □

Die so entdeckte Verteilung spielt eine wichtige Rolle und wir nennen Sie die Poisson-Verteilung.

Definition 3.6. Sei $\lambda \geq 0$ und X eine Zufallsvariable mit Werten in $\{0,1,2,\dots\}$ und

$$P[X = k] = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0,1,2,\dots$$

Dann heißt X *Poisson-verteilt* zum Parameter λ. Wir schreiben $X \sim$Poi(λ).

Satz 3.7. *Sei $\lambda \geq 0$ und $X \sim Poi(\lambda)$. Dann gilt*

$$E[X] = \mathrm{Var}[X] = \lambda.$$

Beweis. Wir schreiben, wegen $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$

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

Mit Satz 2.21 folgt ebenso

$$\begin{aligned}\mathrm{Var}[X] &= E[X(X-1)] + E[X] - E[X]^2 \\ &= \Big(\sum_{k=0}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!}\Big) + \lambda - \lambda^2 \\ &= \Big(\lambda^2 e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}\Big) + \lambda - \lambda^2 = \lambda. \qquad \square\end{aligned}$$

Bemerkung 3.8 (Der Checkpot im Lotto)**.** Um nun das Gesetz der kleinen Zahlen aus Satz 3.5 anwenden zu können, betrachten wir das Lotto-Spiel 6 aus 49. Die Gewinnchancen für sechs Richtige sind hierbei wie in Beispiel 3.1

$$p := \frac{1}{\binom{49}{6}} = \frac{1}{13\,983\,816} \approx 7.15 \cdot 10^{-8}.$$

Geht man davon aus, dass an einer Ausspielung 10^7 Spiele abgegeben werden, gibt es (in Erwartung) etwa $\lambda := 0.715$ Sechser. Genauer können wir sagen, dass etwa die Anzahl X der gespielten Sechser Poisson-verteilt ist zum Parameter λ. Deshalb gilt z.B. für die Wahrscheinlichkeit, dass der Checkpot nicht geknackt wird

$$P[X = 0] \approx e^{-\lambda} \approx 0.49.$$

Man erhält aus dem Gesetz der kleinen Zahl, dass für genügend großes n und genügend kleines p_n die Binomialverteilung durch eine Poissonverteilung approximiert werden kann. Hierbei wählt man den Parameter λ so, dass der Erwartungswert der beiden Verteilungen gleich sind. Dieses Verfahren nennt man *Poissonapproximation*.

Bemerkung 3.9 (Güte der Poissonapproximation)**.** Möchte man eine Binomialverteilung nun durch eine Poissonverteilung approximieren, so macht man natürlich einen Fehler. Das Gesetz der kleinen Zahl sagt nichts über den Fehler. Allerdings kann man zeigen, dass für $X_n \sim \mathrm{bin}(n, p_n)$ und $Y_n \sim \mathrm{poi}(\lambda)$ gilt:

$$\sum_{k=0}^{\infty} \big|P[X_n = k] - P[Y_n = k]\big| \le 2np_n^2. \tag{3.2}$$

Für $p_n = \lambda/n$ ergibt sich $2np_n^2 = 2\lambda^2/n$. Den Beweis und eine Diskussion der Güte der Approximation findet man in Kapitel 5.4 in dem schönen Lehrbuch (**?**).

3.3 Die geometrische und die Exponentialverteilung

Man betrachte einen (unendlichen) p-Münzwurf $X_1, X_2, ...$ Wir wissen zwar schon, dass die Anzahl der Köpfe in den ersten n Würfen gerade $B(n,p)$ verteilt ist, jedoch noch nicht, wie lange man auf das erste Mal Kopf warten muss. Die Verteilung dieser Wartezeit ist die geometrische Verteilung.

Definition 3.10. Sei $p \in (0,1)$ und X eine $\mathbb{N}$-wertige Zufallsvariable mit

$$P[X = k] = (1-p)^{k-1}p, \qquad k = 1, 2, ...,$$

so heißt die Verteilung von X *geometrische Verteilung* zum Parameter p und wir schreiben $X \sim \text{geo}(p)$.

Satz 3.11. *Sei $X \sim geo(p)$. Dann ist*

$$\begin{aligned} P[X > k] &= (1-p)^k, \qquad k = 1, 2, ... \\ E[X] &= \frac{1}{p}, \\ Var[X] &= \frac{1-p}{p^2}. \end{aligned}$$

Beweis. Zunächst ist

$$P[X > k] = \sum_{i=k+1}^{\infty} (1-p)^{i-1}p = p\sum_{i=k}^{\infty}(1-p)^i = p(1-p)^k \sum_{i=0}^{\infty}(1-p)^i = (1-p)^k.$$

Für den Erwartungswert und die Varianz verwenden wir Lemma 2.12 und schreiben

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} P[X > k] = \sum_{k=0}^{\infty}(1-p)^k = \frac{1}{p}, \\ E[X(X-1)] &= 2\sum_{k=0}^{\infty} k \cdot P[X > k] \\ &= 2\sum_{k=0}^{\infty} k(1-p)^k = 2\frac{1-p}{p}\sum_{k=0}^{\infty} k(1-p)^{k-1}p = 2\frac{1-p}{p}E[X] = 2\frac{1-p}{p^2}, \end{aligned}$$

so dass wir leicht errechnen:

$$\text{Var}[X] = E[X(X-1)] + E[X] - E[X]^2 = 2\frac{1-p}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}. \qquad \square$$

Ist der Erfolgsparameter p einer geometrisch verteilten Zufallsvariable klein, so muss man recht lange auf den ersten Erfolg warten. Dies lässt sich zu einer Grenzwertaussage formalisieren.

Lemma 3.12. *Sei $X_n \sim geo(p_n)$ für $n = 1, 2, ...$ und (p_n) eine Nullfolge positiver Zahlen, so dass $np_n \to \lambda > 0$. Dann gilt für $x \geq 0$, dass*

$$\lim_{n\to\infty} P[X_n > nx] = e^{-\lambda x}.$$

Beweis. Wir schreiben einfach

$$P[X_n > nx] = (1 - p_n)^{nx} = \left(1 - \frac{np_n}{n}\right)^{nx} \xrightarrow{n\to\infty} e^{-\lambda x}. \qquad \square$$

Aus dem obigen Lemma erhält man ein Verfahren, wie man die geometrische Verteilung durch eine (im Folgenden einzuführende) Exponentialverteilung approximieren kann. Eine solche Vorgehensweise nennt man *Exponentialapproximation*.

Dies motiviert die folgende, wichtige Exponentialverteilung. Sie ist ein typisches Beispiel für eine Wartezeit unter Idealannahmen.

Definition 3.13. Sei $\lambda > 0$ und X eine $\mathbb{R}_+$-wertige Zufallsvariable mit Dichte

$$\lambda e^{-\lambda x} \mathbb{1}_{\{x\geq 0\}},$$

so heißt die Verteilung von X *Exponentialverteilung* zum Parameter λ und wir schreiben $X \sim \text{Exp}(\lambda)$.

Satz 3.14. *Für $\lambda > 0$ sei X nach Exp(λ) verteilt. Dann gilt*

$$E[X] = \frac{1}{\lambda},$$
$$\text{Var}[X] = \frac{1}{\lambda^2}.$$

Beweis. Mittels partieller Integration gilt

$$E[X] = \lambda \int_0^\infty x e^{-\lambda x} dx = -x e^{-\lambda x}\big|_0^\infty + \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda},$$
$$E[X^2] = \lambda \int_0^\infty x^2 e^{-\lambda x} dx = -x^2 e^{-\lambda x}\big|_0^\infty + \int_0^\infty 2x e^{-\lambda x} dx = \frac{2}{\lambda^2},$$
$$\text{Var}[X] = E[X^2] - E[X]^2 = \frac{1}{\lambda^2}. \qquad \square$$

3.4 Die Normalverteilung

Die Normalverteilung – besonders bekannt durch die Gauss'sche Glockenkurve – spielt in statistischen Anwendungen eine große Rolle. Grund hierfür ist der zentrale Grenzwertsatz, den wir in Satz 4.3 und Bemerkung 4.2 kennen lernen werden. Es folgt zunächst nur eine Definition.

Definition 3.15. Sei $\mu \in \mathbb{R},\ \sigma^2 > 0$ und X eine reellwertige Zufallsvariable mit Dichte

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2}\Big), \tag{3.3}$$

so heißt die Verteilung von X *Normalverteilung* mit Parametern μ und σ^2 und wir schreiben $X \sim N(\mu, \sigma^2)$.

Bemerkung 3.16. Aus der Analysis bekannt ist

$$\int \exp\Big(- \frac{x^2}{2}\Big) dx = \sqrt{2\pi}.$$

Damit zeigt man durch Substitution $z = (x-\mu)/\sqrt{\sigma^2}$ leicht, dass es sich bei (3.3) um eine Dichte handelt.

Satz 3.17. *Sei $X \sim N(\mu, \sigma^2)$. Dann ist*

$$\frac{X-\mu}{\sqrt{\sigma^2}} \sim N(0,1)$$

und

$$\begin{aligned} E[X] &= \mu, \\ \mathrm{Var}[X] &= \sigma^2. \end{aligned}$$

Beweis. Für die erste Aussage schreiben wir mit $Z \sim N(0,1)$

$$\begin{aligned} P\Big(\frac{X-\mu}{\sqrt{\sigma^2}} \le x\Big) &= P\Big(X \le x\sqrt{\sigma^2} + \mu\Big) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{x\sqrt{\sigma^2}+\mu} \exp\Big(- \frac{(y-\mu)^2}{2\sigma^2}\Big) dy \\ &= \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz = P(Z \le x). \end{aligned}$$

Zunächst ist

$$\int_{-\infty}^{\infty} x \exp\Big(- \frac{x^2}{2}\Big) dx = \exp\Big(- \frac{x^2}{2}\Big)\Big|_{-\infty}^{\infty} = 0.$$

Mit partieller Integration folgt, dass

$$\int_{-\infty}^{\infty} x^2 \exp\Big(-\frac{x^2}{2}\Big)dx = -x\exp\Big(-\frac{x^2}{2}\Big)\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \exp\Big(-\frac{x^2}{2}\Big)dx = \sqrt{2\pi},$$

also $E[Z] = 0, \mathrm{Var}[Z] = 1$. Daraus und aus der ersten Aussage folgt

$$E[X] = \sqrt{\sigma^2}E\Big[\frac{X-\mu}{\sqrt{\sigma^2}}\Big] + \mu = \sqrt{\sigma^2}E[Z] + \mu = \mu,$$
$$\mathrm{Var}[X] = \sigma^2 \mathrm{Var}\Big[\frac{X-\mu}{\sqrt{\sigma^2}}\Big] = \sigma^2 \mathrm{Var}[Z] = \sigma^2. \qquad \square$$

3.5 Erzeugung von Zufallszahlen

Will man ein stochastischen System numerisch untersuchen bieten sich mehrere Ansätze an. Als Beispiel stelle man sich die (zufällige) Tiefe eines Baumes im Quicksort Algorithmus vor. Die Verteilung dieser Tiefe ist schwer explizit zu bestimmen, jedoch kann man den Quicksort-Algorithmus (mit zufälliger Eingabesequenz) einfach simulieren, und damit auch die Verteilung der Tiefe des Baumes.

Um stochastische Systeme simulieren zu können, ist es notwendig Zufallsvariablen mit bestimmten Verteilungen mittels Zufallsgeneratoren zu ziehen. Während wir hier nicht über die Güte von Zufallsgeneratoren diskutieren wollen, stellen wir fest, dass diese meist uniform auf $[0, 1]$ verteilte Zufallszahlen liefern. Ausgehen von diesen kann man mittels des nächsten Satzes nach beliebigen Verteilungen verteilte Zufallszahlen generieren. Eine Illustration hierzu findet sich in Abbildung 3.1.

Abbildung 3.1: Illustration zum Simulationslemma, Theorem 3.18.

Theorem 3.18 (Simulationslemma)**.** *Sei X eine (diskrete oder stetige) reellwertige Zufallsvariable mit Verteilungsfunktion F. Wir definieren die Pseudoinverse von F für $x \in [0, 1]$ durch*

$$F^{-1}(x) := \inf\{q : F(q) \geq x\}.$$

Dann ist $F^{-1}(U)$ genauso verteilt wie X.

Beweis. Wir verwenden, dass die Verteilungsfunktion F die Verteilung der Zufallsvariable X eindeutig bestimmt. Da F nicht notwendigerweise injektiv ist, muss F^{-1} nicht die

Umkehrfunktion von F sein. Es gilt jedoch wegen der Konstruktion $F^{-1}(x) \leq q$ genau dann, wenn $x \leq F(q)$. Daraus folgt

$$P[F^{-1}(U) \leq q] = P[U \leq F(q)] = F(q).$$

Das bedeutet, dass $F^{-1}(U)$ die Verteilungsfunktion F hat, woraus die Aussage folgt. □

Bemerkung 3.19 (Anwendung)**.** Angenommen, von einer Verteilung ist die Verteilungsfunktion F (und damit die Pseudoinverse F^{-1}) bekannt. Wir wollen unabhängige Zufallsvariablen $X_1, X_2, ...$ nach dieser Verteilung erzeugen. Verwenden dürfen wir hierzu jedoch nur die vom Computer bereit gestellten uniform verteilten Zufallsvariablen $U_1, U_2, ...$ Das Simulationslemma besagt, dass in diesem Fall $F^{-1}(U_1), F^{-1}(U_2), ...$ genau die gewünschte Verteilung besitzen.

Bemerkung 3.20 (Alternativen)**.** Während das Verfahren aus Theorem 3.18 allgemeingültig ist, gibt es für spezielle Verteilungen schnellere Verfahren, Zufallszahlen zu erzeugen. Wir erwähnen hier (ohne Beweis) nur eine, nämlich die Erzeugung normalverteilter Zufallszahlen. Hierzu seien U_1, U_2 zwei unabhängige Zufallszahlen. Setzt man nun

$$X := \cos(2\pi U_1)\sqrt{-2\log(U_2)},$$

so ist $X \sim N(0, 1)$.

4. Grenzwertsätze

Wir beschäftigen uns nun mit der Verteilung der Summe unabhängiger und identisch verteilter Zufallsvariablen. Dies ist eine Situation, die in der Praxis häufig vorkommt, etwa wenn man die Anzahl der Gewinne in einem Spiel zählt. Für unabhängige, identisch verteilte Zufallsvariable $X_1, X_2, ...$ sei nun $Y_n := X_1 + ... + X_n$. Selbst bei bekannter Verteilung der X_i ist meistens die Verteilung von Y nicht bekannt. Allerdings gilt (falls Erwartungswert bzw. Varianz von X_1 existieren)

$$\begin{aligned} E[Y_n] &= E\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} E[X_i] = nE[X_1], \\ \mathrm{Var}[Y_n] &= \mathrm{Var}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathrm{Var}[X_i] = n\,\mathrm{Var}[X_1]. \end{aligned} \tag{4.1}$$

Geht man also davon aus, dass die typische Schwankung der Zufallsvariable Y_n in etwa ihrer Standardabweichung entspricht, sieht man das $\sqrt{n}$-Gesetz: Y_n streut ihre Werte typischerweise in einem Bereich der Größenordnung $\sqrt{n}$.

4.1 Das schwache Gesetz der großen Zahlen

Das schwache Gesetz der großen Zahlen ist eine Konvergenzaussage für den Mittelwert von unabhängigen, identisch verteilten Zufallsgrößen. Als Beispiel betrachte man die relative Häufigkeit $\frac{1}{n}\sum_{i=1}^{n} X_i$ der Köpfe in einem p-Münzwurf $X_1, X_2, ...$ Intuitiv klar ist, dass

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx p$$

gelten sollte. (Die relative Häufigkeit der Köpfe entspricht also in etwa der Wahrscheinlichkeit Kopf zu werfen.) In welchem Sinne diese Konvergenz zu verstehen ist, werden wir hier zwar nicht beleuchten können, das schwache Gesetz der großen Zahlen formalisiert diese Aussage jedoch. Eine Illustration findet sich in Abbildung 4.1. Zunächst benötigen wir zwei wichtige Ungleichungen.

Satz 4.1. *Sei X eine Zufallsvariable mit $X \geq 0$. Dann gilt für alle $\varepsilon > 0$*

$$P[X \geq \varepsilon] \leq \frac{1}{\varepsilon} E[X]. \qquad \text{(Markov-Ungleichung)}$$

Sei Y eine Zufallsvariable mit $\mathrm{Var}[X] < \infty$. Dann gilt für alle $\varepsilon > 0$

$$P[|X - \mu| \geq \varepsilon] \leq \frac{\mathrm{Var}[X]}{\varepsilon^2}. \qquad \text{(Tschebyschow-Ungleichung)}$$

Beweis. Zum Beweis der Markov-Ungleichung bemerken wir $X \geq \varepsilon \cdot \mathbb{1}_{\{X \geq \varepsilon\}}$. Damit ist wegen der Monotonie des Erwartungswertes

$$\varepsilon \cdot P[X \geq \varepsilon] = E[\varepsilon \cdot 1_{X \geq \varepsilon}] \leq E[X].$$

Die Tschebyschow-Ungleichung folgt nun aus der Markov-Ungleichung, wenn man die Zufallsvariable $(X - \mu)^2$ betrachtet. □

Theorem 4.2 (Schwaches Gesetz großer Zahlen)**.** *Seien $X_1, X_2, \ldots$ reellwertige, unabhängige, identisch verteilte Zufallsvariable mit endlichem Erwartungswert $E[X_1] = \mu$ und endlicher Varianz. Dann gilt für alle $\varepsilon > 0$*

$$\lim_{n \to \infty} P\Big[\Big|\frac{X_1 + \cdots + X_n}{n} - \mu\Big| \geq \varepsilon\Big] = 0. \tag{4.2}$$

Die Konvergenz in (4.2) nennt man Konvergenz in Wahrscheinlichkeit und schreibt oft

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

Beweis. Der Beweis erfolgt mittels der Tschebyschow-Ungleichung und (4.1), denn

$$P\Big(\Big|\frac{X_1 + \cdots + X_n}{n} - \mu\Big| \geq \varepsilon\Big) \leq \frac{\mathrm{Var}\left[\frac{1}{n}(X_1 + \cdots + X_n)\right]}{\varepsilon^2} = \frac{\mathrm{Var}[X_1]}{n\varepsilon} \xrightarrow{n \to \infty} 0. \qquad \square$$

4.2 Der zentrale Grenzwertsatz und die Normalverteilung

Während das Gesetz der großen Zahlen eine Aussage über die Konvergenz des Mittelwertes von unabhängigen, identisch verteilten Zufallsvariablen $X_1, X_2, \ldots$ gegen deren Erwartungswert trifft, beschäftigt sich der zentrale Grenzwertsatz mit den Schwankungen

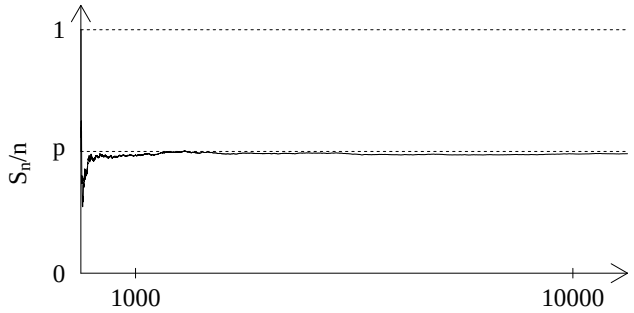


Abbildung 4.1: Illustration zum Gesetz der großen Zahlen, Theorem 4.2.

in dieser Konvergenz. Da $\mathrm{Var}[\frac{1}{n}(X_1 + \cdots + X_n)] = \mathrm{Var}\, X_1/n$, eine typische Schwankung also von der Größenordnung $1/\sqrt{n}$ ist, werden wir hier

$$\frac{\frac{1}{n}\sum_{i=1}^n X_i - E[X_i]}{\sqrt{n\,\mathrm{Var}[X_1]}}$$

betrachten. Obwohl es möglich ist, ein allgemeines Resultat zu beweisen, beschränken wir uns auf den Fall von $B(1,p)$-verteilten Zufallsgrößen $X_1, X_2, ...$ und bemerken, dass $X_1 + \cdots + X_n \sim B(n,p)$. Ohne Beweis werden wir außerdem die *Stirling Formel* zur Approximation von Fakultäten,

$$n! \approx \big(\frac{n}{e}\big)^n \sqrt{2\pi n}$$

verwenden. Hierbei bedeutet $\approx$, dass

$$\frac{n!}{\big(\frac{n}{e}\big)^n \sqrt{2\pi n}} \xrightarrow{n\to\infty} 1.$$

Eine Illustration des Satzes findet sich in Abbildung 4.2.

Abbildung 4.2: Illustration zum Satz von deMoivre-Laplace, Satz 4.3. Gezeigt sind jeweils die Verteilungsfunktionen von $N(0,1)$ (die glatte Kurve) und eine Transformation von $B(100, 0.5)$.

Satz 4.3 (Satz von de Moivre-Laplace)**.** *Sei Z eine $N(0,1)$ verteilte Zufallsvariable und $X_1, X_2, ...$ eine Folge binomialverteilter Zufallsvariable mit* $\mathrm{Var}[X_n] \xrightarrow{n\to\infty} \infty$*. Dann ist für* $-\infty \le c < d \le \infty$

$$P\Big[c \le \frac{X_n - E[X_n]}{\sqrt{\mathrm{Var}[X_n]}} \le d\Big] \xrightarrow{n\to\infty} P[c \le Z \le d].$$

Diese Konvergenz nennen wir *Konvergenz in Verteilung*, da die Verteilungsfunktionen konvergieren. Dies ist eine schwächere Aussage als die Konvergenz in Wahrscheinlichkeit und wir schreiben mit $Y_n := (X_n - E[X_n])/\sqrt{\mathrm{Var}[X_n]}$,

$$Y_n \xrightarrow{\mathscr{L}} N(0,1).$$

Man sieht leicht, dass die linke Seite für jedes n Erwartungswert 0 und Varianz 1 hat.

Beweisskizze. Mit der Stirling-Formel und $\eta(t) = t\ln\frac{t}{p} + (1-t)\ln\frac{1-t}{q}$ ist

$$\begin{aligned} P[X=k] = \binom{n}{k} p^k q^{n-k} &\approx \sqrt{\frac{n}{2\pi k(n-k)}}\Big(\frac{np}{k}\Big)^k\Big(\frac{nq}{n-k}\Big)^{n-k} \\ &= \frac{1}{\sqrt{2\pi n\frac{k}{n}\frac{n-k}{n}}}\exp\Big(-n\eta\Big(\frac{k}{n}\Big)\Big) \\ &\approx \frac{1}{\sqrt{2\pi n\frac{k}{n}\frac{n-k}{n}}}\exp\Big(-\frac{1}{2}\Big(\frac{k-np}{\sqrt{npq}}\Big)^2\Big), \end{aligned}$$

da $\eta(p) = \eta'(p) = 0, \eta''(p) = \frac{1}{pq}$ mit einer Taylor-Entwicklung von η um p. Damit gilt für $\mu_n = np_n, \sigma_n^2 = np_nq_n$ und $z_{x,n} = \frac{x-\mu_n}{\sqrt{\sigma_n^2}}$

$$\begin{aligned} P\Big(c \le \frac{X_n - E[X_n]}{\sqrt{\mathrm{Var}[X_n]}} \le d\Big) &= \sum_{k:c\le z_{k,n}\le d} P[X_n = k] \\ \approx \sum_{k:c\le z_{k,n}\le d} \frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\Big(-\frac{z_{k,n}^2}{2}\Big) &\approx \frac{1}{\sqrt{2\pi}}\int_c^d e^{-\frac{z^2}{2}}dz, \end{aligned}$$

wobei wir die letzte Summe als Riemannsumme interpretiert haben. □

Bemerkung 4.4. Als Anwendunge für den obigen Zentralen Grenzwertsatz betrachten wir folgendes Beispiel: Sei X eine $B(n,p)$ verteilte Zufallsgröße mit n groß, npq groß,

sowie $c, d \in \mathbb{Z}$. Dann gilt

$$P(c \le X \le d) \approx \sum_{k=c}^{d} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{z_{k,n}^2}{2}\Big)(z_{k+\frac{1}{2},n} - z_{k-\frac{1}{2},n})$$
$$\approx \frac{1}{\sqrt{2\pi}} \int_{z_{c-1/2,n}}^{z_{d+1/2,n}} \exp\Big(-\frac{z^2}{2}\Big)dz = \Phi\Big(\frac{d+\frac{1}{2}-np}{\sqrt{npq}}\Big) - \Phi\Big(\frac{c-\frac{1}{2}-np}{\sqrt{npq}}\Big).$$

Man beachte hierbei jeweils die Verschiebung um $\frac{1}{2}$ in der Binomialverteilung, die so genannte *Stetigkeitskorrektur*. Mit ihr erreicht man eine bessere Approximationsgüte. Diese Korrektur wurde etwa auch in Abbildung 4.2 angewandt.

Allgemeiner als der Satz von de Moivre-Laplace gilt auch folgender *Zentraler Grenzwertsatz*.

Satz 4.5. *Seien $X_1, X_2, ...$ reellwertige, unabhängige, identisch verteilte Zufallsvariablen mit endliche endlicher Varianz. Dann gilt für*

$$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - E[X_1]}{\sqrt{\mathrm{Var}[X_1]}},$$

dass

$$S_n \xrightarrow[n\to\infty]{\mathscr{L}} N(0,1). \tag{4.3}$$

5. Markov-Ketten

Wir betreten nun den Bereich der stochastischen Prozesse. Ein stochastischer Prozess ist hierbei eine Familie $(X_t)_{t\in I}$ von Zufallsvariablen mit Indexmenge $I \subseteq \mathbb{R}$. Wir werden im Folgenden die Index-Menge $I = \mathbb{Z}_+$ betrachten. Markov-Ketten zeichnen sich dadurch aus, dass X_{t+1} nur von X_t abhängt (und nicht von den Zufallsvariablen $X_0, ..., X_{t-1}$). Diese Abhängigkeit wird durch bedingte Wahrscheinlichkeiten beschrieben.

5.1 Bedingte Wahrscheinlichkeiten

Definition 5.1 (Bedingte Wahrscheinlichkeit)**.** Seien A,B Ereignisse, X_1 und X_2 reellwertige Zufallsvariable und A_1, A_2 Intervalle.

(i) Die *bedingte Wahrscheinlichkeit* von A gegeben B ist definiert durch

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

Ist $P[B] = 0$, so definieren wir die rechte Seite als 0.

(ii) Die Abbildung $A_2 \mapsto P[X_2 \in A_2|X_1 \in A_1]$ heißt die *bedingte Verteilung von X_2, gegeben* $\{X_1 \in A_1\}$.

(iii) Ist X_1 diskret mit Werten in $\{a_1, a_2, \dots\}$, so ist die bedingte Verteilung von X_2 gegeben X_1 die zufällige Abbildung $A_2 \mapsto f(A_2, X_1)$ mit

$$f(A_2, a) := P[X_2 \in A_2|X_1 = a].$$

Lemma 5.2 (Einfache Eigenschaften bedingter Wahrscheinlichkeiten)**.** *Seien X_1 und X_2 Zufallsvariablen und A_1, A_2 Intervalle.*

(i) Es gilt

$$P[X_2 \in A_2|X_1 \in A_1] = \sum_{a_2 \in A_2} P[X_2 = a_2|X_1 \in A_1].$$

(ii) Die Zufallsvariablen X_1 und X_2 sind genau dann unabhängig, wenn

$$P[X_2 \in A_2 | X_1 \in A_1] = P[X_2 \in A_2]$$

für alle A_1, A_2 gilt. Ist X_1 diskret, so ist dies genau dann der Fall, wenn

$$P[X_2 \in A_2 | X_1] = P[X_2 \in A_2]$$

für alle A_2 gilt.

Beweis. Beide Eigenschaften folgen aus der Definition der bedingten Wahrscheinlichkeit. □

B 5.1 *Gedächtnislosigkeit der geometrischen Verteilung*: Sei $X \sim \text{geo}(p)$. Dann gilt

$$P[X > i + j | X > i] = P[X > j].$$

Dies interpretiert man so: Man führt unabhängige Experimente mit Erfolgswahrscheinlichkeit p hintereinander durch. Wenn i Versuche erfolglos verliefen, ist die Wahrscheinlichkeit für mindestend weitere j erfolglose Versuche genauso groß wie am Anfang.
Denn: Mit $q = 1 - p$ schreiben wir

$$\begin{aligned} P[X > i + j | X > i] &= \frac{P[X > i + j, X > i]}{P[X > i]} = \frac{q^{i+j}}{q^i} = q^j \\ &= P[X > j]. \end{aligned}$$

Eine analoge Aussage gilt übrigens auch für Exponentialverteilungen.

Theorem 5.3 (Formel für die totalen Wahrscheinlichkeit und Bayes'sche Formel)**.** *Seien X_1, X_2 diskrete Zufallsvariable. Dann gilt die Formel von der totalen Wahrscheinlichkeit*

$$P[X_2 \in A_2] = \sum_{a_1} P[X_2 \in A_2 | X_1 = a_1] \cdot P[X_1 = a_1].$$

für Intervalle A_2. Weiter gilt die Bayes'sche Formel*: für ein Intervall A_1 mit $P[X_1 \in A_1] > 0$ ist*

$$P[X_1 \in A_1 | X_2 \in A_2] = \frac{P[X_2 \in A_2 | X_1 \in A_1] \cdot P[X_1 \in A_1]}{\sum_{a_1} P[X_2 \in A_2 | X_1 = a_1] \cdot P[X_1 = a_1]}.$$

Beweis. Die Formel von der totalen Wahrscheinlichkeit ergibt sich durch Einsetzen der Definition von $P[X_2 \in A_2 | X_1 = a_1]$ in

$$P[X_2 \in A_2] = \sum_{a_1} P[X_2 \in A_2, X_1 = a_1].$$

In der Bayes'schen Formel ist der Zähler der rechten Seite gleich $P[X_1 \in A_1, X_2 \in A_2]$ nach der Definition der bedingten Wahrscheinlichkeit, und der Nenner ist

$$\sum_{a_1} P[X_2 \in A_2 | X_1 = a_1] \cdot P[X_1 = a_1] = \sum_{a_1} P[X_2 \in A_2, X_1 = a_1] = P[X_2 \in A_2].$$

□

B 5.2 *Reihenuntersuchungen*: In einer Reihenuntersuchung werden Personen auf eine bestimmte Krankheit getestet. Dabei kann es fälschlicherweise vorkommen, dass gesunde Personen durch den Test als krank eingestuft werden, oder dass kranke Personen nicht als solche erkannt werden.
In einer Population sind insgesamt 0.8% der Personen erkrankt. Eine kranke Person wird in 90% der Fälle positiv getestet (der Test fällt also positiv aus), eine gesunde Person mit 7%. Wie groß ist die Wahrscheinlichkeit, dass eine positiv getestete Person wirklich krank ist?
Um dies zu beantworten, setzen wir $X = 1$ wenn die Person erkrankt ist (sonst $X = 0$) und $Y = 1$ wenn die Person positiv getestet wird (sonst $Y = 0$). Die genannten Angaben übersetzen wir zu

$$P[X = 1] = 0.008, \qquad P[Y = 1|X = 1] = 0.9, \qquad P[Y = 1|X = 0] = 0.07.$$

Mit der Formel von Bayes ergibt sich für die Wahrscheinlichkeit, dass eine positiv getestete Person wirklich erkrankt ist

$$\begin{aligned} P[X = 1|Y = 1] &= \frac{P[Y = 1|X = 1] \cdot P[X = 1]}{P[Y = 1|X = 1] \cdot P[X = 1] + P[Y = 1|X = 0] \cdot P[X = 0]} \\ &= \frac{0.9 \cdot 0.008}{0.9 \cdot 0.008 + 0.07 \cdot 0.992} \approx 0.0939 \end{aligned}$$

Ein positiver Test ist also nicht unbedingt ein sicheres Zeichen dafür, ob eine Person erkrankt ist.

Bemerkung 5.4 (Wahrscheinlichkeitsbäume)**.** Bedingte Wahrscheinlichkeiten kann man durch Wahrscheinlichkeitsbäume darstellen. Siehe Abbildung 5.1.

5.2 Grundlegendes zu Markov-Ketten

Eine besondere Form der Abhängigkeit von Zufallsvariablen tritt in Markov-Ketten zu Tage. Hier werden der Reihe nach Zufallsvariablen $X_0, X_1, \ldots$ realisiert, und zwar so, dass X_{t+1} nur von X_t abhängt. Man sagt auch, X_{t+1} ist unabhängig von $X_1, \ldots, X_{t-1}$, wenn

units ¡.95cm,.95cm¿ x from -.4 to 16, y from 1 to 6.5 2 2 8 6 14 2 / 6 2 4 3.333 / 10 2 12 3.333 / krank
[cC] at 5 5 nicht krank [cC] at 11.5 5 0.008 [cC] at 6.8 4.5 0.992 [cC] at 9.2 4.5 0.9 [cC] at 2.5 3 0.1 [cC] at 5.5 3 0.07 [cC] at 10.5 3 0.93 [cC] at 13.5 3 positiv [cC] at 1.4 2.5 negativ [cC] at 6.7 2.5 positiv [cC] at 9.4 2.5 negativ [cC] at 14.7 2.5 krank positiv [cC] at 2.5 1.3 krank negativ [cC] at 6.5 1.3 nicht krank positiv [cC] at 10 1.3 nicht krank negativ [cC] at 14 1.3 $0.008 \cdot 0.9$ [cC] at 2.5 .3 $0.008 \cdot 0.1$ [cC] at 6.5 .3 $0.992 \cdot 0.07$ [cC] at 10 .3 $0.992 \cdot 0.93$ [cC] at 14 .3

Abbildung 5.1: Ein Wahrscheinlichkeitsbaum für das Beispiel 5.2.

X_t bekannt ist. Diese Form der Abhängigkeit wird häufig zur stochastischen Modellierung verwendet. Beispielsweise könnte $X_0, X_1, \dots$ der Preis einer Aktie an Tagen $0, 1, \dots$ sein. Die Markov-Eigenschaft besagt in diesem Fall, dass die Verteilung der Kursänderungen am Tag $t+1$ nur davon abhängt, wie der Kurs X_t am Tag t war.

Wir beginnen mit der Definition von Markov-Ketten. Wir werden vor allem den Fall von homogenen Markov-Ketten behandeln. Bei solchen gibt es eine sich zeitlich nicht ändernde stochastische Übergangsvorschrift, wie die Verteilung des Zustandes X_{t+1} ist, wenn der Zustand X_t der Kette zur Zeit t bekannt ist. Diese Vorschrift wird mit Hilfe einer Matrix zusammengefasst, der Übergangsmatrix.

Definition 5.5. Seien I und E Mengen.

(i) Ein (E-wertiger) stochastischer Prozess (mit Indexmenge I) ist eine Familie von Zufallsvariablen $X = (X_t)_{t \in I}$ mit Wertebereich E. Die Menge E heißt auch *Zustandsraum* von X und I seine *Indexmenge*.

(ii) Sei $I = \{0, 1, 2, \dots\}$, E abzählbar und $X = (X_t)_{t \in I}$ ein E-wertiger stochastischer Prozess. Falls

$$P[X_{t+1} = i \mid X_0, \dots, X_t] = P[X_{t+1} = i \mid X_t] \tag{5.1}$$

für alle $i \in E$, so heißt X eine *Markov-Kette*. Sie heißt *endlich*, falls E endlich ist.

(iii) Sei $X = (X_t)_{t \in I}$ eine E-wertige Markov-Kette. Existiert eine Matrix $P = (P_{ij})_{i,j \in E}$ mit

$$P_{ij} := P[X_{t+1} = j \mid X_t = i],$$

so heißt X *zeitlich homogen* und P heißt *Übergangsmatrix* von X.

Typischerweise wird es kaum Verwechslungsmöglichkeiten zwischen der Übergangsmatrix P und dem Wahrscheinlichkeitsmaß geben, man sollte diese beiden aber strikt unterscheiden.

Bemerkung 5.6. Zur Interpretation der Definition:

(i) Die Eigenschaft (5.1) bedeutet in Worten: die zukünftige Entwicklung von X nach t hängt von $X_1, \dots, X_t$ nur durch den aktuellen Zustand X_t ab.

(ii) Sei P die Übergangsmatrix einer homogenen Markov-Kette X mit Zustandsraum E. Dann gilt

$$\begin{aligned} 0 \le P_{ij} \le 1, \qquad & i, j \in E, \\ \sum_{j \in E} P_{ij} = 1, \qquad & i \in E. \end{aligned} \tag{5.2}$$

Die erste Eigenschaft ist klar, da die Einträge in P Wahrscheinlichkeiten sind. Außerdem ist

$$1 = P[X_{t+1} \in E \mid X_t = i] = \sum_{j \in E} P[X_{t+1} = j \mid X_t = i] = \sum_{j \in E} P_{ij}.$$

Matrizen P mit den Eigenschaften (5.2) heißen *stochastische Matrizen.*

(iii) Sei X eine homogene Markov-Kette mit Übergangsmatrix P. Definiere einen gewichteten, gerichteten Graphen (E, K, W) wie folgt: die Menge der Knoten ist E, die Menge der (gerichteten) Kanten ist $K := \{(i, j) : P_{ij} > 0\}$. Das Gewicht der Kante (ij) ist $w_{(ij)} := P_{ij}$ und $W = (w_{(ij)})_{(ij) \in K}$. Der Graph (E, K, W) heißt *Übergangsgraph* von X.

B 5.3 *Irrfahrt im Dreieck*: Betrachte eine homogene Markov-Kette X mit Zustandsraum $\{1, 2, 3\}$ und Übergangsmatrix

$$P = \begin{pmatrix} 0 & p & q \\ q & 0 & p \\ p & q & 0 \end{pmatrix} \tag{5.3}$$

für $p \in (0, 1)$ und $q := 1 - p$. Die Kette X veranschaulicht man sich am besten anhand des Übergangsgraphen; siehe Abbildung 5.2. In jedem Zustand $1, 2, 3$ ist die Wahrscheinlichkeit, im Uhrzeigersinn zu wandern p, und die Wahrscheinlichkeit gegen den Uhrzeigersinn zu gehen ist q. □

Abbildung 5.2: Übergangsgraph der Markov-Kette aus Beispiel 5.3.

B 5.4 *Ruinproblem*: Betrachte folgendes Ruinproblem: zwei Spieler spielen gegeneinander. Spieler 1 startet mit n Euro, Spieler 2 mit $N-n$ Euro. Bei jedem Spiel gewinnt Spieler 1 mit Wahrscheinlichkeit p einen Euro, mit Wahrscheinlichkeit $q = 1-p$ verliert er einen Euro. Das Spiel endet, wenn einer der beiden pleite ist.

Sei X_t das Vermögen von Spieler 1 nach dem t-ten Spiel. In dieser Situation ist $X = (X_t)_{t=0,1,2,\ldots}$ eine endliche, homogene Markov-Kette mit Zustandsraum $\{0,\ldots,N\}$ und Übergangsmatrix

$$P = \begin{pmatrix} 1 & 0 & & & & \\ q & 0 & p & & & \\ & q & 0 & p & & \\ & & \ddots & \ddots & \ddots & \\ & & & q & 0 & p \\ & & & & 0 & 1 \end{pmatrix}.$$

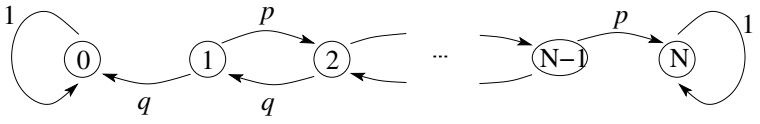


Abbildung 5.3: Übergangsgraph der Markov-Kette aus Beispiel 5.4.

Der Übergangsgraph ist in Abbildung 5.3 dargestellt. Klar ist, dass nach langer Zeit entweder Spieler 1 oder Spieler 2 gewonnen hat, dass also $X_t \xrightarrow{t\to\infty} 0$ oder $X_t \xrightarrow{t\to\infty} N$. Die Frage ist nur:

Mit welcher Wahrscheinlichkeit gewinnt Spieler 1?

Wir bezeichnen mit

$$p_n := P[X_t \xrightarrow{t\to\infty} N \mid X_0 = n]$$

die Wahrscheinlichkeit, dass Spieler 1 gewinnt. Wir werden zeigen, dass

$$p_n = \begin{cases} \dfrac{n}{N}, & p = q = \frac{1}{2}, \\ \dfrac{1-(q/p)^n}{1-(q/p)^N}, & \text{sonst.} \end{cases} \tag{5.4}$$

Zunächst gilt für $n = 1, \ldots, N-1$

$$\begin{aligned} p_n &= q \cdot P[X_t \xrightarrow{t\to\infty} N \mid X_0 = n, X_1 = n-1] + p \cdot P[X_t \xrightarrow{t\to\infty} N \mid X_0 = n, X_1 = n+1] \\ &= q \cdot P[X_t \xrightarrow{t\to\infty} N \mid X_0 = n-1] + p \cdot P[X_t \xrightarrow{t\to\infty} N \mid X_0 = n+1] \\ &= q \cdot p_{n-1} + p \cdot p_{n+1}, \end{aligned}$$

also mit $\Delta p_n := p_n - p_{n-1}$

$$q\Delta p_n = p\Delta p_{n+1}.$$

Im Fall $p = q = \frac{1}{2}$ folgt mit $\sum_{m=1}^{N} \Delta p_m = p_N - p_0 = 1$ daraus bereits

$$p_n = \frac{\sum_{m=1}^{n} \Delta p_m}{\sum_{m=1}^{N} \Delta p_m} = \frac{n\Delta p_1}{N\Delta p_1} = \frac{n}{N}.$$

Im Fall $p \neq q$ setzen wir $u := \frac{q}{p}$ und berechnen iterativ

$$\Delta p_n = u\Delta p_{n-1} = u^2 \Delta p_{n-2} \cdots = u^{n-1}\Delta p_1 = u^{n-1} p_1.$$

Weiter ist

$$1 = \sum_{m=1}^{N} \Delta p_m = p_1 \sum_{m=0}^{N-1} u^m = p_1 \frac{1-u^N}{1-u}.$$

Also

$$p_n = \sum_{m=1}^{n} \Delta p_m = p_1 \sum_{m=1}^{n} u^{m-1} = \frac{1-u}{1-u^N}\frac{1-u^n}{1-u} = \frac{1-u^n}{1-u^N}$$

und die Behauptung ist gezeigt. □

B 5.5 *Ehrenfest'sche Urne*: Betrachte folgendes Urnenmodell: In einer Urne gibt es zwei durch eine Trennwand getrennte Kammern. Insgesamt liegen n Kugeln in den beiden Kammern. Wir ziehen eine Kugel rein zufällig aus der Urne und legen sie anschließend in die andere Kammer zurück. In Abbildung 5.4 findet sich eine Illustration.

Wir betrachten den $\{0, \ldots, n\}$-wertigen stochastischen Prozess $X = (X_t)_{t=0,1,2,\ldots}$, wobei X_t die Anzahl der Kugeln in der linken Kammer nach dem t-Schritt darstellt. Dann ist X eine homogene Markov-Kette, denn der Ausgang des Schrittes $t+1$ hängt nur von X_t ab. Die Überganmatrix von X ist gegeben durch

$$P_{ij} = \begin{cases} \frac{i}{n}, & j = i-1, \\ \frac{n-i}{n}, & j = i+1, \\ 0, & \text{sonst.} \end{cases} \tag{5.5}$$

□

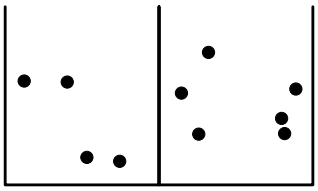


Abbildung 5.4: Die Ehrenfest'sche Urne mit $n = 10$ Kugeln aus Beispiel 5.5. Im nächsten Schritt wird mit Wahrscheinlichkeit $\frac{4}{10}$ eine Kugel von der linken Kammer in die rechte und mit Wahrscheinlichkeit $\frac{6}{10}$ von der rechten in die linke Kammer gelegt.

Wir kommen nun zu einem wichtigen Zusammenhang zwischen der Verteilung einer Markov-Ketten zum Zeitpunkt t und Potenzen der Übergangsmatrix.

Theorem 5.7. *Sei $X = (X_t)_{t=0,1,2,\dots}$ eine homogene Markov-Kette mit Übergangsmatrix P und $\mu^{(t)} = (\mu^{(t)}(i))_{i\in E}$,*

$$\mu^{(t)}(i) = P[X_t = i]$$

für $t = 0, 1, 2, \dots$ Dann gilt

$$\mu^{(t)} = \mu^{(0)} P^t, \tag{5.6}$$

für $t = 0, 1, 2, \dots$, wobei die rechte Seite als Multiplikation des Zeilenvektors $\mu^{(0)}$ und der Matrix $P^t = \underbrace{P \cdots P}_{t \text{ mal}}$ zu verstehen ist.

Beweis. Der Beweis geht mittels Induktion über t. Für $t = 0$ ist die Aussage klar. Ist (5.6) für t gezeigt, so gilt

$$\begin{aligned}\mu^{(t+1)}(j) = P[X_{t+1} = j] &= \sum_{i\in E} P[X_{t+1} = j \mid X_t = i] \cdot P[X_t = i] \\ &= \sum_{i\in E} \mu^{(t)}(i) \cdot P_{ij} = \sum_{i\in E} (\mu^{(0)} P^t)_i \cdot P_{ij} = (\mu^{(0)} P^{t+1})_j.\end{aligned}$$

□

B 5.6 *Irrfahrt im Dreieck*: Betrachte die Markov-Kette X aus Beispiel 5.3 und Übergangsmatrix P aus (5.3). Sei $X_0 = 1$, die Markov-Kette startet also in 1, d.h. $\mu^{(0)} = (1, 0, 0)$. Weiter

berechnen wir

$$P = \begin{pmatrix} 0 & p & q \\ q & 0 & p \\ p & q & 0 \end{pmatrix}, \qquad P^2 = \begin{pmatrix} 2pq & q^2 & p^2 \\ p^2 & 2pq & q^2 \\ q^2 & p^2 & 2pq \end{pmatrix}$$

und damit

$$\mu^{(2)} = (2pq, q^2, p^2).$$

Also ist etwa $P[X_2 = 1] = 2pq$. Dies ist klar, weil $X_2 = 1$ genau dann, wenn die ersten beiden Schritte im Übergangsgraphen aus Abbildung 5.2 einmal mit und einmal entgegen dem Uhrzeigersinn gingen. Es ist $X_2 = 3$ genau dann, wenn zwei Schritte im Uhrzeigersinn realisiert wurden, was Wahrscheinlichkeit p^2 hat. □

5.3 Stationäre Verteilungen

Wir beginnen nun, uns für Markov-Ketten zu interessieren, die schon lange gelaufen sind, also für X_t mit großem t. Solche Markov-Ketten können (mehr dazu im nächsten Abschnitt) so beschaffen sein, dass sich die Verteilung nicht mehr viel ändert. Verteilungen auf dem Zustandsraum E, die invariant sind gegenüber Schritten der Markov-Kette heißen stationär.

Definition 5.8. Sei X eine homogene Markov-Kette mit Übergangsmatrix P und π eine Verteilung auf E, gegeben durch einen Vektor $\pi = (\pi_i)_{i \in E}$.

(i) Gilt

$$\pi P = \pi,$$

so heißt π *stationäre Verteilung* von X.

(ii) Gilt

$$\pi_i P_{ij} = \pi_j P_{ji}$$

für alle i, j, so heißt π reversible Verteilung.

Bemerkung 5.9 (Interpretation von reversiblen Verteilungen)**.** Sei π reversibel, $P[X_t = i] = \pi_i$ und $i_t, \ldots, i_{t+s} \in E$. Dann gilt

$$\begin{aligned} P[X_t = i_t, \ldots, X_{t+s} = i_{t+s}] &= \pi_{i_t} P_{i_t, i_{t+1}} \cdots P_{i_{t+s-1}, i_{t+s}} \\ &= \pi_{i_{t+1}} P_{i_{t+1}, i_{t+2}} \cdots P_{i_{t+s-1}, i_{t+s}} \cdot P_{i_{t+1}, i_t} = \cdots \\ &= \pi_{i_{t+s}} P_{i_{t+s}, i_{t+s-1}} \cdots P_{i_{t+1}, i_t} \\ &= P[X_t = i_{t+s}, \ldots, X_{t+s} = i_t]. \end{aligned}$$

Das bedeutet, dass die Wahrscheinlichkeit, die Zustände $i_t, \dots, i_{t+s}$ zu durchlaufen, dieselbe ist wie die, die Zustände in umgekehrter Reihenfolge (reversibel) zu durchlaufen.

Lemma 5.10 (Einfache Eigenschaften stationärer Verteilungen)**.** *Sei $X = (X_t)_{t\in I}$ eine E-wertige Markov-Kette mit Übergangsmatrix P und π eine Verteilung auf E.*

(i) Sei π stationär und $P[X_t = i] = \pi_i$. Dann gilt, dass

$$P[X_{t+1} = j] = P[X_t = j].$$

Das bedeutet, dass X_t und X_{t+1} (und damit auch $X_{t+2}, X_{t+3}, \dots$) identisch verteilt sind.

(ii) Ist π reversibel, dann ist π auch stationär.

Beweis. 1. Wir verwenden Theorem 5.7 und schreiben

$$\begin{aligned} P[X_{t+1} = j] &= \sum_{i\in E} P[X_t = i] \cdot P[X_{t+1} = j \mid X_t = i] = \sum_{i\in E} \pi_i P_{ij} = (\pi P)_j = \pi_j \\ &= P[X_t = j]. \end{aligned}$$

2. Ist π reversibel, so gilt

$$(\pi P)_j = \sum_{i\in E} \pi_i P_{ij} = \sum_{i\in E} \pi_j P_{ji} = \pi_j,$$

da P Zeilensumme 1 hat. □

B 5.7 *Irrfahrt im Dreieck*: Betrachte die Irrfahrt auf dem Dreieck X aus Beispiel 5.3 mit Übergangsmatrix P aus (5.3). Für $\pi = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ ist

$$\pi P = \left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}\right) \begin{pmatrix} 0 & p & q \\ q & 0 & p \\ p & q & 0 \end{pmatrix} = \pi.$$

Damit ist die Gleichverteilung auf $\{1, 2, 3\}$ stationäre Verteilung von X. Allerdings ist π nicht reversibel (außer für $p = \frac{1}{2}$), denn es gilt etwa

$$\pi_1 P_{12} = \tfrac{1}{3} p \neq \tfrac{1}{3} q = \pi_2 P_{21}.$$

Letztere Beobachtung ist nicht erstaunlich: Sei etwa $p > \frac{1}{2}$. Dann erwarten wir, dass X das Dreieck öfter im Uhrzeigersinn durchläuft als umgekehrt. Wäre nun π reversibel, würde daraus folgen (siehe Bemerkung 5.9) , dass die Wahrscheinlichkeit für einen Durchlauf des Dreiecks gegen und im Uhrzeigersinn dieselbe ist. □

B 5.8 *Ruinproblem*: Im Ruinproblem aus Beispiel 5.4 gibt es zwar stationäre Verteilungen, diese sind jedoch nicht sonderlich interessant. Jede Verteilung $\pi = (p', 0, \ldots, 0, q')$ mit $p' \in (0,1)$ und $q' = 1 - p'$ ist stationäre (aber nicht reversible) Verteilung, wie man leicht nachrechnet. Insbesondere sind $(1, 0, \ldots, 0)$ und $(0, \ldots, 0, 1)$ stationär. Das bedeutet, dass die Zustände $X_t = 0$ und $X_t = n$ von der Markov-Kette nicht mehr verlassen werden. Man sagt auch, 0 und n sind *Fallen* für X. □

B 5.9 *Ehrenfest'sche Urne*: Sei $X = (X_t)_{t=0,1,2,\ldots}$ die Anzahl der Kugeln in der linken Kammer einer Ehrenfest'schen Urne, wie in Beispiel 5.5. Hier ist die Binomialverteilung $\pi = B(n, \frac{1}{2})$ reversible Verteilung für X. Es gilt nämlich mit (5.5) für $i = 1, \ldots, n$

$$\pi_i P_{i,i-1} = \binom{n}{i} \frac{1}{2^n} \frac{i}{n} = \binom{n}{i-1} \frac{1}{2^n} \frac{n-i+1}{n} = \pi_{i-1} P_{i-1,i}.$$

Die Tatsache, dass $B(n,p)$ stationär ist, interpretiert man am besten so: nach langer Zeit ist jede der Kugeln oft von der linken in die rechte Kammer und umgelegt worden, so dass jede Kugel mit Wahrscheinlichkeit $\frac{1}{2}$ jeweils in der linken und rechten Kammer liegt. Diese Verteilung ändert sich nicht mehr, weil in jedem Schritt nur eine der n Kugeln nochmals in die andere Kammer gelegt wird. □

Wir geben nun noch – ohne Beweis – den wichtigen Markov-Ketten-Konvergenzsatz an. Dieser stellt Bedingungen auf, wann eine Markov-Kette eine eindeutige stationäre Vertreilung besitzt. Konvergenz der Markov-Kette $X = (X_t)_{t \in I}$ bedeutet dabei, dass $P[X_t = i] \xrightarrow{t \to \infty} \pi(i)$ für diese stationäre Verteilung π gilt, unabhängig von der Verteilung von X_0. Um den Satz zu formulieren, benötigen wir noch eine Definition.

Definition 5.11. Sei X eine E-wertige homogene Markov-Kette.

(i) Der Zustand $i \in E$ *kommuniziert* mit $j \in E$, falls es ein t gibt mit

$$P[X_t = j \mid X_0 = i] > 0.$$

In diesem Fall schreiben wir $i \to j$. Falls $i \to j$ und $j \to i$, schreiben wir $i \leftrightarrow j$.

(ii) Falls $i \leftrightarrow j$ für alle $i, j \in E$, so heißt X *irreduzibel*. Andernfalls heißt X *reduzibel*.

(iii) Ein Zustand $i \in E$ heißt *aperiodisch*, falls

$$d(i) := \mathrm{ggT}\{t : P[X_t = i \mid X_0 = i] > 0\} = 1.$$

Andernfalls heißt i *periodisch* mit Periode $d(i)$.

(iv) Falls alle $i \in E$ aperiodisch sind, so heißt X *aperiodisch*.

Bemerkung 5.12 (Irreduzibilität, Aperiodizität und die Übergangsmatrix)**.**

(i) Die Begriffe der Irreduzibilität und Aperiodizität lassen sich auch mittels der Übergangsmatrix P der Markov-Kette $\mathcal{X}$ erklären. Es gilt etwa $i \to j$ genau dann, wenn es ein t gibt mit $(P^t)_{ij} > 0$. Außerdem ist $d(i) = \text{ggT}\{t : P^t_{ii} > 0\}$.

(ii) Es gibt einen einfachen Zusammenhang zwischen dem Begriff der kommunizierenden Zustände und dem Übergangsgraphen: ein Zustand i kommuniziert mit einem Zustand j, wenn man einen Pfad $i \to j$ im Übergangsgraphen der Markov-Kette findet. Das bedeutet, dass es eine endliche Folge $(i, i_1), (i_1, i_2), \ldots, (i_{n-1}, i_n), (i_n, j)$ im Übergangsgraphen gibt, also der Zustand j von i aus durch eine endliche Folge von Kanten erreicht werden kann.

(iii) Sei X eine reduzible Markov-Kette. Der Begriff der Reduzibilität erklärt sich so, dass sich die Markov-Kette X auf einen Zustandsraum $E' \subseteq E$ *reduzieren* lässt. Das bedeutet, dass aus $X_0 \in E'$ folgt, dass $X_t \in E'$ für alle $t = 1, 2, \ldots$

B 5.10 *Irrfahrt auf dem Dreieck*: Sei X die Irrfahrt auf dem Dreieck aus Beispiel 5.3. In Beispiel 5.6 haben wir ausgerechnet, dass $P^2 > 0$ ist. Weiter ist auch $P^3 > 0$, was bereits zeigt, dass X sowohl irreduzibel als auch aperiodisch ist. □

B 5.11 *Ruinproblem*: Für die Markov-Kette X aus dem Ruinproblem, Beispiel 5.4, gilt sicher, dass $0 \not\to i$ für $i = 1, \ldots, N$, und $N \not\to i$ für $i = 0, \ldots, N-1$. Mit anderen Worten: Ist Spieler 1 pleite (d.h. $X_t = 0$ für ein t), so wird er nach den Spielregeln nie wieder Geld bekommen, d.h. alle Zustände $1, \ldots, n$ sind für ihn unerreichbar. Das bedeutet also, X ist reduzibel. □

B 5.12 *Ehrenfest'sche Urne*: Die Markov-Kette X, die die Anzahl der Kugeln in der linken Kammer einer Ehrenfest'schen Urne beschreibt, ist irreduzibel. Betrachtet man nämlich Zustände i, j, etwa mit $i > j$, so genügt es ja, genau $i - j$ Kugeln nacheinander von der linken in die rechte Kammer zu legen. Allerdings ist die Markov-Kette periodisch. Sei etwa $X_0 = 2i$ gerade, dann folgt, dass X_1 ungerade (entweder $2i - 1$ oder $2i + 1$) ist. Damit kann nur zu geraden Zeiten t der Zustand $X_t = 2i$ eintreten und es ist $\text{ggT}\{t : P(X_t = 2i | X_0 = 2i)\} = 2$. □

Nun kommen wir also zum Markov-Ketten-Konvergenzsatz.

Theorem 5.13 (Markov-Ketten-Konvergenzsatz)**.** *Sei $X = (X_t)_{t=0,1,2,\ldots}$ eine endliche, homogene, irreduzible und aperiodische Markov-Kette. Dann hat X genau eine stationäre Verteilung π und es gilt für alle $i \in E$*

$$P[X_t = i] \xrightarrow{t \to \infty} \pi(i). \tag{5.7}$$

Beweis. Der Beweis findet sich in allen gängigen Lehrbüchern, in denen Markov-Ketten behandelt werden. □

Bemerkung 5.14. Vor allem ist an dem Resultat bemerkenswert, dass (5.7) nicht von der Verteilung von X_0 abhängt. Das bedeutet, dass die Markov-Kette nach langer Zeit *vergisst*, in welchem Zustand X_0 sie zur Zeit 0 gestartet ist.

B 5.13 *Irrfahrt auf dem Dreieck*: Von unseren drei Beispielen erfüllt (nur) die Irrfahrt auf dem Dreieck X aus Beispiel 5.3 die Voraussetzungen des Satzes. Nach Beispiel 5.10 ist X irreduzibel und aperiodisch. Nach Beispiel 5.7 und obigem Satz ist die Gleichverteilung auf $\{1, 2, 3\}$ die einzige stationäre Verteilung für $\mathcal{X}$. Außerdem gilt

$$P[X_t = i] \xrightarrow{t\to\infty} \tfrac{1}{3}.$$

□

B 5.14 *Ruinproblem*: Für die Markov-Kette X aus dem Ruinproblem, Beispiel 5.4, gibt es stationäre Verteilungen π, nämlich solche mit $\pi(0) + \pi(N) = 1$. (Bei solchen ist einer der beiden Spielern pleite.) Offenbar gilt auch

$$\lim_{t\to\infty} P[X_t = N | X_0 = n] = p_n$$

mit p_n aus (5.4). Das bedeutet, dass die (Verteilung der) Markov-Kette konvergiert, die Grenzverteilung aber vom Startwert (hier $X_0 = n$) abhängt. Dies ist im Einklang mit Theorem 5.13, da X nicht irreduzibel ist.

B 5.15 *Ehrenfest'sche Urne*: Die Markov-Kette X, die die Anzahl der Kugeln in der linken Kammer einer Ehrenfest'schen Urne beschreibt, ist periodisch, und damit ist Theorem 5.13 nicht anwendbar. Ist etwa $X_0 = 2n$ gerade, so ist X_{2t} auch gerade, also $P[X_{2t} \text{ gerade} | X_0 = 2n] = 1$, aber $P[X_{2t} \text{ gerade} | X_0 = 2n + 1] = 0$. Insbesondere existiert der Grenzwert von $P[X_t = k || X_0 = n]$ nicht.

6. Statistik

Es ist nicht übertrieben zu behaupten, dass in der heutigen Welt immer mehr *Daten* jeglicher Art erhoben werden. Diese zu ordnen und aus Daten Schlussfolgerungen zu ziehen ist Aufgabe der Statistik.

Man teilt dabei diese Aufgaben in zwei Gebiete auf. Die *deskriptive Statistik* dient rein der Beschreibung der Daten, etwa durch geeignete Wahl von Statistiken, die die Daten zusammenfassen. Anders ist dies bei der hier behandelten *schließenden* oder *induktiven Statistik*. Die Aufgabe ist hier, mit Hilfe von stochastischen Modellen Aussagen darüber zu treffen, welchen Annahmen den Daten zugrunde liegen könnten.

6.1 Grundlagen

Bevor wir statistische Konzepte einführen, betrachten wir folgendes Beispiel.

B 6.1 *Erfolgswahrscheinlichkeit beim Münzwurf*: Eine Münze wird 100 mal geworfen. Dabei ist die Wahrscheinlichkeit für *Kopf* unbekannt. Von den 100 Würfen erhalten wir 59 mal Kopf.

Unsere statistischen Überlegungen gehen nun von der Vorstellung aus, dass die 100 Münzwürfe die Realisierung einer Zufallsvariable $X = (X_1, \ldots, X_{100})$ sind, wobei die einzelnen Zufallsvariablen $X_1, \ldots, X_{100}$ unabhängig und identisch verteilt sind mit

$$X_i = \begin{cases} 1, & \text{falls der } i\text{-te Wurf } \textit{Kopf} \text{ zeigt,} \\ 0, & \text{sonst.} \end{cases}$$

Es gilt

$$P[X_i = 1] = p.$$

Jetzt ist $X_1 + \cdots + X_n$ die Anzahl der Kopf-Würfe. Als Summe von n unabhängigen Bernoulli-verteilten Zufallsvariablen ist diese Summe $B(n = 100, p)$-verteilt. Wichtig ist, dass zwar $n = 100$ bereits fest steht (schließlich wissen wir ja, dass wir 100 mal die Münze geworfen haben), nicht jedoch p. In dieser Situation gibt es zwei *statistische Probleme*.

- *Schätzproblem:* Wir versuchen, den Parameter p zu schätzen. Wir werden hierzu aus den Daten (59 Erfolge aus 100 Versuchen) einen Wert $\widehat{p}$ ableiten (*Punktschätzer*).

units ¡1.3cm,.9cm¿ x from -2.5 to 13, y from 1 to 4 (.) axes ratio 2:1 360 degrees from 0 2.1 center at 0 1.5 axes ratio 2:1 360 degrees from 2 2.1 center at 2 1.5 Θ' [cC] at 1.9 1.6 Θ [cC] at 0 1.6 axes ratio 2:1 360 degrees from 6 2.5 center at 6 1.5 ¡0.2cm¿ [0.375,1] from 6 1.5 to 2.2 1.5 ¡0.2cm¿ [0.375,1] from .2 1.5 to 1.7 1.5 T [cC] at 3.5 1.8 q [cC] at 1.0 1.8 E [cC] at 5.5 1.8 X [cC] at 5 3.2 3.5 3 5 2.85 6.5 1.5 / ¡0.2cm¿ [0.375,1] from 6.34 1.7 to 6.5 1.5

Abbildung 6.1: Veranschaulichung von statistischen Modellen: Θ ist der Parameterraum, manchmal interessiert uns aber nicht θ, sondern $q(\theta)$. Diese Abbildung landet in dem Raum Θ'. Eine Statistik ist eine Funktion von X und landet natürlich in dem Raum, der uns interessiert ($q(\theta)$), also Θ'.

Alternativ kann man auch ein Intervall $[a, b]$ angeben, in dem der wahre Parameter p mit hoher Wahrscheinlichkeit liegt (*Intervallschätzer*).

- *Testproblem:* Stellen wir uns vor, der Werfer der Münze behauptet, dass die Münze fair ist, also $p = \frac{1}{2}$ gilt. Dieser Meinung können wir skeptisch gegenüber stehen, da ja sogar 59 aus 100 Würfen ein Erfolg waren. Wir können versuchen, die Hypothese $p = \frac{1}{2}$ zu testen. Das bedeutet, dass wir untersuchen, wie gut die Hypothese mit den Daten in Einklang steht.

Wie im Beispiel gesehen besteht, ein statistisches Experiment aus einer (oder mehreren) Zufallsvariablen (etwa die Anzahl der Erfolge) und verschiedenen (möglichen) Verteilungen der Zufallsvariable; hier $B(n = 100, p)$ für variierendes p. Dies führt zur Definition des statistischen Modells. Siehe auch Abbildung 6.1.

Definition 6.1. Seien E, Θ Mengen. Ein *statistisches Modell* ist ein Paar $(X, (P_\theta)_{\theta\in\Theta})$, wobei X eine Zufallsvariable mit Zielbereich E und $(P_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen ist. Die Menge Θ heißt *Parameterraum*, die Menge E *Beobachtungsraum*. Eine Zufallsvariable $T(X)$ mit $T : E \to \Theta'$ mit einer Menge Θ' heißt *Statistik*.

Hierbei ist also X der Vektor der beobachteten Daten und eine Statistik ist eine Funktion der Daten, die Werte im Parameterraum annimmt. Im Fall wo X eine Dichte hat, erhalten wir also, dass für alle $\theta \in \Theta$ eine Dichte f_θ existiert. Typischerweise betrachten wir Punktschätzer, für welche stets $\Theta' = \Theta$ gilt. Für Intervallschätzer ist Θ' die Menge der Intervalle welche in Θ liegen.

B 6.2 *Erfolgswahrscheinlichkeit beim Münzwurf*: In Beispiel 6.3 ist einfach X die Anzahl der Erolge und $P_p := B(n = 100, p)$ (wobei wir θ durch p ersetzt haben).

6.2 Schätzprobleme

Ein Schätzproblem setzt sich zum Ziel, einen möglichst guten Schätzer zu finden. Hierbei interessiert uns nicht immer direkt θ, sondern oft ist eine Funktion $q(\theta)$ von Interesse. Besonderer Bedeutung kommt der Definition bei, was eigentlich unter einem guten Schätzer zu verstehen ist.

Definition 6.2. Wir betrachten ein statistisches Modell $(X, (P_\theta)_{\theta\in\Theta})$ und eine Funktion $q : \Theta \to \Theta$.

(i) Ein *Punktschätzer* ist eine Statistk $T : E \to \Theta$. Der Punktschätzer $T(X)$ heißt *unverzerrt* für $q(\theta)$, falls
$$E_\theta[T(X)] = q(\theta)$$
für alle $\theta \in \Theta$.

(ii) Ist X für alle $\theta \in \Theta$ diskret, so heißt
$$L(\theta, x) := P_\theta[X = x]$$
Likelihood-Funktion. Ist X für alle $\theta \in \Theta$ stetig mit Dichte f_θ, so heißt
$$L(\theta, x) := f_\theta(x)$$
Likelihood-Funktion. Gilt für einen Schätzer $\widehat{\theta}_{ML}$, dass
$$L(x, \widehat{\theta}_{ML}(x)) \geq L(x, \theta(x)), \quad x \in E$$
für alle Punktschätzer θ, so heißt $\widehat{\theta}_{ML}$ *Maximum-Likelihood-Schätzer* von θ.

Analog kann man natürlich auch ML-Schätzer für $q(\theta)$ definieren. Eine wichtige Eigenschaft ist die Konsistenz, was bedeutet, das für immer größer werdenden Beobachtungsraum der Schätzer gegen den unbekannten Parameter konvergiert.

Definition 6.3. Sei $((P_\theta^n)_{\theta\in\Theta})_{n=1,2,\dots}$ eine Folge statistischer Modelle mit Beobachtung (X^n) und $\mathbf{T} := (T_1, T_2, \dots)$ eine Folge von Schätzern für $q(\theta)$. Die Folge $\mathbf{T}$ heißt *konsistent*, falls
$$P_\theta^n\left[|T_n(X^n) - q(\theta)| \geq \varepsilon\right] \xrightarrow{n\to\infty} 0$$
für alle $\varepsilon > 0, \theta \in \Theta$.

Bemerkung 6.4. Sei X diskret und $X = x$. Ein Maximum-Likelihood-Schätzer ist ein Parameter θ, unter dem die Wahrscheinlichkeit, die Daten $X = x$ zu beobachten – das ist $P_\theta[X = x]$ – maximal ist.

B 6.3 *Erfolgswahrscheinlichkeit beim Münzwurf*: In der Situation aus Beispiel 6.1 hieß der Parameter p anstatt θ. Es ist $E = \{0, \dots, 100\}$ und X die Anzahl der Erfolge in $n = 100$ Versuche. Damit ist $P_\theta = B(n = 100, \theta)$ mit $\theta \in \Theta = [0, 1]$.

Wählen wir $T(x) = {}^{x}/_{n} = {}^{x}/_{100}$, so ist $\widehat{\theta} := T(X)$ ein Schätzer für den unbekannten Parameter θ. Wird 59-mal Kopf beobachtet, so ist

$$\widehat{\theta} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{59}{100} = 0.59.$$

Da $\widehat{\theta}$ von den Daten abhängt, ist der Schätzer zufällig. Unter welchen Umständen ist der Schätzer $\widehat{\theta}$ gut? Nehmen wir an, wir wüssten den wahren Parameter p. Dann leistet $\widehat{\theta}$ zumindest im Mittel das gewünschte[1]

$$E_\theta[\widehat{\theta}] = \frac{1}{n} E_\theta[X_1 + \cdots + X_n] = \theta.$$

Wir sagen auch, der Schätzer $\widehat{\theta}$ ist erwartungstreu (oder unverzerrt oder unbiased).

Eine weitere wünschenwerte Eigenschaft eines Schätzers ist, dass er immer besser wird, je größer die zu Grunde liegende Datenmenge ist. Eine große Datengrundlage bedeutet in unserem Fall, dass die Münze oft geworfen wurde, also n groß ist. Aus dem schwachen Gesetz großer Zahlen wissen wir, dass

$$P_\theta\big[|\widehat{\theta} - \theta| \geq \varepsilon\big] = P_\theta\Big[\Big|\frac{X_1 + \cdots + X_n}{n} - E_\theta[X_1]\Big| \geq \varepsilon\Big] \xrightarrow{n\to\infty} 0$$

für alle $\varepsilon > 0$. Die Eigenschaft, dass $\widehat{\theta}$ mit hoher Wahrscheinlichkeit immer näher am wahren Wert θ liegt, wenn mehr Daten zur Verfügung stehen, hatten wir Konsistenz genannt.

B 6.4 *Erfolgswahrscheinlichkeit beim Münzwurf*: Betrachten wir wieder das Beispiel der Schätzung der Erfolgswahrscheinlichkeit beim Münzwurf. Wir hatten $X = 59$ Erfolge bei $n = 100$ Erfolgen verzeichnet. Unter P_θ ist X nach $B(n = 100, \theta)$ verteilt, also ist

$$L(X, \theta) = \binom{n}{X} \theta^X (1 - \theta)^{n-X}.$$

die Likelihood-Funktion; siehe auch Abbildung 6.2. Um diese für gegebenes X zu maximieren, berechnen wir den Wert θ, für den $\log L(x, \theta)$ maximal ist. Die Bestimmung des Maximums der log-Likelihood-Funktion $\log L(x, \theta)$ genügt, da log eine streng monotone Funktion ist. Wir berechnen

$$\frac{\partial \log L(x, \theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\left(\log \binom{n}{x} + x \log \theta + (n - x) \log(1 - \theta)\right) = \frac{x}{\theta} - \frac{n - x}{1 - \theta}.$$

[1] Wir schreiben E_θ für den Erwartungswert unter dem Maß P_θ.

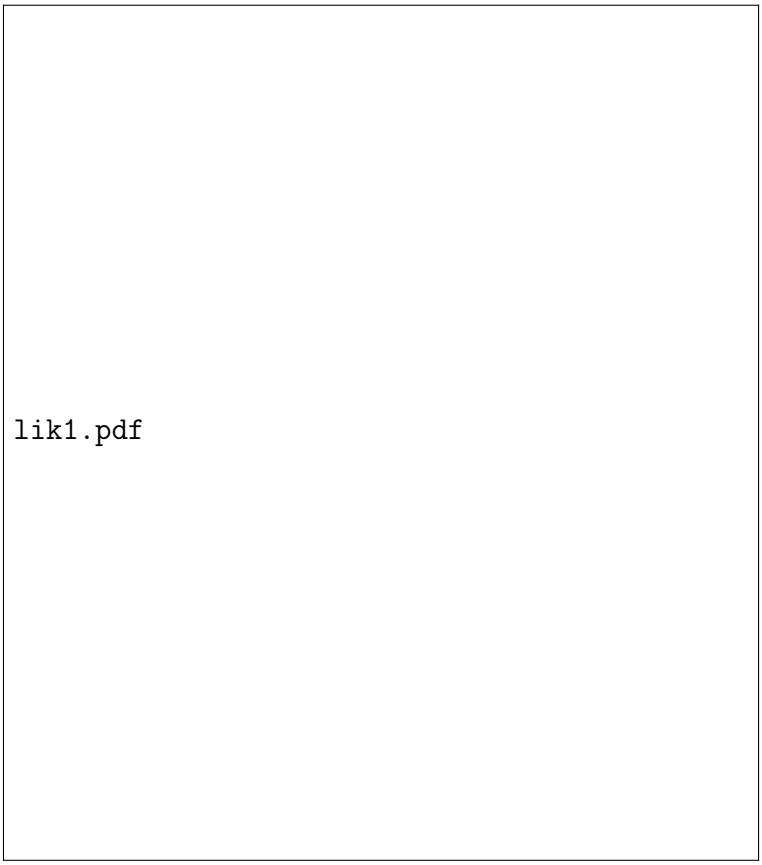


Abbildung 6.2: Die Likelihood-Funktion beim Münzwurf für $X = 23$ aus Beispiel 6.4.

Am Maximum $\widehat{\theta}_{ML}$ muss $\frac{\partial \log L(x,\theta)}{\partial \theta} = 0$ sein, also ist

$$(1 - \widehat{\theta}_{ML})X = \widehat{\theta}_{ML}(n - X), \qquad \widehat{\theta}_{ML} = \frac{X}{n}$$

ein Maximum-Likelihood-Schätzer für θ. Da es nur ein einziges Maximum der Likelihood gibt, ist dies auch der einzige Maximum-Likelihood-Schätzer.

Soeben haben wir gesehen, dass der Maximum-Likelihood-Schätzer für eine Summe von Zufallsvariablen (was der Anzahl an Erfolgen im Münzwurf entspricht) gerade durch den Mittelwert gegeben ist. Die Verwendung des Mittelwertes für eine solche Schätzung ist generell eine gute Idee, wie wir nun zeigen werden.

Definition 6.5. Sei $X = (X_1, \ldots, X_n)$ ein Vektor von Zufallsvariablen. Dann heißt

$$\bar{X} := \frac{1}{n}(X_1 + \cdots + X_n)$$

Mittelwert von X und

$$s^2(X) := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Stichprobenvarianz von X.

Bemerkung 6.6. Der Faktor $\frac{1}{n-1}$ in der Stichprobenvarianz sorgt dafür, dass sie unter der i.i.d.-Annahme unverzerrt ist, das heißt der Erwartungswert der Stichprobenvarianz ist die Varianz. Wird der Faktor $\frac{1}{n}$ verwendet, ist das nicht mehr der Fall.

Theorem 6.7.
Sei $(P_\theta)_{\theta\in\Theta}$ ein statistisches Modell, wobei $X_1, \dots, X_n$ unter allen P_θ identisch verteilt sind und $\mu_\theta := E_\theta[X_1] < \infty$.

(i) Der arithmetische Mittelwert $\bar{X}$ ist ein unverzerrter Schätzer für μ_θ. Sind außerdem $X_1, \dots, X_n$ unter allen P_θ unabhängig mit $\mathrm{Var}_\theta[X_1]$, so ist $\bar{X}$ auch konsistent.

(ii) Sei $n \geq 2$ und $X = (X_1, \cdots, X_n)$ so, dass $X_1, \dots, X_n$ unter allen P_θ paarweise unkorreliert und identisch verteilt sind mit $\sigma_\theta^2 := \mathrm{Var}_\theta[X_1] < \infty$. Dann ist die empirische Varianz $s^2(\mathbf{X})$ ein unverzerrter Schätzer für σ_θ^2.

Beweis. Zu Teil (i): Zunächst ist

$$E_\theta[\bar{X}] = \tfrac{1}{n}(E_\theta[X_1] + \cdots + E_\theta[X_n]) = E_\theta[X_1] = \mu_\theta,$$

was bereits die Unverzerrtheit von $\bar{X}$ als Schätzer von μ_θ zeigt. Für die Konsistenz berechnen wir für $\varepsilon > 0$

$$P_\theta[|\bar{X} - \mu_\theta| \geq \varepsilon] = P_\theta\left[\left|\frac{X_1 + \cdots + X_n}{n} - E_\theta[X_1]\right| \geq \varepsilon\right] \xrightarrow{n\to\infty} 0$$

nach dem schwachen Gesetz der großen Zahlen. Für (ii) schreiben wir zunächst

$$E_\theta[s^2(\underline{X})] = \tfrac{1}{n-1}\sum_{i=1}^{n} E_\theta[X_i^2 - 2X_i\bar{X} + \bar{X}^2] = \tfrac{n}{n-1}E_\theta[X_1^2 - 2X_1\bar{X} + \bar{X}^2].$$

Nun ist

$$\begin{aligned} E_\theta[X_1^2] &= \mu_\theta^2 + \sigma_\theta^2, \\ E_\theta[X_1\bar{X}] &= \mu_\theta^2 + \tfrac{1}{n}\sigma_\theta^2, \\ E_\theta[\bar{X}^2] &= E_\theta[X_1\bar{X}], \end{aligned}$$

also

$$E_\theta[s^2(\underline{X})] = \tfrac{n}{n-1}E_\theta[X_1^2 - X_1\bar{X}] = \sigma_\theta^2,$$

was die Unverzerrtheit bereits zeigt. □

B 6.5 *Erfolgswahrscheinlichkeit beim Münzwurf*: Betrachten wir noch einmal das einführende Beispiel der Schätzung des Erfolgsparameters in einem Münzwurf in n Versuchen. Wir wählen dazu die Notation aus den Beispielen 6.1 und 6.3. Der Schätzer $\widehat{\theta} = \widehat{\theta}(X) = Xn^{-1}$ ist unverzerrt und konsistent, wie wir soeben gezeigt haben.

Ein weiterer Schätzer für θ wäre $\widehat{\theta'} = X_1$. In diesem Fall nimmt $\widehat{\theta'}$ nur die beiden Werte 0 und 1 an. Dieser Schätzer ist ebenfalls erwartungstreu, denn

$$E_\theta[\widehat{\theta'}] = E_\theta[X_1] = \theta.$$

Allerdings ist $\widehat{\theta'}$ nicht konsistent, da

$$P_\theta[|\widehat{\theta'} - \theta| > \varepsilon] = 1,$$

falls $\varepsilon < \min(\theta, 1-\theta)$, unabhängig von n.

B 6.6 *Maximum-Likelihood-Schätzer bei Normalverteilungen*: Wir betrachten den Fall einer unabhängigen, normalverteilten Stichprobe. Sei $(X, (P_{(\mu,\sigma^2)})_{\mu\in\mathbb{R},\sigma^2\in\mathbb{R}_+})$ so, dass $X = (X_1, \ldots, X_n)$ und $X_1, \ldots, X_n$ unter $P_{(\mu,\sigma^2)}$ unabhängig und identisch nach $\mathcal{N}(\mu, \sigma^2)$ verteilt sind. Der Parameter θ ist in diesem Fall der Vektor (μ, σ^2).

Wir berechnen nun die Maximum-Likelihood-Schätzer für μ und σ^2. Genau wir im letzten Beispiel berechnen wir zunächst die log-Likelihood-Funktion

$$\begin{aligned}\log L((X_1, \ldots, X_n), (\mu, \sigma^2)) &= \log\left(\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^2}\Big)\right)\\ &= -n\log\sigma - \sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^2} + C,\end{aligned}$$

wobei C weder von μ noch von σ abhängt. Ableiten nach μ und σ ergibt

$$\begin{aligned}\frac{\partial \log L((X_1, \ldots, X_n), (\mu, \sigma^2))}{\partial \mu} &= \sum_{i=1}^n \frac{X_i - \mu}{\sigma^2},\\ \frac{\partial \log L((X_1, \ldots, X_n), (\mu, \sigma^2))}{\partial \sigma} &= -\frac{n}{\sigma} + \sum_{i=1}^n \frac{(X_i-\mu)^2}{\sigma^3}.\end{aligned}$$

Für die Maximum-Likelihood-Schätzer $\widehat{\mu}_{ML}$ und $\widehat{\sigma}^2_{ML}$ gilt notwendigerweise

$$\begin{aligned}\sum_{i=1}^n (X_i - \widehat{\mu}_{ML}) &= 0,\\ \frac{n}{\widehat{\sigma}_{ML}} - \sum_{i=1}^n \frac{(X_i - \widehat{\mu}_{ML})^2}{\widehat{\sigma}^3_{ML}} &= 0.\end{aligned}$$

Die Maximum-Likelihood-Schätzer sind also gegeben durch

$$\widehat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X},$$
$$\widehat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}s^2(X).$$

Insbesondere sehen wir, dass $\bar{X}$ nicht nur erwartungstreu und konsistent (siehe Theorem 6.7) ist, sondern auch ein Maximum-Likelihood-Schätzer für μ. Allerdings ist der Maximum-Likelihood-Schätzer für σ^2 nicht erwartungstreu, wie man aus Theorem 6.7 abliest. Immerhin ist $\widehat{\sigma}^2_{ML}$ für große n annähernd erwartungstreu, da $\widehat{\sigma}^2_{ML} - s^2(X) \xrightarrow{n\to\infty} 0$.

6.3 Intervallschätzer oder Konfidenzintervalle

Sei $T(X)$ ein Schätzer von $q(\theta) \in \mathbb{R}$. Für eine vernünftige Schätzung ist es essenziell, neben dem Schätzwert auch ein Maß für die Präzision des Schätzverfahrens anzugeben. Ziel dieses Abschnittes ist, die Präzision oder den Fehler von T zu bestimmen. Dabei gehen wir folgendem Ansatz nach: Wir suchen zufällige Grenzen $\underline{T}(X) \le q(\theta) \le \overline{T}(X)$, so dass die Wahrscheinlichkeit, dass $q(\theta)$ von $[\underline{T}(X), \overline{T}(X)]$ überdeckt wird, ausreichend hoch ist. Ein solches zufälliges Intervall nennen wir *Zufallsintervall*.

Definition 6.8. Ein durch $\underline{T}(X) \le \overline{T}(X)$ gegebenes Zufallsintervall $[\underline{T}(X), \overline{T}(X)]$ für welches für alle $\theta \in \Theta$ gilt, dass

$$P_\theta\big[q(\theta) \in [\underline{T}(X), \overline{T}(X)]\big] \ge 1 - \alpha, \tag{6.1}$$

heißt $(1-\alpha)$-*Konfidenzintervall* für $q(\theta)$ zum *Konfidenzniveau* $1 - \alpha \in [0, 1]$.

Hierbei verwenden wir folgenden Sprachgebrauch: Ein $(1-\alpha)$-Konfidenzintervall bedeutet ein $(1-\alpha)\cdot 100$ %-Konfidenzintervall; ist etwa $\alpha = 0.05$, so verwenden wir synonym die Bezeichnung 0.95-Konfidenzintervall und 95%-Konfidenzintervall.

Wir sind natürlich daran interessiert, für ein vorgegebenes Konfidenzniveau das kleinste Intervall zu finden, welches die Überdeckungseigenschaft (6.1) erfüllt.

Handelt es sich um ein symmetrisches Intervall, so nutzen wir die Schreibweise

$$a \pm b := [a - b, a + b].$$

B 6.7 *Normalverteilung, σ bekannt: Konfidenzintervall*: Seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\theta, \sigma^2)$ und σ^2 sei bekannt. Als Schätzer für den Erwartungswert θ verwenden wir den ML-Schätzer

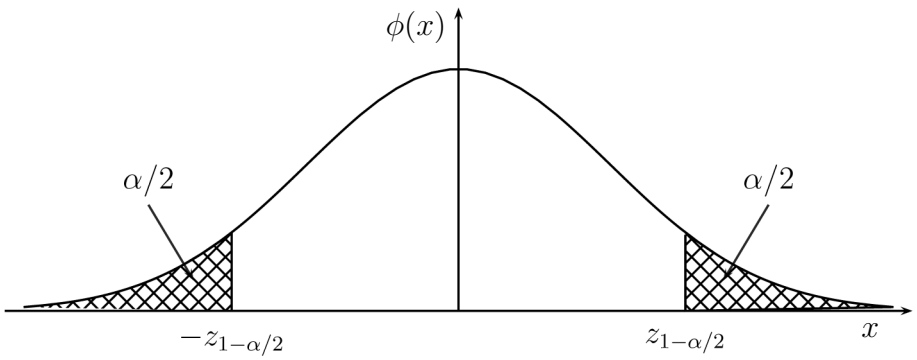


Abbildung 6.3: Dichte der Standardnormalverteilung mit den $\alpha/2$ und $1-\alpha/2$-Quantilen.

$\bar{X}$. Da die $\mathcal{N}(\theta, \sigma^2)$-Verteilung symmetrisch um θ ist, liegt es nahe als Konfidenzintervall ein symmetrisches Intervall um $\bar{X}$ zu betrachten. Für $c > 0$ gilt

$$P_\theta \left[\bar{X} - c\frac{\sigma}{\sqrt{n}} \leq \theta \leq \bar{X} + c\frac{\sigma}{\sqrt{n}} \right] = P_\theta \left[\left| \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} \right| \leq c \right].$$

Da $\frac{\bar{X}-\theta}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, folgt

$$P_\theta \left[\left| \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} \right| \leq c \right] = \Phi(c) - \Phi(-c) = 2\Phi(c) - 1.$$

Da wir das kleinste Konfidenzintervall suchen, welches die Überdeckungseigenschaft (6.1) erfüllt, suchen wird ein $c > 0$ so, dass $2\Phi(c) - 1 = 1 - \alpha$ gilt. Mit

$$z_a := \Phi^{-1}(a)$$

sei das *a-Quantil* der Standardnormalverteilung bezeichnet. Dann ist das symmetrische Intervall

$$\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

ein $(1-\alpha)$-Konfidenzintervall für θ; siehe Abbildung 6.3. Da $z_{0.975} = 1.96$ gilt, ist in einer Stichprobe mit $\bar{x} = 5$, $\sigma = 1$, $n = 100$ das 95%-Konfidenzintervall für θ gegeben durch 5 ± 1.96.

Die χ^2 und die t-Verteilung. Die χ^2-Verteilung entsteht als Summe von quadrierten, normalverteilten Zufallsvariablen.

Lemma 6.9. *(und Definition) Sind $X_1, \dots, X_n$ unabhängig und standardnormalverteilt, heißt*

$$V := \sum_{i=1}^{n} X_i^2$$

χ^2-verteilt *mit n Freiheitsgraden, kurz χ^2_n-verteilt. Die Dichte von V ist gegeben durch*

$$p_{\chi^2_n}(x) = \mathbb{1}_{\{x>0\}} \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}. \tag{6.2}$$

Hierbei verwenden wir die *Gamma-Funktion*, definiert durch

$$\Gamma(a) := \int_0^\infty t^{a-1} e^{-t} dt, \qquad a > 0.$$

Dann ist $\Gamma(n) = (n-1)!,\ n \in \mathbb{N}$ und $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Weiterhin gilt $E[V] = n$ und $\mathrm{Var}[V] = 2n$.

Bemerkung 6.10. Die Darstellung der Dichte in (6.2) zeigt, dass die χ^2_n-verteilte Zufallsvariable V für $n = 2$ exponentialverteilt ist mit Parameter $\frac{1}{2}$. Aus dem zentralen Grenzwertsatz folgt, dass

$$\frac{\chi^2_n - n}{\sqrt{2n}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1).$$

Möchte man ein Konfidenzintervall für den Mittelwert einer Normalverteilung mit unbekannter Varianz bilden, so muss man diese schätzen. Dabei taucht die Wurzel einer Summe von Normalverteilungsquadraten (mit Faktor $\frac{1}{n}$) im Nenner auf. Hierüber gelangt man zur t-Verteilung, welche oft auch als Student-Verteilung oder Studentsche t-Verteilung bezeichnet wird.

Definition 6.11. Ist X standardnormalverteilt und V χ^2_n-verteilt und unabhängig von X, so heißt die Verteilung von

$$T := \frac{X}{\sqrt{\frac{1}{n}V}} \tag{6.3}$$

die *t-Verteilung* mit n Freiheitsgraden, kurz t_n-Verteilung.

Lemma 6.12. *Die Dichte der t_n-Verteilung ist gegeben durch*

$$p_{t_n}(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(n/2)\Gamma(1/2)\sqrt{n}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

für alle $x \in \mathbb{R}$.

B 6.8 *Normalverteilung, μ und σ unbekannt: Konfidenzintervall*: Die Zufallsvariablen $X_1, \ldots, X_n$ seien i.i.d. mit $X_1 \sim \mathcal{N}(\mu, \sigma^2)$. Gesucht ist ein Konfidenzintervall für den Mittelwert μ, aber auch σ ist unbekannt. Wir setzen $c := t_{n-1,1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden. Man erhält mit $\theta := (\mu, \sigma^2)^\top$, dass

$$P_\theta\left[\bar{X} - \frac{cs_n}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{cs_n}{\sqrt{n}}\right] = P_{m\theta}\left[\left|\frac{\bar{X}-\mu}{s_n/\sqrt{n}}\right| \leq c\right].$$

Nun kann man zeigen, dass $\bar{X}$ von $s_n^2(X)$ unabhängig ist und $(n-1)\frac{s_n^2(X)}{\sigma^2} \sim \chi^2_{n-1}$. Wir erhalten nach Definition 6.11, dass

$$T_{n-1}(X) := \frac{\sqrt{n}(\bar{X}-\mu)}{s_n(X)} = \frac{\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}}{\sqrt{\frac{1}{n-1}\frac{(n-1)s_n^2(\mathbf{X})}{\sigma^2}}}$$

t_{n-1}-verteilt ist, also unabhängig von θ! Somit ergibt sich folgendes Konfidenzintervall für μ:

$$\bar{X} \pm \frac{s_n}{\sqrt{n}}\, t_{n-1,1-\alpha/2}.$$

6.4 Testprobleme

Bisher haben wir Schätzverfahren betrachtet und entwickelt, welche man beispielsweise nutzen kann, um aus den Daten die Wirksamkeit einer Therapie zu schätzen. Allerdings ist man oft nicht direkt an dem Schätzwert interessiert, sondern man möchte entscheiden, ob diese Therapie hilft oder nicht. Hierfür wird man wegen der Zufälligkeit des Problems keine absolute Entscheidung treffen können, sondern zu jeder Zeit muss man eine gewisse Wahrscheinlichkeit für eine Fehlentscheidung akzeptieren, ähnlich wie bei den Konfidenzintervallen.

Im Folgenden führen wir das Konzept des statistischen Tests zur Überprüfung von Hypothesen auf Basis einer Stichprobe ein. Stets gehen wir von einem statistischen Modell $\{P_\theta : \theta \in \Theta\}$ mit $X \sim P_\theta$ aus. Allerdings zerlegt die betrachtete Fragestellung den Parameterraum disjunkt in die zwei Hypothesen Θ_0 und Θ_1 mit $\Theta = \Theta_0 \oplus \Theta_1$, was gleichbedeutend ist mit $\Theta_0 \cap \Theta_1 = \emptyset$ und $\Theta_0 \cup \Theta_1 = \Theta$. Die beiden Parameterbereiche Θ_0 und Θ_1 stehen für unterschiedliche Hypothesen. Im obigen Beispiel würde man Θ_0 als den Bereich wählen, in welchem die Therapie nicht hilft; in dem Bereich Θ_1 hilft hingegen die Therapie. Wir verwenden die folgenden Bezeichnungen:

$H_0 = \{\theta \in \Theta_0\}$ heißt *Null-Hypothese* und

$H_1 = \{\theta \in \Theta_1\}$ heißt *Alternative*.

Oft schreiben wir hierfür $H_0 : \theta \in \Theta_0$ gegen $H_1 : \theta \in \Theta_1$. Die Bezeichnung Null-Hypothese stammt vom englischen Begriff *to nullify* = entkräften, widerlegen. Wie wir später sehen werden, ist die Hypothese, die widerlegt werden soll, stets als Null-Hypothese zu wählen.

Besteht Θ_0 aus einem einzigen Element, $\Theta_0 = \{\theta_0\}$, so spricht man von einer *einfachen* Hypothese, ansonsten handelt es sich um eine *zusammengesetzte* Hypothese. Ist $\Theta \subset \mathbb{R}$ und die Alternative von der Form $\Theta_1 = \{\theta : \theta \neq \theta_0\}$, so nennt man sie *zweiseitig*; ist sie von der Form $\Theta_1 = \{\theta : \theta > \theta_0\}$, so heißt sie *einseitig*.

B 6.9 *Erfolgswahrscheinlichkeit beim Münzwurf*: Nehmen wir an, der Werfer der Münze behauptet, sie sei fair, also $\theta = \frac{1}{2}$. Wir sind also interessiert an einem Test mit $\Theta_0 = \{\frac{1}{2}\}$. Können wir diese Hypothese aufgrund der Daten verwerfen? Zunächst stellen wir fest, dass wir prinzipiell zwei Arten von Fehlern mit unserer Entscheidung machen können. Wenn wir die Hypothese verwerfen, könnte sie doch wahr sein, und wenn wir die Hypothese nicht verwerfen, könnte sie doch richtig sein.

Da wir nicht grundlos dem Werfer der Münze widersprechen wollen, wollen wir die Wahrscheinlichkeit, dass wir die Hypothese ablehnen (wir dem Werfer der Münze widersprechen), obwohl sie wahr ist (die Hypothese des Werfers richtig ist), kontrollieren. Das bedeutet, dass

$$P_{p=1/2}\big[\text{Hypothese verwerfen}\big] \leq \alpha$$

für ein anfangs gewähltes $\alpha \in (0, 1)$ sein soll. Klar ist, dass damit die Hypothese umso seltener abgelehnt werden kann, je kleiner α ist. Nun kommen wir zu der Regel, mit der wir die Hypothese ablehnen wollen. In unserem Beispiel haben wir für die Hypothese $p = \frac{1}{2}$ zu viele (59 von 100) Erfolge. Wir würden die Hypothese ablehnen wollen, wenn

$$P_{p=1/2}[\text{Daten mindestens so extrem wie tatsächliche Daten}] \leq \alpha. \tag{6.4}$$

Wir wählen $\alpha = 5\,\%$. Für $X \sim B(n = 100, p = \frac{1}{2})$ verteilt, erwarten wir 50 Erfolge. Um die Wahrscheinlichkeit einer Abweichung, die größer ist als die der Daten zu berechnen,

betrachten wir eine nach $N(0,1)$ verteilte Zufallsvariable Z und berechnen

$$\begin{aligned} &P_{p=1/2}\big[|X_1+\cdots+X_n-50|\geq 9\big] \\ &\qquad = 1-P_{p=1/2}\bigg[-\frac{9}{\sqrt{np(1-p)}}<\frac{Y_n-np}{\sqrt{np(1-p)}}<\frac{9}{\sqrt{np(1-p)}}\bigg] \\ &\qquad \approx 1-P_{p=1/2}\big[-1.8\leq Z\leq 0.1.8\big]\approx 7.19\,\% \end{aligned}$$

Da dieser Wert größer als $\alpha = 5\,\%$ ist, kann die Hypothese nicht verworfen werden, siehe (6.4).

Wir beginnen mit der Einführung wichtiger Begriffe wie Teststatistik, Nullhypothese, Alternative, Ablehnungsbereich, Signifikanzniveau und p-Wert.

Definition 6.13. Sei $(P_\theta)_{\theta\in\Theta}$ ein statistisches Modell, E der Zielbereich der Zufallsvariable X, und $\Theta_0,\Theta_1\subseteq\Theta$ disjunkt mit $\Theta_0\cup\Theta_1=\Theta$.

(i) Ein Paar (T,C) mit einer Statistik $T:E\to E'$ und $C\subseteq E'$ heißt *statistischer Test* von

$$H_0:\theta\in\Theta_0 \text{ gegen } H_1:\theta\in\Theta_1.$$

Hier heißt C *kritischer Bereich* des Tests, H_0 heißt *Nullhypothese* und H_1 heißt *Alternative*. Man sagt, der Test (T,C) hat *(Signifikanz-)Niveau* $\alpha\in[0,1]$, falls

$$\sup_{\theta\in\Theta_0} P_\theta[T(X)\in C]\leq\alpha.$$

Falls $T\in C$ ist, sagt man, dass H_0 verworfen wird. Falls $T\notin C$, sagt man, dass H_0 akzeptiert wird.

Sei nun (T,C) ein Test der Nullhypothese H_0 gegen H_1.

(ii) Die Hypothese $H:\theta\in\Theta_0$ (die entweder H_0 oder H_1 sein kann) heißt *einfach* wenn $\Theta_0=\{\theta_0\}$ für ein $\theta_0\in\Theta$. Andernfalls heißt H *zusammengesetzt*.

(iii) Ist $\Theta=(\underline{\theta},\overline{\theta})$ ein Intervall (wobei $\underline{\theta}=-\infty$ und $\overline{\theta}=\infty$ zugelassen sind und die Intervalle auch abgeschlossen sein können), so heißt der Test (T,C) *einseitig*, falls $C=(\underline{\theta},\theta^*)$ oder $C=(\theta^*,\overline{\theta})$. Falls $C=(\underline{\theta},\theta^*)\cup(\theta^*,\overline{\theta})$, so heißt der Test *zweiseitig*.

(iv) Der Test (T,C) heißt *unverfälscht*, falls

$$P_{\theta_0}[T(X)\in C]\leq P_{\theta_1}[T(X)\in C]$$

für alle $\theta_0\in\Theta_0,\theta_1\in\Theta_1$ gilt.

Bemerkung 6.14 (Interpretation und Fehler eines Tests)**.**

(i) Einen statistischen Test hat man sich am besten so vorzustellen (siehe auch das nächste Beispiel): die Daten sind gegeben durch die Zufallsvariable X. Diese Daten fasst man durch die meist reellwertige Funktion T zusammen zur Teststatistik $T(X)$. Die Daten können entweder nach P_θ mit $\theta \in \Theta_0$ (d.h. die Nullhypothese ist richtig) oder mit $\theta \in \Theta_1$ (d.h. die Alternative ist richtig) verteilt sein. Ziel ist es, die Nullhypothese genau dann (anhand der Daten X) abzulehnen, wenn H_1 richtig ist. Der Ablehnungsbereich C ist so gewählt, dass H_0 genau dann abgelehnt wird, wenn $Y \in C$. Dabei können zwei verschiedene Arten von Fehler auftreten; siehe auch Tabelle 6.1.

	H_0 abgelehnt	H_0 nicht abgelehnt
H_0 richtig	Fehler erster Art	richtige Entscheidung
H_0 falsch	richtige Entscheidung	Fehler zweiter Art

Tabelle 6.1: Die möglichen Fehler eines statistischen Tests.

Gehen wir zunächst davon aus, dass $\theta \in \Theta_0$. Hat der Test ein Niveau α, so wissen wir, dass $P_\theta[T(X) \in C] \leq \alpha$. Da H_0 genau dann abgelehnt wird, wenn $T(X) \in C$, wissen wir also, dass die Nullhypothese höchstens mit Wahrscheinlichkeit α abgelehnt wird, wenn sie zutrifft. Damit hat man also die Wahrscheinlichkeit, die Nullhypothese abzulehnen, falls sie zutrifft, durch α beschränkt. Falls $T(X) \in C$, aber $\theta \in \Theta_0$, die Nullyhpothese also irrtümlicherweise verworfen wird, sprechen wir über einen *Fehler erster Art* (dessen Wahrscheinlichkeit durch α kontrolliert wird).

Geht man davon aus, dass $\theta \in \Theta_1$, liegt eine Fehlentscheidung genau dann vor, wenn $T(X) \notin C$, die Nullhypothese also nicht abgelehnt wird. In diesem Fall sprechen wir von einem *Fehler zweiter Art*. Das Niveau des Tests liefert keinen Anhaltspunkt dafür, mit welcher Wahrscheinlichkeit ein solcher Fehler auftritt.

(ii) Es ist zu beachten, dass H_0 und H_1 nicht symmetrisch behandelt werden. Dies liegt daran, dass man gleichzeitig nur einen Fehler (erster oder zweiter Art) kontrollieren kann. Man stellt den Test so auf, dass der Fehler 1. Art kontrolliert wird. Aus diesem Grund bekommt der Nullhypothese eine besondere Rolle zu: Wegen dieser Praxis ist die Nullhypothese genau so zu wählen, dass eine Ablehnung der Nullhypothese möglichst sicher auf die Richtigkeit der Alternative zurückzuführen ist. Wir betrachten das Beispiel der Münzwürfe aus Beispiel 6.1. Bevor wir die Daten des Experimentes erhoben haben, haben wir die Vorstellung, dass der Würfel gezinkt

sein könnte. Außerdem legen wir ein Signifikanzniveau α fest (was in der Praxis oft $\alpha = 5\,\%$ ist). Um unsere Vorstellung über die Münze zu überprüfen, testen wir

$$H_0 : \text{die Münze ist nicht gezinkt, } p = \tfrac{1}{2}$$
$$\text{gegen}$$
$$H_1 : \text{die Münze ist gezinkt, } p \neq \tfrac{1}{2}.$$

Kommt es nämlich jetzt zu einer Ablehnung von H_0, so wissen wir, dass dies mit Wahrscheinlichkeit höchstens α dann passiert, wenn H_0 wahr ist, die Münze also nicht gezinkt ist. Damit können wir uns relativ sicher sein, dass die Ablehnung der Nullhypothese darauf zurückzuführen ist, dass H_1 zutrifft. Damit ist unsere Vorstellung, dass die Münze gezinkt ist, bei Ablehnung der Nullhypothese höchstwahrscheinlich bestätigt.

(iii) Die Forderung von unverfälschten Tests ist klar zu verstehen: Da wir H_0 dann ablehnen, wenn $T(X) \in C$, soll zumindest die Wahrscheinlichkeit, dass H_0 abgelehnt wird, unter P_{θ_1}, $\theta_1 \in \Theta_1$ größer sein als für P_{θ_0}, $\theta_0 \in \Theta_0$.

Bemerkung 6.15 (p-Werte und alternative Definition eines Tests)**.**

(i) Wir betrachten das Ereignis $T(X) = t$, d.h. dass die Teststatistik gleich der Beobachtung t ist. Dann heißt der Wert

$$p_t := \sup_{\theta \in \Theta_0} P_\theta(T(X) \text{ mindestens so extrem wie } t)$$

p-Wert des Tests für $T(X) = t$. Dabei hängt die Bedeutung davon, was 'extrem' heißt davon ab, was genau die Alternative ist. (Dies ist oftmals in konkreten Beispielen einfach zu verstehen, siehe etwa den Binomialtest, Abbildung 6.4.) Immer gilt jedoch $p_t \leq p_{t'}$, falls t mindestens so extrem wie t' ist. Es ist wichtig zu beachten, dass es dadurch einen engen Zusammenhang zwischen dem Niveau α des Tests und dem p-Wert gibt. Ist nämlich (T, C) ein Test zum Niveau α und

$$C = \{t : t \text{ mindestens so extrem wie } t_0\}$$

für ein t_0, so wird H_0 genau dann abgelehnt, wenn

$$\alpha \geq \sup_{\theta \in \Theta_0} P_\theta\big[T(X) \text{ mindestens so extrem wie } t_0\big] = p_{t_0}.$$

Ist $T(X) = t$ und gilt $p_t \leq p_{t_0}$, so wird H_0 also abgelehnt. Es genügt also, für einen Test zum Niveau α und $T(X) = t$ den Wert p_t zu bestimmen. Ist $p_t \leq \alpha$, so wird H_0 abgelehnt. Dieses Vorgehen wird bei vielen Statistik-Programmen angewendet, bei denen ausschließlich p-Werte ausgegeben werden. Dabei muss man meist angeben, was genau die Alternative ist (einseitig oder zweiseitig), damit das Programm weiß, in welche Richtungen Abweichungen als extrem zu betrachten sind.

Wir formalisieren nun noch das Eingangsbeispiel 6.1, was uns zum Binomialtest führt.

Satz 6.16 (Binomialtest)**.** *Sei $\alpha \in [0,1], n \in \mathbb{N}$ und $(P_p)_{p\in[0,1]}$ ein statistisches Modell, so dass X unter P_p nach $B(n,p)$ verteilt ist.*

(a) Ist $\Theta_0 = p^*$, $\Theta_1 = \Theta \setminus \Theta_0$, so ist $(X, \{0,\dots,k\} \cup \{l,\dots,n\})$ ein unverfälschter Test zum Niveau α, falls

$$P_{p^*}[X \le k] \le \alpha/2, \qquad P_{p^*}[X \ge l] \le \alpha/2.$$

(b) Ist $\Theta_0 = [0,p^*]$, $\Theta_1 = \Theta \setminus \Theta_0$, so ist $(X, \{k,\dots,n\})$ ein unverfälschter Test zum Niveau α, falls

$$P_{p^*}[X \ge k] \le \alpha.$$

(c) Ist $\Theta_0 = [p^*,1]$, $\Theta_1 = \Theta \setminus \Theta_0$, so ist $(X, \{0,\dots,k\})$ ein unverfälschter Test zum Niveau α, falls

$$P_{p^*}[X \le k] \le \alpha.$$

Beweis. Wir beweisen nur (c), die anderen beiden Aussagen folgen analog. Klar ist, dass der Test unverfälscht ist. Es ist außerdem

$$\sup_{p\in\Theta_0} P_p[X \in \{0,\dots,k\}] = P_{p^*}[X \le k] \le \alpha$$

nach Voraussetzung. Also folgt bereits die Aussage. □

B 6.10 *Binomialtest*: Sei $\alpha = 5\,\%$, $n = 100$ und $(P_p)_{p\in[0,1]}$ wie in obigem Satz. Wir wollen nun

$$H_0 : p = 1/2 \text{ gegen } H_1 : p \neq 1/2$$

testen, wenn wir in 100 Versuchen 59 Erfolge erzielt haben. Nach Satz 6.16 ist der kritische Bereich von der Form $\{0,\dots,k\} \cup \{l,\dots,100\}$. Es ist

$$P_{p=1/2}[X \le 40] + P_{p=1/2}[X \ge 60] \approx 5.69\,\%.$$

Da $40 < 59 < 60$, liegt 59 nicht im Ablehnungsbereich von H_0. Damit kann die Nullhypothese aufgrund der Daten ($X = 59$) nicht abgelehnt werden. Auf dasselbe Ergebnis kommt man mit Hilfe des p-Wertes. Es ist

$$P_{p=1/2}[X \text{ mindestens so extrem wie } 59] = P_{p=1/2}[X \le 41] + P_{p=1/2}[X \ge 59] \approx 8.86\,\%.$$

Da dieser Wert größer als $\alpha = 5\,\%$ ist, wird die Nullhypothese akzeptiert. Das Experiment zeigt keinen signifikanten Widerspruch dazu, dass der wahre Parameter $\theta = 0.5$ ist.

Der Binomialtest

überprüft, ob bestimmte Erfolgswahrscheinlichkeiten einer Binomialverteilung angenommen werden.

Statistisches Modell	X unter P_p nach $B(n,p)$ verteilt
Hypothesen	(a) $H_0 : p = p^*$ gegen $H_1 : p \neq p^*$ (b) $H_0 : p \leq p^*$ gegen $H_1 : p > p^*$ (c) $H_0 : p \geq p^*$ gegen $H_1 : p < p^*$
Teststatistik	X unter P_p verteilt nach $B(n,p)$
Ablehnungsbereich	(a) $\{0,\dots,k,l,\dots,n\}$ mit $P_{p^*}(X \leq k), P_{p^*}(X \geq l) \leq \alpha/2$, (b) $\{l,\dots,n\}$ mit $P_{p^*}(X \geq l) \leq \alpha$ (c) $\{0,\dots,k\}$ mit $P_{p^*}(X \leq k) \leq \alpha$,
p-Wert, falls $X = x$	(a) $P_{p^*}(X \leq x \wedge (2p^*n - x)) + P_{p^*}(X \geq x \vee (2p^*n - x))$ (b) $P_{p^*}(X \geq x)$ (c) $P_{p^*}(X \leq x)$

Abbildung 6.4: Schematische Darstellung des Binomialtests aus Proposition 6.16

6.4.1 Der einfache t-Test

Bemerkung 6.17 (Transformation von $\mathcal{N}(\mu, \sigma^2)$-verteilten Zufallsvariablen)**.** Sei $n > 1$ und $X = (X_1, ..., X_n)$ unabhängig und nach $N(\mu, \sigma^2)$ verteilt. Man überlegt sich leicht, dass dann $Y_1, ..., Y_n$ mit $Y_i := \frac{X_i - \mu}{\sqrt{\sigma^2}} \sim N(0, 1)$. Man kann mit einer längeren Rechnung zeigen, dass

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \cdot \frac{\sigma}{\sqrt{s^2(X)}} \sim t(n-1).$$

Satz 6.18 (Einfacher t-Test)**.**
Sei $\alpha \in [0, 1]$, $\mu^* \in \mathbb{R}$ und $(P_{\mu,\sigma^2})_{\mu\in\mathbb{R},\sigma^2\in\mathbb{R}_+}$ ein statistisches Modell, so dass $X_1, \ldots, X_n$ unter P_{μ,σ^2} unabhängig und nach $\mathcal{N}(\mu, \sigma^2)$ verteilt sind. Weiter sei

$$T := \sqrt{n}\frac{\bar{X} - \mu^*}{\sqrt{s^2(X)}}.$$

und $t_{n,p}$ für $p \in [0, 1]$ das p-Quantil von $t(n)$.

(a) Ist $\Theta_0 = \{\mu^*\} \times \mathbb{R}_+$, $\Theta_1 = \Theta \setminus \Theta_0$, so ist $(T, (-\infty, t_{n-1,\alpha/2}) \cup (t_{n-1,1-\alpha/2}, \infty))$ ein unverfälschter Test zum Niveau α.

(b) Ist $\Theta_0 = (-\infty, \mu^*] \times \mathbb{R}_+$, $\Theta_1 = \Theta \setminus \Theta_0$, so ist $(T, [t_{n-1,1-\alpha}, \infty))$ ein unverfälschter Test zum Niveau α.

(c) Ist $\Theta_0 = [\mu^*, \infty) \times \mathbb{R}_+$, $\Theta_1 = \Theta \setminus \Theta_0$, so ist $(T, (-\infty, t_{n-1,\alpha}])$ ein unverfälschter Test zum Niveau α.

Beweis. Wieder beweisen wir nur (c), da die anderen beiden Aussagen analog folgen. Klar ist, dass der Test unverfälscht ist. Nach obiger Bemerkung folgt, dass T unter P_{μ^*,σ^2} gerade $t(n-1)$ verteilt ist. Damit gilt

$$\sup_{\mu\geq\mu^*} P_{\mu,\sigma^2}(T \leq t_{n-1,\alpha}) = P_{\mu^*,\sigma^2}(T \leq t_{n-1,\alpha}) = \alpha,$$

woraus die Behauptung sofort folgt. □

B 6.11 *Laufzeittest bei einem Algorithmus*: Sie haben einen neuen Algorithmus programmiert und stellen ihn im Vergleich zu dem bisheringen Algorithmus. Die Laufzeit ist von den Eingabedaten abhängig und Sie simulieren sich zufällig Eingabedaten. Sie messen 10 mal und notieren die Laufzeitdifferenzen (Alter Algorithmus - Neuer Algorithmus). Sie erhlaten

$$1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4.$$

Der einfache t-Test

überprüft, ob der Erwartungswert einer Normalverteilung (oder irgendeiner Verteilung bei approximativ unendlich großer Samplegröße) gleich einer vorgegebenen Größe μ^* ist, wenn die Varianz unbekannt ist.

Statistisches Modell	$X = (X_1, \ldots, X_n)$ und $X_1, \ldots, X_n$ unter P_{μ,σ^2} unabhängig und nach $N(\mu, \sigma^2)$ verteilt
Hypothesen	(a) $H_0 : \mu = \mu^*$ gegen $H_1 : \mu \neq \mu^*$
	(b) $H_0 : \mu \leq \mu^*$ gegen $H_1 : \mu > \mu^*$
	(c) $H_0 : \mu \geq \mu^*$ gegen $H_1 : \mu < \mu^*$
Teststatistik	$T = \dfrac{\bar{X} - \mu^*}{\sqrt{s^2(X)/n}}$ unter P_{μ^*,σ^2} verteilt nach $t(n-1)$
Ablehnungsbereich	(a) $(-\infty, t_{n-1,\alpha/2}) \cup (t_{n-1,1-\alpha/2}, \infty)$,
	(b) $(t_{n-1,1-\alpha}, \infty)$
	(c) $(-\infty, t_{n-1,\alpha})$
p-Wert, falls $T = t$	(a) $2(1 - P[T \leq \lvert t \rvert])$, wenn T nach $t(n-1)$ verteilt ist
	(b) $P[T \geq t]$
	(c) $P[T \leq t]$

Abbildung 6.5: Schematische Darstellung des einfachen t-Tests aus Proposition 6.18

Sie möchten nun testen, ob die Laufzeiten unterschiedlich sind. Die Beobachtungen fassen Sie dazu Realisierungen von Zufallsvariablen $X_1, \ldots, X_{10}$, die nach $N(\mu, \sigma^2)$ verteilt sind, auf, wobei weder μ noch σ^2 bekannt sind. Sie erhalten

$$\bar{X} = 2.33, \qquad s^2(X) = 4.01.$$

Damit ist

$$T = \sqrt{10} \cdot 2.33/\sqrt{4.01} \approx 3.68.$$

Testen Sie also

$$H_0 : \mu = 0 \text{ gegen } \mu \neq 0,$$

auf dem Niveau 5 %, so ist dessen Ablehnungsbereich[2] $C = (-\infty, -2.262) \cup (2.262, \infty)$. Da $3.68 \in C$, kann man H_0 auf dem Niveau 5 % verwerfen. Das bedeutet, dass der neue Algorithmus eine signifikant andere Laufzeit hat: Und zwar ist Laufzeit alter Algorithmus abzüglich Laufzeit neuer Algorithmus im Mittel 3.68, der neue Algorithmus ist also signifikant schneller.

B 6.12 *Normalverteilte Schlafdauern*: Wir untersuchen alternativ ein Beispiel aus der Medizin. Ein Medikament wir daraufhin untersucht, ob es den Schlaf von Probanden *verlängert*[3]. Dazu wird jeweils die Schlafdauerdifferenz bei zehn Patienten notiert. Man erhält

$$1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4.$$

Diese Beobachtungen betrachten wir als Realisierungen von Zufallsvariablen $X_1, \ldots, X_{10}$, die nach $N(\mu, \sigma^2)$ verteilt sind, wobei weder μ noch σ^2 bekannt sind. Wir berechnen

$$\bar{X} = 2.33, \qquad s^2(X) = 4.01.$$

Damit ist

$$T = \sqrt{10} \cdot 2.33/\sqrt{4.01} \approx 3.68.$$

Als Nullhypothese (abzulehnen) wählen wir, dass die Schlafdauer nicht beeinflusst oder verkürzt wurde. Testet man also

$$H_0 : \mu \leq 0 \text{ gegen } \mu > 0,$$

auf dem Niveau 5 %, so ist dessen Ablehnungsbereich[4] $C = (1.833, \infty)$. Da $3.68 \in C$, kann man H_0 auf dem Niveau 5 % verwerfen. Das bedeutet, dass eine signifikante Verlängerung der Schlafdauer durch Einnahme des Medikamentes zu beobachten war.

[2] Das $t_{9,0.25}$-Quantil ist -2.262 wie man leicht in R mit `qt(0.025,9)` berechnet

[3] Wir stehen dem Medikament kritisch gegenüber und wollen nachweisen, dass das Medikament die Schlafdauer verlängert.

[4] Das $t_{9,0.95}$-Quantil ist 1.833 wie man leicht in R mit `qt(0.95,9)` berechnet.

Bemerkung 6.19 (Kontingenztabellen und Unabhängigkeitstests)**.** Wir kommen nun zum letzten hier behandelten Test, Fisher's exaktem Test. Gegeben seien Objekte mit zwei Merkmalen. (Beispielsweise Pflanzenblätter, die sowohl eine Farbe als auch eine bestimmte Form haben.) Solche Daten lassen sich in einer Kontingenztafel zusammenfassen; siehe unten. Falls beide Merkmale genau zwei verschiedene Möglichkeiten haben, lässt sich mit Hilfe des exakten Tests von Fisher bestimmen, ob beide Ausprägungen unabhängig sind.

	Merkmal 2, Möglichkeit 1	Merkmal 2 Möglichkeit 2	$\sum$
Merkmal 1, Möglichkeit 1	S_{11}	S_{12}	$S_{1\bullet}$
Merkmal 1, Möglichkeit 2	S_{21}	S_{22}	$S_{2\bullet}$
$\sum$	$S_{\bullet 1}$	$S_{\bullet 2}$	S

Satz 6.20 (Fisher's exakter Test)**.** *Sei $\alpha \in [0,1]$, $\mathcal{I}, \mathcal{J}$ Mengen mit jeweils zwei Elementen und*

$$\Theta = \{p = (p_{ij})_{i\in\mathcal{I}, j\in\mathcal{J}} \text{ Verteilungsgewichte einer Verteilung auf } \mathcal{I}\times\mathcal{J}\}.$$

Weiter sei $(P_p)_{p\in\Theta}$ ein statistisches Modell, so dass $(X_1, Y_1), \ldots, (X_n, Y_n)$ unter P_p unabhängig und nach p verteilt sind (d.h. $P(X_k = i, Y_k = j) = p_{ij}$). Setze für $i \in \mathcal{I}, j \in \mathcal{J}$

$$\begin{aligned} S_{ij} &:= |\{k : X_k = i, Y_k = j\}|, \\ S_{i\bullet} &:= |\{k : X_k = i\}| = \sum_{j\in\mathcal{J}} S_{ij}, \\ S_{\bullet j} &:= |\{k : Y_k = j\}| = \sum_{i\in\mathcal{I}} S_{ij} \end{aligned}$$

und $p_{i\bullet} := \sum_{j\in\mathcal{J}} p_{ij}, p_{\bullet j} := \sum_{i\in\mathcal{I}} p_{ij}$ für $p \in \Theta$. Sei

$$\Theta_0 = \{p \in \Theta : p_{ij} = p_{i\bullet} \cdot p_{\bullet j} \text{ für } i \in \mathcal{I}, j \in \mathcal{J}\},$$

$\Theta_1 = \Theta \setminus \Theta_0$. Dann ist (S_{11}, C) ein unverfälschter Test zum Niveau α, falls

$$\mathrm{Hyp}(S_{1\bullet}, S, S_{\bullet 1})(C) \leq \alpha. \tag{6.5}$$

Die Gleichung (6.5) meint, dass die Menge C unter der Hypergeometrischen Verteilung mit den Parametern $(S_{1\bullet}, S, S_{\bullet 1})$ ein Wahrscheinlichkeit kleiner oder gleich α hat.

Beweis. Gegeben sind Daten von S Objekten. Wir gehen davon aus, dass $S_{1\bullet}, S_{2\bullet}, S_{\bullet 1}$ und $S_{\bullet 2}$ bekannt sind. In diesem Fall gilt es, die S Objekte so auf die Kontingenztafel zu verteilen, dass die Randeinträge stimmen. Wir stellen uns vor, die $S_{\bullet 1}$ Elemente mit Merkmal 2, Möglichkeit 1 seien weiß, die anderen $S_{\bullet 2}$ schwarz. Aus diesen wählen wir $S_{1\bullet}$ aus. Sind die Merkmale wirklich unabhängig, so erhalten wir eine $\mathrm{Hyp}(S_{1\bullet}, S, S_{\bullet 1})-$ verteilte Zufallsgröße als Eintrag von S_{11}. Ist diese entweder zu groß oder zu klein, können wir die Unabhängigkeit verwerfen. □

Bemerkung 6.21. In (6.5) besteht eine vermeintliche Unsymmetrie zwischen dem ersten und zweiten Merkmal, da die Parameter der hypergeometrischen Verteilung $S_{1\bullet}$ die Anzahl der gezogenen und $S_{\bullet 1}$ die Anzahl der weißen Kugeln bedeutet. Allerdings gilt

$$\begin{aligned}\mathrm{Hyp}(S_{1\bullet}, S, S_{\bullet 1})(k) &= \frac{\binom{S_{\bullet 1}}{k}\binom{S_{\bullet 2}}{S_{1\bullet}-k}}{\binom{S}{S_{1\bullet}}} = \frac{S_{\bullet 1}!S_{\bullet 2}!S_{1\bullet}!S_{2\bullet}!}{S!k!(S_{\bullet 1}-k)!(S_{1\bullet}-k)!(S-S_{\bullet 1}-S_{1\bullet}+k)!}\\ &= \mathrm{Hyp}(S_{\bullet 1}, S, S_{1\bullet})(k),\end{aligned}$$

was diese vermeintliche Unsymmetrie erklärt.

B 6.13 *Freiburger Fussballfans*: In der Freiburger Innenstadt werden nach den Spielen des FC Fahrradkontrollen gemacht und nötigenfalls Alkoholtests angeordnet. Die zu untersuchende Frage ist, ob das Ergebnis einen Einfluss auf den Alkoholisiserungsgrad der Fahrradfahrer hat. Per Se haben wir keine Vermutung, wenden also einen zweiseitigen Test an. Es werden 400 Personen untersucht. Das Ergebnis ist

	nüchtern	alkoholisiert	$\sum$
gewonnen	41	74	115
verloren	96	189	285
$\sum$	137	263	400

Kann aufgrund dieser Daten die Hypothese der Unabhängigkeit auf einem Signifikanzniveau von 5% verworfen werden? Wären die beiden Merkmale unabhängig, so wäre $E[S_{11}] = 400 \cdot \frac{137}{400} \cdot \frac{115}{400} \approx 39.39$. Beobachtet wurden aber $S_{11} = 41$. Wir bestimmen den (zweiseitigen) kritischen Bereich Zunächst stellen wir fest, dass

$$H(115, 400, 137)\{0, ..., 30\} \approx 0.018 < 0.032 = H(115, 400, 137)\{0, ..., 31\}$$

und

$$H(115, 400, 137)\{49, ..., 115\} \approx 0.018 < 0.030 = H(115, 400, 137)\{48, ..., 115\}.$$

Fisher's exakter Test

überprüft, ob zwei Merkmale, die in jeweils zwei möglichen Ausprägungen vorliegen, stochastisch unabhängig sind.

Statistisches Modell	$(X_1, Y_1), \ldots, (X_n, Y_n)$ unabhängig, identisch verteilt, $$S_{ij} := \lvert\{k : X_k = i, Y_k = j\}\rvert,$$ $$S_{i\bullet} := \lvert\{k : X_k = i\}\rvert = \sum_{j \in \mathcal{J}} S_{ij},$$ $$S_{\bullet j} := \lvert\{k : Y_k = j\}\rvert = \sum_{i \in \mathcal{I}} S_{ij}.$$
Hypothesen	$H_0 : P(X_k = i, Y_k = j) = P(X_k = i) \cdot P(Y_k = j)$ für alle i, j $H_1 : P(X_k = i, Y_k = j) \neq P(X_k = i)P(Y_k = j)$ für mindestens ein Paar i, j
Teststatistik	S_{11} ist – gegeben $S_{1\bullet}, S_{2\bullet}, S_{\bullet 1}$ und $S_{\bullet 2}$ – nach $H(S_{1\bullet}, S, S_{\bullet 1})$ verteilt.
Ablehnungsbereich	$S_{11} \in C := \{0, ..., k, \ell, ..., S_{1\bullet} \wedge S_{\bullet 1}\}$ falls $H(S_{1\bullet}, S, S_{\bullet 1})(C) \leq \alpha$.

Abbildung 6.6: Schematische Darstellung von Fisher's exaktem Test

Damit ist

$$C = \{0, ..., 30, 49, ..., 115\}$$

der Ablehnungsbereich für Fisher's exakten Test zum Fehlerniveau von 0.05. Die Nullhypothese der Unabhängigkeit kann nicht abgelehnt werden. Der Anteil der alkoholisierten Fahrer ist also in beiden Fällen gleich hoch.

Die Angaben in diesem Beispiel sind natürlich frei erfunden.

7. Stochastische Optimierung

Stochastische Optimierung meint das Optimieren mit stochastischen Methoden.[1] Das abstrakte Problem ist dabei wiefolgt: Gegeben sei eine endliche Menge E und eine Funktion $f : E \to \mathbb{R}$. Finde s_* (oder s^*) mit

$$f(s_*) = \min_{s \in E} f(s) \qquad \big(\text{oder } f(s^*) = \max_{s \in E} f(s)\big).$$

Minimierungs- und Maximierungsprobleme kann man dabei durch den Übergang zur Funktion $-f$ ineinander umwandeln.

B 7.1 *Traveling Salesman Problem (TSP)*: Ein Handlugnsreisender will m Städte besuchen und dann nach Hause zurückkehren. Der Abstand der Städte ist dabei durch eine (symmetrische) $m \times m$-Matrix D gegeben. Bei seiner Rundreise will er seine zurückgelegte Wegstrecke minimieren. Hier ist

$$E = \mathcal{S}_m$$

die Menge aller Permutationen von $\{1, \ldots, m\}$ und

$$f(\xi) = \sum_{i=1}^{m-1} D_{\xi_i, \xi_{i+1}} + D_{\xi_m, \xi_1}. \tag{7.1}$$

Legt man einen Wert d fest und fragt, ob es eine Permutation gibt mit $f(\xi) \leq d$, so handelt es sich um das TSP-Entscheidungsproblem. Dieses ist NP-vollständig: Bis heute sind keine deterministischen Algorithmen bekannt, die das TSP-Entscheidungsproblem in einer Zeit, die polynomial ist in m, entscheiden können. Ein Ansatz, der hier weiterführen kann, ist es, stochastische Algorithmen zu untersuchen, die zumindest solche Wege finden, deren Länge kurz ist. Dabei kann man nur hoffen, mit stochastischen Methoden eine möglichst gute Lösung zu erhalten, jedoch nicht hoffen, mit stochastischen Methoden einen Beweis für die Optimalität einer Lösung zu finden.

Zwei Methoden der stochastischen Optimierung, die man auf obige Probleme anwenden kann, ist *Simulated Annealing* und *genetische Algorithmen*.

[1] Oft wird unter *stochastischer Optimierung* auch das Optimieren unter unsicheren (stochastischen) Daten verstanden. Das wollen wir hier nicht behandeln.

7.1 Simulated Annealing

Der Begriff des *Annealing* kommt aus der Metallverarbeitung und beschreibt das *Auskühlen* von Metallen. Dort sieht man sich dem Problem gegenüber, ein Metall so auszukühlen, dass lokal keine Spannungen entstehen, die das Material instabil werden lassen. Mit anderen Worte will man die Energie des Metalls durch das Ausglühen minimieren.

Die Idee des *Simulated Annealing* ist die folgende: Starte eine Markoff-Kette mit Zustandsraum E und Übergangsmatrix P_1 und einer stationären Verteilung, die großes Gewicht auf Zustände $s \in E$ mit kleinem $f(s)$ legt. Warte eine Zeit N_1, bis diese annähernd im Gleichgewicht ist. Starte nun in diesem Zustand eine neue Markoff-Kette mit Übergangsmatrix P_2, die noch größeres Gewicht auf $s \in E$ mit kleinem $f(s)$ legt. Warte wieder eine Zeit N_2 bis diese Markoff-Kette im Gleichgewicht ist. Setzt man dieses Vorgehen fort, so besteht die Hoffnung, dass sich die Markoff-Kette am Ende in einem Zustand mit minimalem $f(s)$ befindet. Formal wird dies durch eine zeitinhomogene Markoff-Kette mit Übergangsmatrix

$$P^{(n)} = \begin{cases} P_1, & 0 < n \leq N_1 \\ P_2, & N_1 < n \leq N_1 + N_2 \\ \dots & \end{cases}$$

beschrieben.

Zunächst wollen wir die *Boltzmann-Verteilung* für eine Funktion $f : E \to \mathbb{R}$ definieren, die obige Forderung, hohes Gewicht auf Zustände s mit kleinem $f(s)$ zu legen, erfüllt. Im Anschluss definieren wir eine Markov-Kette, die die Boltzmann-Verteilung als stationäre Verteilung hat.

Definition 7.1 (Boltzmann-Verteilung)**.** Sei $f : E \to \mathbb{R}$ nach unten beschränkt und $T > 0$. Dann ist die *Boltzmann-Verteilung* $\pi_{f,T}$ gegeben durch

$$\pi_{f,T}(s) = \frac{1}{Z_{f,T}} \exp\Big(-\frac{f(s)}{T}\Big),$$
$$Z_{f,T} = \sum_{s \in E} \exp\Big(-\frac{f(s)}{T}\Big).$$

Der Parameter T heißt *Temperatur*.

Je kleiner T als Parameter in der Boltzmann-Verteilung, desto größer sind die relativen Unterschiede für Zustände $s, t \in E$ mit $f(s) \neq f(t)$. Je größer $f(t)$, desto kleiner ist das Gewicht $\pi_{f,T}(t)$. Noch genauer gilt das folgende Resultat.

Lemma 7.2. *Sei $T > 0$, E eine endliche Menge und $f : E \to \mathbb{R}$ beliebig. Weiter sei*

$$E_* := \{s \in E : f(s) = \min_{t \in E} f(t)\}.$$

Dann gilt

$$\lim_{T \to 0} \pi_{f,T}(E_*) = 1.$$

Beweis. Sei $E_* = \{s_1, \ldots, s_n\}$, $a = f(s)$ und $b = \min_{t \in E \setminus E_*} f(t)$. Es gilt

$$\lim_{T \to 0} \exp\Big(\frac{a-b}{T}\Big) = 0$$

und damit

$$\begin{aligned}\pi_{f,T}(E_*) &= \frac{|E_*| \exp\big(-\frac{a}{T}\big)}{\sum_{t \in E} \exp\big(-\frac{f(t)}{T}\big)} = \frac{|E_*| \exp\big(-\frac{a}{T}\big)}{|E_*| \exp\big(-\frac{a}{T}\big) + \sum_{t \in E \setminus E_*} \exp\big(-\frac{f(t)}{T}\big)} \\ &\geq \frac{|E_*| \exp\big(-\frac{a}{T}\big)}{|E_*| \exp\big(-\frac{a}{T}\big) + (|E| - |E_*|) \exp\big(-\frac{b}{T}\big)} = \frac{1}{1 + (\frac{|E|}{|E_*|} - 1) \exp\big(\frac{a-b}{T}\big)} \xrightarrow{T \to 0} 1.\end{aligned}$$

□

Bemerkung 7.3 (Herleitung der Metropolis-Kette)**.** Für eine Verteilung π auf E benötigen wir nun noch eine Markov-Kette $(X_t)_{t=0,1,\ldots}$ mit stationärer Verteilung π. Diese wird durch die Metropolis-Kette geliefert, für die π sogar reversibel ist (siehe Definition 5.8). Diese erhält man wiefolgt:

- Wähle eine Übergangsmatrix Q einer irreduziblen Markov-Kette.
- Wähle die Übergangsmatrix P der gesuchten Markov-Kette so, dass

  $$\pi_i P_{ij} = \min(\pi_i Q_{ij}, \pi_j, Q_{ji})$$

  für $i \neq j$. Dann ist wegen der Symmetrie der rechten Seite π reversible Verteilung für die Markov-Kette mit Übergangsmatrix P.
- Auflösen der letzten Gleichung liefert

  $$P_{ij} = \min\Big(1, \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}}\Big) Q_{ij}.$$

  Dies ist die gesuchte Übergangsmatrix für

  $$P_{ii} = 1 - \sum_{j \neq i} P_{ij}.$$

Da nämlich $\sum_{j\neq i} Q_{ij} \leq 1$, muss auch $\sum_{j\neq i} P_{ij} \leq 1$ gelten.

Definition 7.4 (Metropolis-Kette)**.** Sei Q die Übergangsmatrix einer irreduziblen Markov-Kette mit Zustandsraum E und π eine Verteilung auf E. Dann ist die Metropolis-Kette $X = (X_t)_{t=0,1,\ldots}$ die Markov-Kette mit Übergangsmatrix P mit

$$P_{ij} = \min\Big(1, \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}}\Big) Q_{ij}$$

und

$$P_{ii} = 1 - \sum_{j\neq i} P_{ij}.$$

B 7.2 *Traveling Salesman Problem*: Wir wollen nun für das TSP eine Metropolis-Kette konstruieren, die die Boltzmann-Verteilung $\pi_{f,T}$ zur Funktion f aus Beispiel 7.1 als stationäre Verteilung hat. Zur Definition der Metropolis-Kette benötigen wir zunächst einen Graphen mit der Punktmenge E. Wir müssen also angeben, welche Permutationen ξ und η benachbart sein sollen. Dazu definieren wir für $i < j$

$$\sigma_{i,j}(\xi) = (\xi_1, \ldots, \xi_{i-1}, \xi_j, \xi_{j-1}, \ldots, \xi_{i+1}, \xi_i, \xi_{j+1}, \ldots, \xi_m).$$

Also ist $\sigma_{i,j}(\xi)$ die Permutation, die entsteht, indem man in ξ den Abschnitt $\xi_i, \ldots, \xi_j$ in umgekehrter Reihenfolge durchläuft. Damit definieren wir

$$\xi \sim \eta :\Longleftrightarrow \exists i < j : \eta = \sigma_{i,j}(\xi).$$

Damit hat jede Permutation genau $\binom{m}{2}$ Nachbarn. Wir betrachten die Übergangsmatrix Q mittels

$$Q_{\xi,\eta} = \begin{cases} \frac{1}{\binom{m}{2}}, & \xi \sim \eta \\ 0, & \xi \not\sim \eta. \end{cases} \tag{7.2}$$

Die Markov-Kette mit dieser Übergangsmatrix ist irreduzibel, weil jede Permutation als Komposition von Transpositionen dargestellt werden kann. Für die dazugehörigen Boltzmann-Verteilung entsteht die Metropolis-Kette nun durch Definition 7.4 aus

$$P_{\xi,\eta} = \begin{cases} \frac{1}{\binom{m}{2}} \min\Big(\exp\Big(\frac{f(\xi)-f(\eta)}{T}\Big), 1\Big), & \xi \sim \eta \\ 0, & \xi \not\sim \eta, \xi \neq \eta \\ 1 - \sum_{\xi'\sim\xi} \frac{1}{\binom{m}{2}} \min\Big(\exp\Big(\frac{f(\xi)-f(\xi')}{T}\Big), 1\Big), & \xi = \eta. \end{cases}$$

Ist die Markoff-Kette im Zustand $X_t = \xi$, so verfährt sie also folgendermaßen:

- Wähle zuerst nach der Gleichverteilung zwei Positionen $1 \leq i < j \leq m$.
- Ist $f(\sigma_{i,j}(\xi)) \leq f(\xi)$, so ist $X_{t+1} = \sigma_{i,j}(\xi)$. Ist jedoch $f(\sigma_{i,j}(\xi)) > f(\xi)$, so ist

$$X_{t+1} = \begin{cases} \sigma_{i,j}(\xi) & \text{mit Wahrscheinlichkeit } \exp\left(\frac{f(\xi)-f(\eta)}{T}\right), \\ \xi & \text{mit Wahrscheinlichkeit } 1 - \exp\left(\frac{f(\xi)-f(\eta)}{T}\right). \end{cases}$$

Es ist wichtig zu betonen, dass wir zur Berechnung der Übergangswahrscheinlichkeiten der Metropolis-Kette die Normalisierungskonstanten $Z_{f,T}$ nicht berechnen mussten. Das liegt an der allgemeinen Eigenschaft der Metropolis-Ketten, dass zu deren Berechnung nur Quotienten der Werte der stationären Verteilung bekannt sein müssen.

Die Markov-Kette X ist irreduzibel und aperiodisch, da $P_{\xi,\xi} > 0$ für manche ξ (falls f nicht konstant ist). Durch Theorem 5.13 und da $\pi_{f,T}$ reversibel für P ist, wissen wir, dass $P[X_t = i] \xrightarrow{t\to\infty} \pi_{f,T}(i)$ gelten muss. Schön wäre es nun auszunutzen, dass für E_* wir in Lemma 7.2

$$P[X_t \in E_*] \xrightarrow{t\to\infty} \pi_{f,T}(E_*) \xrightarrow{T\to 0} 1.$$

Allerdings sind hier zwei Grenzübergänge zu betrachten. Dies wird durch eine Wahl von $T_1, T_2, \ldots$ und $N_1, N_2, \ldots$ durchgeführt: Ziel ist es immer, die Temperatur zu erniedrigen, um dem Minimum näher zu kommen. Passiert dies jedoch zu schnell, so läuft man Gefahr, nur ein lokales Minimum von f zu finden. Ein Theorem von Geman und Geman besagt: Gilt für die Temperatur $T^{(t)}$ im t-ten Schritt

$$T^{(t)} \geq \frac{|E|(\max_{s\in E} f(s) - \min_{s\in E} f(s))}{\log t},$$

so ist

$$P[X_t \in E_*] \xrightarrow{t\to\infty} 1.$$

Leider ist das Annealing-Schema, das durch dieses Theorem gegeben wird, sehr langsam. (Beispielsweise ist schon $|E|$ sehr groß.) In der Praxis wird man mehrere Annealing-Schemas ausprobieren, um sagen zu können, welches in akzeptabler Zeit gute Resultate liefert.

B 7.3 *Zu schnelles Abkühlen*: Um ein einfaches Beispiel anzugeben, in dem ein zu schnelles Abkühlen der Markoff-Kette zu einem falschen Ergebnis führt, betrachten wir $E = \{s_1, s_2, s_3, s_4\}$. Weiter haben wir

$$f(s_1) = 1, \quad f(s_2) = 2, \quad f(s_3) = 0, \quad f(s_4) = 2.$$

Um eine Metropolis-Kette zu definieren, müssen wir angeben, welche Punkte benachbart sind. Dies ist durch nachfolgende Grafik veranschaulicht (jeweils mit Überganswahrscheinlichkeiten $\frac{1}{2}$:

units ¡1cm,1cm¿ x from 0 to 4, y from 0 to 4 1 1 3 1 3 3 1 3 1 1 / • [cC] at 1 1 • [cC] at 1 3 • [cC] at 3 3 • [cC] at 3 1 s_1 [cC] at 1 3.2 s_2 [cC] at 3 3.2 s_3 [cC] at 3 .8 s_4 [cC] at 1 .8

Damit ergibt sich die Übergangsmatrix der Metropolis-Kette

$$P = \begin{pmatrix} 1-e^{-1/T} & \frac{1}{2}e^{-1/T} & 0 & \frac{1}{2}e^{-1/T} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2}e^{-2/T} & 1-e^{-2/T} & \frac{1}{2}e^{-2/T} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

Wir starten die Markoff-Kette in $X_0 = s_1$ und verwenden ein Annealing- Schema $T^{(1)}, T^{(2)}, \ldots$. Sie wird sicher dann nicht in das globale Minimum s_3 konvergieren, wenn sie immer im Zustand s_1 bleibt. Es gilt

$$P[X_0 = X_1 = \cdots = s_1] = \lim_{t\to\infty} \prod_{i=1}^{t} (1 - e^{-1/T^{(i)}}) = \prod_{i=1}^{\infty} (1 - e^{-1/T^{(i)}}).$$

Damit gilt nach einem Satz der Analysis

$$P[X_0 = X_1 = \cdots = s_1] > 0 \quad \text{genau dann wenn} \quad \sum_{i=1}^{\infty} e^{-1/T^{(i)}} < \infty.$$

Damit besteht beispielsweise für $T^{(i)} = \frac{1}{i}$ die Möglichkeit, dass die Markoff-Kette immer im Zustand s_1 bleibt und somit nicht in das globale Minimum s_3 konvergiert.

Allgemein kann man sagen, dass ein solches Verhalten auf zwei Dinge zurückzuführen ist: ein zu schnelles Annealing-Schema und ein Punkt s, der zwar ein lokales, aber kein globales Minimum von f ist.

7.2 Genetische Algorithmen

Bei genetischen Algorithmen handelt es sich um Algorithmen, die durch eine ähnliche Dynamik wie sie bei der Fortpflanzung von Individuen beobachtet wird, versuchen, eine Zielfunktion zu maximieren. Diese Zielfunktion heist meist *Fitness*funktion.

Bemerkung 7.5 (Vererbung)**.** In der Biologie geht man davon aus, dass alle Individuen durch ihr genetisches Material, also ihre Chromosomen, bestimmt werden. Chromosomen bestehen aus DNA, also aus der aneinandergerehten Aminosäuren. Mathematisch können Chromosomen als endliche Folge von Buchstaben aus einem Alphabet modelliert werden. Jeder solchen Folge wird der Wert einer Fitness-Funktion zugeordnet. Genetische Algorithmen modellieren nun die Fortpflanzung dieses genetischen materials. Pflanzen sich

Organismen fort, so liegt von beiden Eltern je ein Chromosom vor. Beispielsweise liegen zwei Chromosomen mit der Codierung

$$\begin{array}{c} 1\ 0\ 0\ 1\ 0\ 1\ 1\ 0 \\ \text{und} \\ 0\ 1\ 0\ 0\ 1\ 1\ 1\ 1 \end{array}$$

vor; jedes Chromosom besteht also aus acht Aminosäuren, die wir auch Bits nennen wollen. (Dabei haben wir angenommen, dass es nur genau zwei verschiedene Aminosäuren gibt, was eine Vereinfachung darstellt.) Nun findet *Mutation, Rekombination* und *Vererbung mit Selektion* statt.

- *Mutation:* Es besteht die Möglichkeit, dass sich einzelne Bits zufällig ändern. Passiert die beispielsweise an der vierten Stelle des Chromosoms, so ergibt sich

  $$1\ 0\ 0\ \mathbf{1}\ 1\ 1\ 10$$

- *Rekombination* (oder engl. *Crossover*): Das Chromosom des Nachkommens besteht aus Stücken beider Elternteile. Beispielsweise werden die ersten drei und die letzten beiden Bits des ersten und die restlichen Bits des zweiten Elternteils genommen. Dies ergibt das Chromosom

  $$\begin{array}{cccc|ccc|cc} \mathbf{1} & \mathbf{0} & \mathbf{0} & & 1 & 0 & 1 & \mathbf{1} & \mathbf{0} \\ 0 & 1 & 0 & & \mathbf{0} & \mathbf{1} & \mathbf{1} & 1 & 1 \end{array} \longrightarrow 1\ 0\ 0\ 0\ 1\ 1\ 1\ 0$$

  Die Punkte des Chromosoms, an denen die Rekombination stattfindet, heißen Crossover-Punkte.

- *Reproduktion mit Selektion:* Es werden zwei Individuen i und j der Population gezogen, das Individuum i bekommt einen Nachkommen, und j stirbt. Die Wahrscheinlichkeit, dass sich i fortpflanzen darf ist gegeben durch proportional zur Fitness von Individuum i.

Die Evolutionstheorie sagt voraus, dass sich Organismen, die besser an ihre Umwelt angepasst sind oder auf andere Art und Weise *fitter* als andere, sich öfter fortpflanzen werden, was durch den letzten Schritt realisiert ist.

Bemerkung 7.6 (In silico Evolution)**.** Ziel ist es, das Maximum einer Funktion $f : E \to \mathbb{R}$ zu finden. Hierzu sein eine Population der Größe N gegeben. Individuum i hat einen Typ $\alpha(i) \in E$ im Typenraum E. Die Fitness von Typ α ist nun gegeben durch $f(\alpha)$. Da sich gute Typen eher durchsetzen als schlechte, betrachtet man eine Markov-Kette $X = (X_t)_{t=0,1,...}$ mit $X_t = (\alpha_t(1), ..., \alpha_t(N))$, wobei $\alpha_t(i)$ der Typ von Individuum i zur Zeit t ist. Man geht nun folgendermaßen vor, um das gesuchte Maximum zu finden: Mit bestimmten Wahrscheinlichkeiten findet in jedem Schritt eine der folgenden drei Mechanismen statt.

- *Mutation:* Wähle ein Individuum i und setze $\alpha_{t+1}(i) = \beta$ mit Wahrscheinlichkeit $Q_{\alpha_t(i),\beta}$ für eine stochastische Matrix Q.
- *Rekombination:* Wähle zwei Individuen i, j aus und setze $\alpha_{t+1}(i) = \beta$ mit Wahrscheinlichkeit $Q'_{\alpha_t(i),\alpha_t(j);\beta}$ für eine stochastische Matrix Q'.
- *Reproduktion und Selektion:* Wähle ein Individuum j rein zufällig aus. Weiter wähle Individuum i mit Wahrscheinlichkeit $\frac{f(i)}{\sum_{j\in N} f(j)}$ aus. Setze nun $\alpha_{t+1}(j) = \alpha_t(i)$. (Das bedeutet, dass Individuum i stirbt und durch einen Nachkommen von i ersetzt wird.)

Dieses Schema führt hoffentlich dazu, dass sich die Population aus immer fitteren Individuen zusammensetzt.

B 7.4 *Traveling Salesman Problem (TSP)*: Wir wollen nun einen genetischen Algorithmus angeben, mit dem man eventuell das TSP lösen kann. Wir müssen hierzu definieren, was genau ein Individuum ist und welche Mutations- Rekombinationsmechanismen zugelassen sind und welche Fitness-Funktion wir verwenden.

Der Typenraum ist hier der Raum $E = \mathcal{S}_m$ aller Permutationen der Länge m. Die Fitness von $\xi \in E$ (die es ja zu maximieren gilt) ist z.B. gegeben durch

$$g(\xi) = \frac{1}{f(\xi)}$$

mit f aus (7.1).

- *Mutation:* Eine Mutation besteht aus der zufälligen Vertauschung zweier besuchter Orte, d.h. Q ist gegeben durch (7.2).
- *Rekombination:* Sei $i \in \{1, ..., m\}$ rein zufällig gewählt. Um ξ, η an der Stelle i zu rekombinieren, gehen wir folgendermaßen vor: Bis Ort i werden die Orte in der Reihenfolge von i besucht. Im Anschluss werden die Orte, die noch nicht besucht worden, in der Reihenfolge besucht, wie sie durch η vorgegeben sind. Sei etwa $m = 5$, $\xi = (3, 5, 2, 1, 4), \eta = (5, 3, 1, 4, 2)$, so rekombinieren die beiden an Ort 5 zu $(3, 5, 1, 4, 2)$.
- *Reproduktion mit Selektion:* Dies folgt wieder dem allgemeinen Schema aus Bemerkung 7.6.